Language and Literature

http://lal.sagepub.com

Book Review: Corpus-Based Language Studies: An Advanced Resource Book by Tony McEnery, Richard Xiao and Yukio Tono, 2006. New York: Routledge, pp. xx +386 ISBN 0415286239 (pbk)

Robin Straaijer

Language and Literature 2009; 18; 394
DOI: 10.1177/0963947009018004070l

Corpus-Based Language Studies: An Advanced Resource Book by Tony McEnery, Richard Xiao and Yukio Tono, 2006. New York: Routledge, pp. xx +386 ISBN 0415286239 (pbk)

The preface of *Corpus-Based Language Studies* states that it 'covers the "how to" as well as the "why"' (p. xvii) of corpus linguistics, and on the whole it succeeds in that aim. The book is set up with three parts: Section A 'Introduction', Section B 'Extension', and Section C 'Exploration', with each section being made up of several units. The first seven units of Section A introduce corpus linguistics as a methodology and a practice, and focus on several key aspects. Representativeness and sampling are discussed in Unit A2, and corpus mark-up and annotation in units A3 and A4. Dealing with the statistical implications of quantitative corpus studies is the topic of Unit A6, and units A8 and A9 explain how to go about creating a corpus. Unit A10 discusses how the use of corpora can be incorporated into a number of areas of linguistics, ranging from dialectology to semantics to discourse analysis.

By means of a series of extracts from previously published articles Section B elaborates on a number of the subjects introduced in section A. Units B1 and B2 deal primarily with methodological issues, while the remaining four units deal with the use of corpora in some of the areas introduced in Unit A10, specifically lexical and grammatical studies, language variation studies, contrastive and diachronic studies, and language teaching and learning. The benefit of incorporating these excerpts is that it allows for a relatively large variety of subjects to be presented in a concise fashion.

Section C of the book takes the reader through six case studies in corpus linguistics. In a way, this section is an elaboration from McEnery and Wilson (2001), which presented a single case study. These illustrations of the use of various corpora, concordancers and related applications such as statistical packages is one of its greatest strengths.

The book's target audience, as given in the series editors' preface, is upper graduates and postgraduates. Considering that, however, the book seems to assume a relatively low entry-level of knowledge of corpus linguistics from the reader as evinced by the appearance of Unit A1.3 'What is a corpus?' There are a few aspects that could perhaps have been elaborated on a little more.

Unit A1.7 deals with the corpus-based versus the corpus-driven approach, and only mentions in passing that the corpus-driven approach advocates the inclusion of whole texts. This, according to the authors, makes it 'nearly unavoidable that a small number of texts may seriously affect, either by theme or in style, the balance of a corpus' (p. 9). In defence of the corpus-driven approach it is possible to argue, as probably its proponents will, that when a cumulatively balanced corpus is skewed, it is likely that this is because it is not yet big enough to have achieved cumulative representativeness. Another argument for including whole texts, which is not mentioned, is that while it is not necessary to include whole texts for all kinds of analyses, it is of course absolutely vital to do so when the features of text types themselves is the subject of the investigation.

More and more individual researchers are creating specialized corpora for their own use, but there is no mention of these in the book, whether they are called 'focused corpora', 'mini corpora' or 'third generation corpora'. Since this is a book meant for individuals as well as groups to be able to use corpora in general, not just the *prêt-a-porter* corpora such as the BNC or the Helsinki Corpus, it might perhaps not be amiss to pay some more attention to corpus creation in the unit on DIY-corpora. It would be useful to read more about the methods involved in collecting materials for a corpus – apart from downloading from the internet – and about whether or not sampling is required and whether or not to annotate and if so, to what extent.

An example of this is illustrated by the following remark in Unit A4.2 'all of the four criticisms of corpus annotation can be dismissed … quite safely' (p. 32). This would lead one to believe that corpus annotation is virtually a requirement for all corpora. In the case of small DIY corpora, however, this is not straightforward methodologically and often problematic practically. For very small corpora, say those with less than a few thousand words, it may still be possible to annotate by hand. However, for larger-sized 'small' corpora – those having between 10,000 and 100,000 words, which still qualify as mini-corpora – annotation and tagging by hand is likely to be too time-consuming for the individual researcher. In addition, the accuracy of automated taggers, while relatively high, may still lead to an unacceptably high number of errors in small corpora.

The use of automated semantic tagging software with small corpora may present another kind of potential problem. In order to be able to get any use out of the application, the individual researcher needs to be in agreement with the designers of the tagging software about the parameters of the semantic categories. Where corpora of modern English are concerned, this need not cause any problem, but when we are dealing with corpora of historical language, it cannot be assumed that the historical semantic categories are the same as the modern ones. The same is true, though to a lesser extent, for automated POS-tagging of historical data.

The following remark from Unit A10.15 is rather interesting:

> it is important to keep in mind that the findings based on a particular corpus only tell us what is true in that corpus, though a representative corpus allows us to make reasonable generalizations about the population from which the corpus was sampled. Nevertheless, unwarranted generalizations can be misleading. (p. 121)

It seems that it is explicitly mentioned as something important to keep in mind for a quantitative approach such as corpus linguistics. However, is not this applicable to all forms of analysis, quantitative and qualitative, corpus-based or introspection-based? And is not then a more representative corpus to be preferred to examples taken from introspection, when there is a choice between the two?

Generally speaking, Section B gives a good overview of some areas of and issues in corpus linguistics. As such, it is a good addition to works such as Sampson and McCarthy (2005). Personally, I would like to have seen a little more of the author's opinions in the text in which the abstracts are embedded. It would be useful to have more of the author's view on the issues presented on the topics discussed, which would also create more cohesion within the units.

'A badly designed corpus, or indeed, even a well-designed corpus, when used for a purpose it is not designed for, may provide misleading results' (p. 262). I think this remark from Unit C3.5 'Conversational speech in American English' is important enough to merit more prominence. Perhaps it ought to have been more mentioned in Unit A2 'Representativeness, balance and sampling' or A8 'Going solo: DIY corpora' and elaborated on in that same section.

Probably one of the most useful case studies is Unit C5, in which Xiao and McEnery present their application of Biber's multi-feature multi-dimensional (MF/MD) approach (Biber 1988) to genre analysis. It also very usefully presents WordSmith KeyWords as a possible alternative for Biber's often daunting procedure. The contrast between the two approaches is illustrative and enlightening, and provides the reader with an informed choice of methodologies.

For a budding corpus linguist, the book is certainly illuminating. Its main strength is that it brings many opinions of a theoretical and methodological nature together, thereby providing a good overview of corpus linguistics, which may not be so easy to glean from the literature someone with a new-found interest in the subject might happen across. The three sections of the book are well balanced with respect to each other and the increasing level of complexity they exhibit works well. The list of web addresses at the end of the book is useful, as such lists generally are, but URLs are prone to change, as is also the case in this book. The URL for Beaugrande's article on page 140 is a dead link, which is a bit disappointing for anyone who would like to read the full article. In addition, 10 out of the 26 URLs given were unfortunately no longer functioning at the time of the review, including most notably, those for the book's companion websites.

On the whole the book achieves the aims stated in the preface, not just with regard to the 'how to' and 'why' of corpus linguistics, but also in its aim 'to bring readers up to date with the latest developments in corpus-based language studies' and to teach readers 'when and how to combine these approaches with other methodologies' (p. xvii). It also appears to succeed in making the reader 'be able to become a corpus linguist having read the book' (p. xvii).

References

Biber, D. (1988) *Variation Across Speech and Writing*. Cambridge: Cambridge University Press.

McEnery, T. and Wilson, A. (2001) *Corpus Linguistics: An Introduction.* Edinburgh: Edinburgh University Press.

Sampson, G. and McCarthy, D. (2005) *Corpus Linguistics: Readings in a Widening Discipline*. London: Continuum.

Robin Straaijer
Leiden University, The Netherlands

The Body in Flannery O'Connor's Fiction
by Donald E. Hardy, 2007. Columbia, South Carolina: University of South Carolina Press, pp. ix + 188
ISBN 978 1 57003 698 9 (hbk)

This book has two main aims. The first aim is to build on existing critical literature relating to the fiction of Flannery O'Connor, much of which has focused on O'Connor's exploration of the sacramental, the incarnational, and the grotesque (see Lake, 2005; Srigley, 2004; as well as Hardy's previous book *Narrating Knowledge in the Fiction of Flannery O'Connor*, 2002). It is worth considering briefly what these terms mean, as they are central to Hardy's study. The sacramental and the incarnational refer to the interconnectedness of spirit and matter, mind and body – a theme that is central in the work of Flannery O'Connor. Specifically, sacramentalism is the belief that spiritual reality exists and is present in the physical world; the incarnational is the way in which this spiritual presence reveals itself, that is 'the *method* of signalling the incarnation of spirit in