Book Review

Discourse on the Move: Using Corpus Analysis to Describe Discourse Structure

Douglas Biber, Ulla Connor, and Thomas A. Upton
(Northern Arizona University and Indiana University–Indianapolis)

John Benjamins Publishing (Studies in corpus linguistics, edited by Elena Tognini-Bonelli, volume 28), 2007, xii+289 pp; hardbound, ISBN 978-90-272-2302-9, $142.00, €105.00

Reviewed by
Marina Santini
University of Glasgow

The study of discourse can be undertaken from different perspectives (e.g., linguistic, cognitive, or computational) with differing purposes in mind (e.g., to study language use or to analyze social practices). The aim of *Discourse on the Move* is to show that it is possible and profitable to join quantitative and qualitative analyses to study discourse structures. Whereas corpus-based quantitative discourse analysis focuses on the distributional discourse patterns of a corpus as a whole with no indication of how patterns are distributed in individual texts, manual qualitative analysis is always carried out on a small number of texts and does not support large generalizations of the findings. The book proposes two methodological approaches—top-down and bottom-up—that combine the quantitative and qualitative views into a *corpus-based description of discourse organization*. Such a description provides detailed analyses of individual texts and the generalization of these analyses across all the texts of a genre-specific corpus.

Top-down is the more traditional (not necessarily corpus-based) approach in which researchers establish functional–qualitative methods to develop an analytical framework capable of describing the types of discourse units in a target corpus. In this approach, linguistic–quantitative analyses come as a later step to facilitate the interpretation of discourse types. In contrast, the bottom-up approach begins with a linguistic–quantitative analysis based on the automatic segmentation of texts into discourse units on the basis of vocabulary distributional patterns. In this approach, the functional–qualitative analysis that provides an interpretation of the discourse types is performed as a later step.

Both top-down and bottom-up analyses can be broken down into seven procedural steps, but the order of the steps in the two approaches is not the same. The steps to be followed in top-down methods are these:

1. Determination of communicative/functional categories.
2. Segmentation.
3. Classification.
4. Linguistic analysis of each unit.
5. Linguistic description of discourse categories.
6. Text structure analysis.
7. Description of discourse organizational tendencies.

In the top-down studies presented in this book, the communicative/functional categories used to segment texts into meaningful units of discourse are identified through **move analysis** and **appeals analysis**. Moves segment texts according to the communicative functions of texts (Swales 1981, 1990), whereas the primary role of appeals—derived from the Aristotelian theory of persuasion and employed by Perelman (1982) to develop his theory of "new rhetoric"—is to make the reader 'act'. For this reason, appeals analysis is often applied to persuasive texts.

The steps to be followed in bottom-up methods are these:

1. Automatic segmentation.
2. Linguistic analysis of each unit.
3. Classification.
4. Linguistic description of discourse categories.
5. Determination of communicative/functional categories.
6. Text structure analysis.
7. Description of discourse organizational tendencies.

In the bottom-up studies presented in this book, the computational methods used to automatically identify vocabulary-based discourse units (VBDUs) are based on Hearst's (1994, 1997) TextTiling procedure. TextTiling is a quantitative procedure that compares the words used in contiguous segments of text. If the vocabulary in two segments is very similar, the two segments are analyzed as belonging to the same discourse unit; otherwise they are analyzed as two distinct units.

Therefore, one major difference between the two approaches is the unit of analysis. In the top-down case, the units of analysis (i.e., moves and appeals) are directly interpretable by discourse analysts, whereas bottom-up VBDUs are more complex to describe. However, the bottom-up method can be easily applied to large corpora and is replicable, whereas the top-down approach is more subjective. In both approaches, the linguistic analysis of discourse units relies on multidimensional analysis, the well-known statistical method developed by Biber to study linguistic variation.

Instructively, the book ends with a comparison of two independent analyses—one top-down and one bottom-up, described in Chapters 4 and 7, respectively—carried out on two different corpora, one containing biochemistry research articles, the other including articles about the more general discipline of biology. Although the expectation that both methods would reveal and underpin a similar inherent structure in the articles of both corpora is met to some extent, there are still several aspects that are problematic and have no ready explanation. The authors acknowledge that additional research is needed to shed more light on these aspects.

The book is easy to read and well structured. It consists of a preface, nine chapters—divided into two parts—and two appendices. Top-down methods (based on move analysis and appeals analysis) and their application to direct-mail letters, biochemistry research articles, and fund-raising letters are described in Chapters 2–5. Bottom-up methods (based on VBDUs) and their application to biology research articles and spoken university lectures are presented in Chapters 6–8. The introductory Chapter 1 contains a synthetic overview of the several perspectives and purposes guiding discourse analysis, a useful distinction between register and genre, the motivation for the book, and its research questions. The final Chapter 9 presents a critical comparison between the two approaches and outlines the many questions to be addressed in further studies and the directions to be explored in future research. The two appendices document the steps included in multidimensional analysis (Appendix I) and list the

lexico-grammatical features identified by the Biber tagger (Appendix II) and used in multidimensional analyses.

Overall, *Discourse on the Move* is interesting and inspiring. It is a valuable work of synthesis, where several previous approaches are combined to produce more extensive and comprehensive findings. The title suggests that discourse analysis is moving forward. This is indeed the impression that we have when reading the last pages of the final chapter. The authors list studies, publication of which is forthcoming, that were carried out using top-down and bottom-up approaches, and possible future directions that range from the investigation of multimodal texts to the integration of "contextual analysis" through, for example, surveys or interviews with informants.

In the list of desiderata, I would personally prioritize systematic comparisons between the findings returned by the top-down and bottom-up analyses on the same genre-specific corpus. This would allow us to assess whether the two methods are basically overlapping and can be used interchangeably, or whether they are complementary, so that it is worth applying them both on the same corpus. Such comparisons would also reveal important details—for example, whether multidimensional analyses are more effective and more easily interpretable on top-down units of analysis (i.e., moves and appeals) or on bottom-up VBDUs, or whether there is a way of marshaling the descriptive labels assigned to factor interpretations. As it is now, if we have a look at the functional labels shown in Table 9.1 (pages 246–247), it is unclear to what extent they are comparable.

The book is aimed at corpus linguists, but it can be informative also for those computational linguists, NLP researchers, and language engineers who are keen on incorporating language variation and genre specificities into computational models. For instance, the identification of regular variational discourse patterns could be helpful to fine-tune parsers and automatic summarizers.

As stated by the authors themselves, this book has only been able to "scratch the surface" of the corpus-based description of discourse organization. We look forward to a rapid growth of this research area.

References

Hearst, Marti. 1994. Multi-paragraph segmentation of expository texts. Technical Report 94/790, Computer Science Division (EECS), University of California, Berkeley.

Hearst, Marti. 1997. TextTiling: segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64.

Perelman, Chaim. 1982. *The Realm of Rhetoric* (William Kluback, translator). University of Notre Dame Press, South Bend, IN.

Swales, John. 1981. *Aspects of Article Introductions*. University of Aston, Birmingham, UK.

Swales, John. 1990. *Genre Analysis: English in Academic and Research Settings*. Cambridge University Press, Cambridge.

Marina Santini is an Honorary Research Fellow in the Humanities Advanced Technology and Information Institute (HATII), University of Glasgow. She is a computational linguist interested in genres, sentiment, and other discourse categories (also known as non-topical descriptors), as well as in Web documents, corpus building, feature extraction, and classification algorithms. Santini's address is: Varvsgatan 25, 117 29 Stockholm, Sweden; e-mail: MarinaSantini.MS@gmail.com.