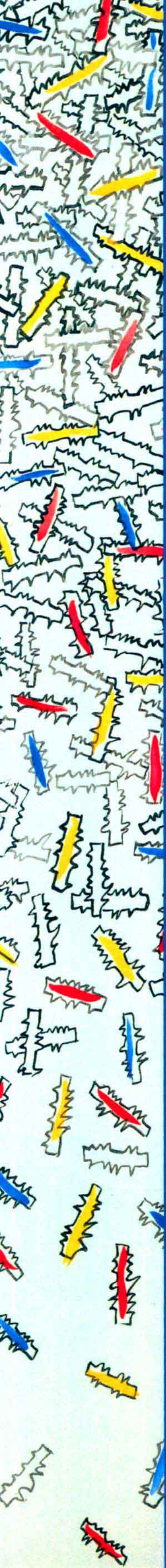

Discourse on the Move

Using corpus analysis to describe discourse structure

Douglas Biber
Ulla Connor
Thomas A. Upton

Studies in Corpus Linguistics 28

JOHN BENJAMINS PUBLISHING COMPANY

Discourse on the Move

Using corpus analysis to describe discourse structure

Douglas Biber
Northern Arizona University

Ulla Connor
Thomas A. Upton
Indiana University – Indianapolis

John Benjamins Publishing Company

Amsterdam / Philadelphia

Library of Congress Cataloging-in-Publication Data

Biber, Douglas.

Discourse on the move : using corpus analysis to describe discourse structure / Douglas Biber, Ulla Connor, Thomas A. Upton.

p. cm. (Studies in Corpus Linguistics, ISSN 1388-0373 ; v. 28)

Includes bibliographical references and index.

1. Discourse analysis--Data processing. I. Connor, Ulla, 1948- II. Upton, Thomas A. (Thomas Albin) III. Title.

P302.3.B53 2007

401'.41--dc22 2007029145

ISBN 978 90 272 2302 9 (Hb; alk. paper)

John Benjamins Publishing Co. · P.O. Box 36224 · 1020 ME Amsterdam · The Netherlands
John Benjamins North America · P.O. Box 27519 · Philadelphia PA 19118-0519 · USA

Discourse on the Move

Studies in Corpus Linguistics (SCL)

SCL focuses on the use of corpora throughout language study, the development of a quantitative approach to linguistics, the design and use of new tools for processing language texts, and the theoretical implications of a data-rich discipline.

Volume 28

Discourse on the Move. Using corpus analysis to describe discourse structure
Douglas Biber, Ulla Connor and Thomas A. Upton

Preface

The idea for this book evolved slowly, emerging from research taking place at several institutions applying different approaches to a single research problem: can discourse structure and organization be investigated from a corpus perspective?

At Northern Arizona University (NAU), research on this topic began in a PhD seminar in 1999. Inspired by the research of Youmans (1991; 1994) on the 'Vocabulary Management Profile', students in that seminar explored ways in which the discourse structure of a text can be discovered automatically by tracking the text-internal use of vocabulary and other linguistic features. This initial effort resulted in a PhD dissertation by Csomay (2002), followed by several other research studies undertaken at NAU that employed the 'TextTiling' methods originally developed by Hearst (1997).

Over the same period, researchers at Indiana University Purdue University Indianapolis (IUPUI) and Georgetown University were exploring a completely different approach to this same research problem: applying the framework of rhetorical move analysis, developed by Swales (1981; 1990) for the detailed analysis of texts, to analyze the general rhetorical and linguistic patterns of discourse structure in a corpus. At IUPUI, this research effort focused primarily on philanthropic discourse, especially grant proposals and fundraising letters. And at Georgetown University, this research culminated in 2003 with the completion of a PhD dissertation by Kanoksilapatham (2003) on the discourse structure of biochemistry research articles.

The actual idea for the present book came about as colleagues from these different institutions would get together at conferences and discuss their different approaches to the study of discourse structure and organization from a corpus perspective. We realized that there had been very little previous research done on this topic, and that by combining and comparing our approaches, we could provide a relatively comprehensive overview of this emerging subfield.

Because the book grew out of relatively independent research efforts, each author has had different primary responsibilities. At the same time, we have been eager to structure the book as a coherent treatment of this subject: an authored book rather than an edited collection of articles. Thus, the three book authors share equal responsibility for revising and editing all chapters, and ultimately the content of all chapters. But on the other hand, each chapter has different primary authors, including several co-authors in addition to the three book authors for Chapters 1–3, 5–7, and 9. Two chapters are invited, single-authored contributions – Chapter 4 by Kanoksilapatham and Chapter 8 by Csomay. The primary authors for each chapter are as follows:

Chapter 1:	Biber, Connor, Upton
Chapter 2:	Connor, Upton, Kanoksilapatham
Chapter 3:	Upton, Connor
Chapter 4:	Kanoksilapatham
Chapter 5:	Connor, Anthony, Gladkov, Upton
Chapter 6:	Biber, Csomay, Jones, Keck
Chapter 7:	Biber, Jones
Chapter 8:	Csomay
Chapter 9:	Biber, Connor, Upton

We would like to thank the numerous colleagues who have made useful suggestions and criticisms over the years in relation to the various research projects that come together in the present book. We also owe a special thanks to Eric Friginal, Bethany Gray, Jack Grieve, Mark Johnson, Erkan Karabacak, YouJin Kim, Poonpon Kornwipa, Jingjing Qin, Angkana Tongpoon, and Faith Young -- the students of ENG 707 (*Seminar on Discourse*) at Northern Arizona University in the fall of 2006, who read the entire book manuscript and made numerous useful comments and suggestions (including the title for our book, suggested by Jack Grieve).

Table of contents

Part 1. Top-down analyses of discourse organization

PART 2. Bottom-up analyses of discourse organization

CHAPTER 1

Discourse analysis and corpus linguistics

1 Discourse and discourse analysis

The study of discourse has become a major focus of research in many disciplines of the humanities, social sciences, and information sciences. Because this area of study can be approached from so many different perspectives, the terms 'discourse' and 'discourse analysis' have come to be used in widely divergent ways.

Several introductory treatments survey the range of definitions given to the term 'discourse' (e.g., Jaworski & Coupland, 1999, pp. 1–7; Schiffrin, 1994, pp. 23–43). Schiffrin, Tannen, and Hamilton (2001) in their introduction to *The Handbook of Discourse Analysis* (p. 1), group previous definitions of 'discourse analysis' into three general categories: 1) the study of language use; 2) the study of linguistic structure 'beyond the sentence'; and 3) the study of social practices and ideological assumptions that are associated with language and/or communication.

The object of study for these three approaches to discourse is increasingly removed from the research goals of traditional structural linguistics. The study of language use focuses on traditional linguistic constructs, such as phrase structures and clause structures, but addresses the problem of why languages have structural variants with nearly equivalent meanings (e.g., particle movement, as in *pick up the book* versus *pick the book up*). By considering factors that are not strictly structural, linguists are able to predict when one or another variant is likely to be used. For example, the length of the direct object noun phrase is an important factor predicting the likelihood of particle movement. Aspects of the discourse context are often important for understanding linguistic variation, especially for linguistic constructions that involve word order variation (such as passives, extraposition, clefts, inversions, existential *there*, etc.). For example, writers will choose passive voice rather than active voice depending on the topical relevance of the 'patient' noun phrase.

The study of linguistic structure 'beyond the sentence' focuses on a larger object of study: extended sequences of utterances or sentences, and how those 'texts' are constructed and organized in systematic ways. Although studies of this type are removed from the traditional concerns of structural linguistics (which focuses mostly on phrasal and clause syntax), the two share a primary focus on linguistic form and how language structures are used for communication.

In contrast, the third approach to discourse is socio-cultural in orientation, and generally not concerned with the description of particular texts or the analysis of language structure and use. Socio-cultural approaches to discourse sometimes focus on the actions of participants in particular communication events, and at other times focus on the general characteristics of speech/discourse communities in relation to issues such as power and gender. Although the socio-cultural approaches are obviously important for understanding the broader role of texts in culture, they typically are not concerned with understanding the linguistic forms used in those texts.

Corpus linguistic studies are generally considered to be a type of discourse analysis because they describe the use of linguistic forms in context. For example, words are described in terms of their typical collocates: the words that normally occur in the discourse context. Grammatical variation is also described in terms of the words and other grammatical structures that occur in the context. As such, corpus linguistic research has fallen squarely under the first approach to discourse: the study of language use.

However, it has been much less common to study discourse organization from a corpus perspective. In fact, these two subfields have research goals and methods that might be considered incompatible: The study of discourse organization – linguistic structure 'beyond the sentence' – is usually based on detailed analysis of a single text, resulting in a qualitative linguistic description of the textual organization. In contrast, corpus studies are based on analysis of all texts in a corpus, utilizing quantitative measures to identify the typical distributional patterns that occur across texts.

In fact, individual 'texts' often have no status whatsoever in corpus investigations. Instead, what we find are comparisons of the distributional patterns in one sub-corpus to the patterns in a second sub-corpus. For example, Scott and Tribble (2006) describe how we can compare the 'keywords' of the spoken versus written sub-corpora from the British National Corpus. Nesselhauf (2005, Chapter 3) describes the 'deviant collocations' in a corpus of learner English essays. And Römer (2005, Chapter 4) documents the variants and distributional patterns of progressive verb phrases in the spoken sub-corpora from the British National Corpus. These studies are typical of corpus-based research on discourse: they describe the typical patterns of language use, considering the systematic ways in which aspects of the lexico-grammatical context tend to occur together with different linguistic variants; but such corpus-based studies usually tell us nothing about the discourse structure of particular texts.

We thus see this interface as one of the current challenges of corpus linguistics: Is it possible to merge the analytical goals and methods of corpus linguistics with those of discourse analysis that focuses on the structural organization of texts? Can a corpus be analyzed to identify the general patterns of discourse organization that are used to construct texts, and can individual texts be analyzed in terms of the general patterns that result from corpus analysis? These are the central issues that we take up in the present book.

1.1 Discourse studies of language use

The first major approach to discourse identified above – the study of language use – has been carried out from several different perspectives, including research in pragmatics, speech act theory, functional linguistics, variationist studies, and register studies. These subfields all investigate how words and linguistic structures are used in discourse contexts to express a range of meanings. Many of these approaches focus on the study of linguistic variation, showing how linguistic choice is systematic and principled when considered in the larger discourse context.

There have been numerous studies of grammar and discourse over the last two decades, as researchers have come to realize that the description of grammatical function is as important as structural analysis. By studying linguistic variation in naturally occurring discourse, researchers have been able to identify systematic differences in the functional use of each variant. An early study of this type is Prince (1978), who compares the discourse functions of WH-clefts and it-clefts. Thompson and Schiffrin have carried out numerous studies in this research tradition; Thompson on detached participial clauses (1983), adverbial purpose clauses (1985), omission of the complementizer that (S. Thompson & Mulac, 1991a, 1991b), relative clauses (Fox & Thompson, 1990); and Schiffrin on verb tense (1981), causal sequences (1985b), and discourse markers (1985a, 1987). Other more recent studies of this type include Ward (1990) on VP preposing, Collins (1995) on dative alternation, and Myhill (1995; 1997) on modal verbs.

Most corpus-based research is discourse analytic in this sense, investigating systematic patterns of language use across discourse contexts, generalized over all the texts in a corpus (see, e.g., Biber, Conrad, & Reppen, 1998; McEnery, Xiao, & Tono, 2006). The advantages of a corpus approach for the study of discourse, lexis, and grammatical variation include the emphasis on the representativeness of the text sample, and the computational tools for investigating distributional patterns across discourse contexts. The recent edited volumes by Connor and Upton (2004b), Meyer and Leistyna (2003), Lindquist and Mair (2004), and Sampson and McCarthy (2004) provide good introductions to work of this type. There are also a number of book-length treatments reporting corpus-based investigations of grammar and discourse: for example, Aijmer (2002) on discourse particles, Collins (1991) on clefts, Granger (1983) on passives, Mair (1990) on infinitival complement clauses, Meyer (1992) on apposition, Römer (2005) on progressive verbs, Tottie (1991) on negation, and several books on nominal structures (e.g., de Haan, 1989; Geisler, 1995; Johansson, 1995; Varantola, 1984). The *Longman Grammar of Spoken and Written English* (1999) applies corpus-based analysis to a more comprehensive grammatical description of English, showing how any grammatical feature can be described for both structural characteristics and discourse patterns of use.

The recent book by Partington (2003) is interesting here in that it combines corpus-based study with an analysis of pragmatics, to investigate the discourse features of White House briefings. A corpus of 48 briefings (250,000 words of running texts) was subjected to computerized concordance and 'keyword' analysis. However, the computational analyses were guided by detailed qualitative analysis: "a summer reading the corpus briefings and making notes" (p. 12). This allowed Partington to check on oddities of computerized collocation analysis, highlighting odd language usage that computerized analysis might not have revealed.

A more specialized corpus-based approach to the study of language use is multi-dimensional (MD) analysis. Unlike most corpus-based research, MD studies investigate language use in individual texts. This approach describes how linguistic features co-occur in each text, resulting in more general patterns of linguistic co-occurrence that hold across all texts of a corpus. The approach can thus be used to show how patterns of linguistic features vary across individual texts, or across registers and genres. MD analysis is used in several chapters in the present book, and so it is introduced more fully in Appendix One.

1.2 Discourse studies of linguistic structure 'beyond the sentence'

The second major approach to discourse analysis identified above – the study of linguistic structure 'beyond the sentence' – is the primary focus of the present book. Previous research on discourse-level structures has been undertaken from linguistic, cognitive, and computational perspectives.

Linguistic Perspectives: Linguistic analyses of discourse structure have focused on lexico-grammatical features that indicate the organization of discourse (see, e.g., the papers in Coulthard, 1994). Focusing on units beyond the sentence-level (e.g., paragraphs in written discourse and episodes in oral discourse), these researchers investigate linguistic devices that signal the underlying discourse structure.

Much research of this type has described the discourse functions of particular words and phrases, referred to as 'discourse markers', 'connectives', 'discourse particles' (Schiffrin, 1994), 'lexical phrases' (Hansen, 1994; Nattinger & DeCarrico, 1992), or 'cue phrases' (Passonneau & Litman, 1996). Other studies discuss the linguistic devices used to mark information structure, topical development, or 'rhetorical' structures in discourse (e.g., Mann, Matthiessen, & Thompson, 1992; Mann & Thompson, 1988; Prince, 1981). Finally, some studies track the use of linguistic devices across a text. For example, discourse 'maps' are used to track verb tense and voice patterns across the sections of research articles (Biber et al., 1998, Chapter 5), while other studies track referential expressions used in anaphoric chains throughout a text (e.g., Biber, 1992; Fox, 1987; Givón, 1983).

A related area of research is the study of textual 'cohesion': the use of lexical and grammatical devices as the 'glue' of a text, holding the text together as discourse rather than an accidental sequence of sentences (see, e.g., Halliday, 1989; Halliday & Hasan, 1976; Hoey, 1991; Phillips, 1985; Tyler, 1995). Linguistic devices used to establish cohesion include anaphoric pronouns, linking adverbials, and the use of lexical repetition and synonymy to establish topical cohesion. Similarly, Tannen (1989) found that repetitions in conversation "operate as a kind of theme-setting" at the beginning of a topical unit and "at the end, forming a kind of coda" (p. 69).

Cognitive perspectives: Cognitive investigations of discourse structure study the factors that make a text 'coherent'. Text coherence refers to the linking of ideas within a text to create meaning for readers. Analyses of textual coherence typically identify the propositions expressed in a text, the logical relations among those propositions, and how listeners/readers are able to construct the overall textual meaning in terms of those propositional relations. In contrast to the study of cohesion, which refers to surface-level patterns, coherence entails the study of larger discourse relationships. Many of these studies describe texts in terms of the coherence relations expressed by clause-level propositions (Bateman & Rondhuis, 1997; Dahlgren, 1996; Hobbs, 1979; Sanders, 1997; Sanders & Noordman, 2000). Related studies also consider other factors that influence coherence, including differences between subject versus presentational matter (Mann & Thompson, 1988), text structural patterns – like problem-solution (Connor, 1987) and given-new (theme-rheme) structures (Cooper, 1988), and the semantic and pragmatic relations between units (Polanyi, 1985, 1988; Sanders, 1997). Several researchers have developed analytical frameworks for the study of coherence relations (e.g., Grosz & Sidner, 1986; McNamara & Kintsch, 1996; Tomlin, Forrest, Ming Pu, & Hee Kim, 1997; Van Dijk, 1981, 1997; Van Dijk & Kintsch, 1983).

The ongoing flow of information is also central to coherence (Grabe & Kaplan, 1996). Studies have approached information flow from various perspectives, including representations of the flow of thought (Chafe, 1994, 1997) or short-term memory (Tomlin et al., 1997).