Corpora and Discourse

The challenges of different settings

Edited by Annelie Ädel
and Randi Reppen

Studies in Corpus Linguistics 31

JOHN BENJAMINS PUBLISHING COMPANY

Corpora and Discourse

The challenges of different settings

Edited by

Annelie Ädel
University of Michigan, USA

Randi Reppen
Northern Arizona University, USA

John Benjamins Publishing Company
Amsterdam / Philadelphia

™ The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences – Permanence of Paper for Printed Library Materials, ANSI Z39.48-1984.

Library of Congress Cataloging-in-Publication Data

Corpora and discourse : the challenges of different settings / edited by Annelie Adel and Randi Reppen.
p. cm. (Studies in Corpus Linguistics, ISSN 1388-0373 ; v. 31)
Includes bibliographical references and index.
1. Discourse analysis--Data processing. 2. Corpora (Linguistics) I. Ädel, Annelie. II. Reppen, Randi.
P302.3.C6683 2008
401'.410285--dc22 2008006978
ISBN 978 90 272 2305 0 (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

Corpora and Discourse

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 31

Corpora and Discourse. The challenges of different settings
Edited by Annelie Ädel and Randi Reppen

Table of contents

The challenges of different settings

An overview

Annelie Ädel and Randi Reppen

Corpus-linguistic studies of discourse

Corpus linguistics has, over the past few decades, undergone a transformation from a "little donkey cart" to a "bandwagon" (Leech 1991:25), and is now at a point at which it "is becoming part of mainstream linguistics" (Mukherjee 2004:118). Mainstream linguistics, however, is very broad and multifaceted, and some subfields are more amenable to corpus-linguistic methodology than others.

If we disregard some basic research issues, such as access to a suitable corpus that gives a reasonably representative sample of the population studied, there are certain generalizations we can make about the compatibility of corpus-based methods with the research questions posed in different linguistic subfields. For example, while lexicographers are often able to use corpus-assisted methods in answering their particular questions about language in relatively straightforward ways, discourse analysts – whether working with speech or writing – are likely to spend a great deal of time finding possible solutions for computerizing their methods. Discourse phenomena, with their frequent dependence on and sensitivity to context, co-text, and interpretation, require rather complex solutions and often a great deal of intervention on the part of the researcher.

Despite the potential difficulties of automatizing data retrieval and analysis, researchers interested in discourse have started to adopt corpus-linguistic methods more seriously over the past few years – a trend to which the current volume bears witness. This is, however, a very recent development. At the end of the twentieth century, Biber et al. (1998:106) described the state of the art as follows: "although nearly all discourse studies are based on analysis of actual texts, they are not typically corpus-based investigations: most studies do not use quantitative methods to describe the extent to which different discourse structures are used, and relatively few of these studies aim to produce generalizable findings that hold across texts." Two other textbooks on corpus linguistics published around the same time –

McEnery & Wilson (1996) and Kennedy (1998) – both point to the comparatively marginal application of corpus-linguistic methods in discourse studies.

However, a couple of years into the new century a slightly different picture of the compatibility of computer-assisted methods with discourse-level phenomena was presented. Comparing the state of the art in 2002 to the early days of corpus linguistics, Conrad (2002:86) gives a positive characterisation, stating that, "[a]s corpus linguistics first developed, it was often thought that it could not be applied to language phenomena that extended beyond clause boundaries. As the field has matured, it has instead become apparent that many studies within corpus linguistics address discourse-level concerns, many showing association patterns or the interactions of variables that would not be apparent without corpus-based techniques."

At this point in time, we are happy to be able to say that things really are changing. For readers who wish to explore why this might be, Partington (2004) offers a summary of explanations (such as the widespread inclusion of text extracts rather than full texts in standard corpora) for the historically slight application of corpus-linguistic methods in studies of text and discourse. As a demonstration of recent shifts in this area, the present volume brings together researchers from diverse areas of text and discourse, all of whom demonstrate the viability of corpus-based research and corpus-assisted tools for discourse studies.

Finding discourse-relevant data

It is interesting to consider the search methods used by the different researchers in this volume to locate linguistic forms in a corpus – usually, in the case of discourse analysis, forms that are linked to a particular function. We believe that a description of commonly used retrieval methods can help others in reflecting on their own studies and the options available to them. Four main methods were used by the authors of these chapters, which we believe to be representative of the field.

The most typical search method can be called *one-to-one searching*, which involves investigating a linguistic form through a search term that only yields relevant hits. A good example of this is Crawford's time and place adverbs *here* and *now* in Chapter 12, where there are no spurious hits, and the entire set that the researcher intends to examine is captured. To use more technical vocabulary, precision and recall are both at 100%. The ease of capturing relevant examples, however, does not necessarily mean that no more work remains for the researcher, who will often go on to examine the different discourse functions or semantic distinctions of the search term in question.

Other search methods, however, need to be used when there is not a simple one-to-one mapping between a search term and the body of relevant hits in a corpus. To mention just a couple of complicating factors familiar to all linguists,

individual linguistic forms can be polysemous, while specific functions of language (such as politeness) can be realized by many different linguistic forms.

The second search method can be called *sampling* (Ädel 2003). It involves the use of one or more search terms that are good examples of the linguistic phenomenon in question. The drawback is that not all instances of the phenomenon, but only a subset, will be captured, although one advantage is that the search terms used tend to yield a high number of relevant hits. When using this method, the researcher cannot claim to have covered all bases or to have mapped out a linguistic function in its entirety, but many valuable insights can still be provided, especially if the search term is a good indicator of the phenomenon under study. Chapter 5 provides a good example of sampling, with Vaughan being able to draw interesting conclusions about the role of humour in the workplace based on occurrences of laughter. Vaughan uses occurrences of laughter, indicated in the transcriptions, as a "proxy" (cf. Garretson & O'Connor 2007: 89) for humour.

The third search method can be called *sifting* (Ädel 2003), since once the initial hits have been retrieved, they need to be sifted through – meaning that a certain proportion will be manually discarded. Using this method, the researcher often needs to put a great deal of time into checking the retrieved data (before the actual analysis can begin). The advantage of this method tends to be that, once the sifting has been done, the remaining set covers all or most of the potential forms of the linguistic phenomenon one is looking for. An example of this method is found in Chapter 9, where Quaglio uses an extensive inventory of linguistic forms that tend to be associated with face-to-face conversation. A small subset of these includes *so* and *really* used as informal intensifiers (but crucially, not anaphoric *so* and not *really* as a news recipient). Although this is part of a multi-dimensional analysis (Biber 1988) that both finds and interprets the co-occurrence of a selection of linguistic features, some of the forms involved can still be said to be retrieved by sifting.

The fourth and final method can be called *frequency-based listing*. It involves the use of a frequency list (of individual words or collocations), specifically based on the corpus under investigation, as a starting point. Using such a list, the researcher goes on to select the relevant search terms that occur with high frequency. This way, the search terms will be tailor-made for the corpus and the particular discourse studied. It is an effective way of using corpus-assisted methods to spot persistent patterns in a specific dataset. A nice example of this method is found in Chapter 2, where Walsh, O'Keeffe & McCarthy are able to identify exactly which expressions of vagueness to focus on based on a frequency list of multi-word clusters. Having identified the relevant expressions, they can go on to concordance and analyze them.

Of course, we live in an increasingly hybridized world, and it would probably be foolish to expect to find only pure examples of each method. Two or more of

these search methods are sometimes combined. The study by Garretson & Ädel reported in Chapter 8, for example, uses both sampling and sifting. Sampling is the overall method: by listing what they call "reporting words" (e.g. the verb lemma STATE, the noun *statement*, and the phrase *according to*), they attempt to capture instances of hearsay evidentiality in their data. Sifting is employed when individual words in the list are ambiguous or polysemous, as in the case of *states* – a highly frequent string in the US newspaper data. The analyst is required to retain examples like *the association states that misconceptions continue to affect law* and reject examples like *two dozen states that allow early voting*, either by manual elimination or through complex computational algorithms.

Any automatic or semi-automatic corpus-based method is restricted to considering surface realizations (whether actual linguistic forms, or units identified by annotation) – and herein lies the challenge for studies of functional categories. The present volume offers many interesting examples of how this challenge can be met.

Overview of the chapters

Rather than organizing the book according to the different methods researchers used for analyses, we chose as the main organizing principle the different contexts of language use. One of the main strengths of this book is its exploration of discourse in various settings, covering discourse in academia, in the workplace, in news and entertainment. Thus, the four sections of the book primarily reflect the different settings of the discourses analyzed.

The theme of the first section is "Exploring discourse in academic settings". The section begins with Walsh, O'Keeffe, and McCarthy taking a close look at the use of vague language in a range of speech events recorded at universities in the Republic of Ireland and Northern Ireland. The chapter brings to light some interesting uses of vague language and how the use of vagueness varies depending on the discourse context. The next two chapters focus on language in academic journals. First, Bondi examines stance and engagement as realized through keyword adverbs in a corpus of English-language journal articles in history and economics. A selection of the adverbials (*significantly*, *undoubtedly* and *invariably*) is studied more closely, from the perspective of collocation and patterns of semantic preference as well as pragmatic and textual functions. Next, Sanderson looks at journal articles drawn from five different disciplines in the humanities and written in German, American English and British English, focusing on the use of pronouns that mark interactivity between writer and reader. Various types of sociological information about the authors were encoded, which enabled her to check the relative influence of variables such as linguistic background, discipline, age, and gender.

The theme of the second section is "Exploring discourse in workplace settings". This section examines language in the workplace, both the contexts of business and public reports, and the context of professional meetings. The section begins with Vaughan's in-depth look at the roles humor plays in institutional interactions of teacher meetings. Using a corpus from two different settings of teacher meetings recorded in Mexico and in Ireland, Vaughan discovers interesting patterns in the use of laughter. The following two chapters explore a variety of aspects of the use of English in Hong Kong. Using a small, specialized corpus of professional reports, Flowerdew analyzes discourse moves, focusing particularly on problem-solution patterns. She also examines a couple of keywords (the lemmas *problem* and *impact*) and how they co-pattern with structural units in the texts. In the next chapter, Cheng and Warren end the second section with "a first attempt at examining the relationship between the phraseological characteristics of language and the communicative role of discourse intonation". They present an innovative investigation of patterns of discourse intonation in frequent three- and four-word combinations based on a corpus of spoken English in Hong Kong.

The theme of the third section is "Exploring discourse in news and entertainment". This section exhibits the greatest diversity of genres, including newspaper reports, a television series, and internet-based discussion boards on hip-hop. As diverse as the genres, so are the techniques used to examine discourse. In Chapter 8, Garretson and Ädel tackle the highly political issue of how hearsay evidentiality is reported in news articles related to the 2004 US presidential election. In a detailed look at how campaign language is reported and attributed, they lead the reader through unexpected insights into how different newspapers report the speech of different individual and collective entities. The next chapter takes us from the serious world of reporting presidential campaigns to a popular American situation comedy, *Friends*. In Chapter 9, Quaglio provides a detailed linguistic investigation of *Friends*, comparing it to a large corpus of natural conversation. It is a data-driven investigation which combines multidimensional methodology with a frequency-based analysis of a large number of linguistic features associated with the typical characteristics of face-to-face conversation. Quaglio indicates how the language of this television show may prove to be a resource for ESL and EFL teachers. The section concludes by moving from the language of television to the internet postings of hip-hop fans. In Chapter 10, Beers Fägersten carefully examines how identity is constructed in the virtual environment of message board postings. She guides the reader through the linguistic construction of identity – through the use of specific openings and closings, slang and taboo terms, and "verbal art" – in this highly specialized use of language.

The theme of the fourth and final section is "Exploring discourse through specific linguistic features". Johansson traces the uses of *it*-clefts diachronically. Using several corpora of diachronic and present-day English, she looks across

several different registers to reveal how the use of *it*-clefts has changed over time. The greatest frequency and the greatest number of variations on the prototypical *it*-cleft pattern are found in manuscripts from trials, where the functions of identifying and clarifying are shown to be important, especially to verify the identification of a person, thing or place. In the final chapter, Crawford analyzes the time and place adverbs *here*, *there*, *now* and *then* in three corpora of learner writing in English and compares that with corpora of English speech and writing produced by native speakers. The adverbs are analysed quantitatively and qualitatively in order to test the hypothesis that the learner writers' language use is closer to that of native-speaker speech rather than native-speaker writing.

Although the investigations represented in this book are quite narrowly focused on English, the reader will learn a great deal about different varieties of English, for example diachronic, international, learner, and non-standard varieties. Not only does this volume offer a rich sample of the spoken and written discourse around the world that takes place in English – with the interesting exceptions of references to German in Chapter 4 – but it also offers a range of topics and methods. The different approaches to the use of corpora are as diverse as the topics investigated. It is our hope that this will encourage other researchers to continue to use corpora in new ways, addressing questions in ways that were previously difficult to imagine.

References

Ädel, A. 2003. *The Use of Metadiscourse in Argumentative Writing by Advanced Learners and Native Speakers of English*. PhD dissertation, University of Göteborg.

Biber, D. 1988. *Variation across speech and writing*. Cambridge: CUP.

Biber, D., Conrad, S. & Reppen, R. 1998. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: CUP.

Conrad, S. 2002. Corpus linguistic approaches for discourse analysis. *Annual Review of Applied Linguistics* 22: 75–95.

Garretson, G. & O'Connor, M. C. 2007. Between the humanist and the modernist: Semi-automated analysis of linguistic corpora. In *Corpus Linguistics Beyond the Word: Corpus research from phrase to discourse*, E. Fitzpatrick (ed.), Amsterdam: Rodopi.

Kennedy, G. 1998. *An Introduction to Corpus Linguistics*. London: Longman.

Leech, G. 1991. The state of the art in corpus linguistics. In *English Corpus Linguistics. Studies in honour of Jan Svartvik*, K. Aijmer & B. Altenberg (eds), 8–29. London: Longman.

McEnery, T. & Wilson, A. 1996. *Corpus Linguistics*. Edinburgh: EUP.

Mukherjee, J. 2004. The state of the art in corpus linguistics: Three book-length perspectives. *English Language and Linguistics* 8(1): 103–119.

Partington, A. 2004. Corpora and discourse, a most congruous beast. In *Corpora and Discourse* [Linguistic Insights: Studies in Language and Communication 9], A. Partington, J. Morley & L. Haarman (eds), 11–20. Frankfurt: Peter Lang.

SECTION I

Exploring discourse in academic settings

'*...post-colonialism, multi-culturalism, structuralism, feminism, post-modernism and so on and so forth*'

A comparative analysis of vague category markers in academic discourse

Steve Walsh, Anne O'Keeffe and Michael McCarthy
Newcastle University, UK / Mary Immaculate College, University of Limerick, Ireland / University of Nottingham, UK

The use of vague language is one of the most common features of everyday spoken English. Speakers regularly use vague expressions to project shared knowledge (e.g., *pens, books, and that sort of thing*) as well as to make approximations (e.g. *around sevenish; he's sort of tall*). Research shows that many of the most common single word items in a core vocabulary form part of vague language fixed expressions (e.g. *thing* in *that kind of thing*). This chapter will address the use of vague language in a new corpus of academic English, the Limerick-Belfast Corpus of Academic Spoken English (LIBEL CASE). The LIBEL corpus consists of one million words of spoken data collected in two universities on the island of Ireland, one in the Republic of Ireland and one in Northern Ireland. Analysis of the LIBEL corpus identified forms and functions of vague language in an academic context and these findings are compared with two corpora of everyday spoken language from the Republic of Ireland and the United Kingdom, the Limerick Corpus of Irish English (LCIE) and the Cambridge and Nottingham Corpus of Discourse in English (CANCODE). Cross-corpora comparison allowed us to look at how forms and frequencies of certain vague language expressions vary across casual and formal/institutional contexts. Within the academic data we build on Walsh's work (see for example Walsh 2002, 2006) to show how vague language use is relative to mode of discourse at any given stage of classroom interaction. We suggest that these qualitative differences are a valuable means of understanding the complex relationship between language and learning.

1. Introduction: Vague categories

The use of vague language is one of the most common features of everyday spoken English. Speakers regularly use vague expressions to project shared knowledge (e.g., *pens, books, and that sort of thing*) as well as to make approximations (e.g. *around sevenish; he's sort of tall*). Research shows that many of the most common single word items in a core vocabulary form part of vague language fixed expressions (e.g. *thing* in *that kind of thing*). Carter and McCarthy (2002), who looked at five million words of spoken British English data, show that vague language items are among the core vocabulary items (see also O'Keeffe et al. 2007). Multi-word units which mark vagueness, such as *and things like that, that sort of thing,* occurred with greater frequency than many single word items. Degrees of variation exist in how vague language is defined. Channell (1994) restricts it to 'purposefully and unabashedly vague' uses of languages while Franken (1997) distinguishes between 'vagueness' and 'approximation'. Zhang (1998) makes a case for four separate categories: 'fuzziness', 'generality', 'vagueness' and 'ambiguity'. Chafe (1982) puts vagueness and hedging in the same category of 'fuzziness' – all of which are seen as 'involvement devices' more prevalent in spoken rather than written language. The notion of vagueness as an involvement device is consistent with the view that vague language is a core feature of the grammar of spoken language (Carter & McCarthy 1995, 2006; McCarthy & Carter 1995; O'Keeffe et al. 2007). As Carter and McCarthy (2006) note, vague language is a strong indicator of assumed shared knowledge which marks in-group membership insofar as the referents of vague expressions can be assumed to be known by the listener. This is consistent with Cutting (2000), who illustrates how discourse communities use vague language as a marker of in-group membership. The interactive aspect of vague language is important to our focus in this chapter where we examine the use of vague language in the learning context of university discourse. In this domain, the use of vague language is part of meaning making within specific learning contexts or *modes* (see Walsh 2006: 111).

We will focus on one type of vague language, namely vague category markers (hereafter VCMs). These non-lexicalised categories are created within interactions, at the moment of speaking. The categories contain exemplars followed by a vagueness tag (*and so on, and that kind of thing, et cetera, and things like that)* and the listener(s) is/are expected and assumed to fill in, or implicitly understand the reference. The example in Extract 1 is taken from a drama lecture in the Limerick Belfast Corpus of Academic Spoken English (LIBEL CASE;[1] see details in Sections 3 and 4):

1. Hereafter, LIBEL CASE will be shortened to LIBEL.