Annual Review of Applied Linguistics (2002) **22**, 75–95. Printed in the USA.

4. CORPUS LINGUISTIC APPROACHES FOR DISCOURSE ANALYSIS

Susan Conrad

This chapter provides an overview of approaches within corpus linguistics that address discourse-level phenomena. The shared characteristics of all corpus-based research are first reviewed. Then four major approaches are covered: (1) investigating characteristics associated with the use of a language feature, for example, analyzing the factors that affect the omission or retention of *that* in complement clauses; (2) examining the realizations of a particular function of language, such as describing all the constructions used in English to express stance; (3) characterizing a variety of language, for example, conducting a multi-dimensional analysis to investigate relationships among the registers used in different settings at universities; and (4) mapping the occurrences of a feature through entire texts, for example, tracing how writers refer to themselves and their audience as they construct authority in memos. For each approach, a variety of studies are reviewed to illustrate the diverse perspectives that corpus linguistics can bring to our understanding of discourse. The chapter concludes with a brief overview of some other foci in corpus linguistics and suggests that two areas require particular attention for the advancement of discourse-oriented corpus studies: the need for more computer tools and computer programmers for corpus linguistics, and the need for further studies about how best to represent language varieties in a corpus.

Corpus linguistics encompasses a great variety of approaches for studying language use. For many readers, concordance listings may be the most familiar form of corpus linguistics. These listings display all the occurrences of a word or structure in a database, with a small amount of context on each side. Overall, they appear quite divorced from a situation of use, and may therefore give the impression that corpus linguistics has little to offer discourse analysis. However, concordance listings represent only a small piece of the work that goes on in corpus linguistics. Full corpus-based studies provide complex information about social and textual factors that influence language choices, and therefore can contribute greatly to our understanding of discourse.

In this chapter, I provide an overview of approaches within corpus linguistics that are especially applicable to discourse analysis. I begin with an introduction to the features that characterize all corpus linguistic work, and then focus on four approaches:

- investigating the characteristics associated with the use of a language feature; for example, what factors are associated with a speaker's use of past perfect ("I'd done a lot") rather than past tense ("I did a lot")?

- examining the realizations of a particular function of language, for example, describing all the constructions in English that are used to express stance.

- characterizing a variety of language, such as describing the similarities and differences between casual conversations and academic writing.

- mapping the occurrence of language features through a text, for example, tracking the terms that writers use to refer to themselves and their audience.

Within each approach, I briefly review studies to illustrate the diverse research that has been conducted. I conclude by mentioning some additional foci in corpus linguistics, and briefly consider the future challenges for corpus linguistics in discourse analysis. Throughout, I emphasize work that has been conducted since the late 1990s; useful earlier work and bibliographies can be found in many sources including Biber, Conrad, and Reppen (1996); McEnery and Wilson (1996); Svartvik (1992); and Thomas and Short (1996). I also give most coverage to studies of English, since the majority of corpus work has been done with English; however, the approaches are equally applicable to other languages.

Characteristics of Research in Corpus Linguistics

Four features, described briefly below, characterize work within corpus linguistics (for more details, see introductory corpus linguistics books such as Biber, Conrad and Reppen, 1998; Kennedy, 1998; and, on statistics in corpus linguistics, Oakes, 1998).

Use of a Corpus

A corpus is a large, principled collection of naturally-occurring texts that is stored in electronic form (accessible on computer). Corpora can include both written and transcribed spoken texts.

Corpus design is crucial to reliable and generalizable results. Although a full discussion of issues in corpus design is beyond the scope of this paper, it is important to note that the size of the corpus, the types of texts included, the number of texts, the sampling procedure, and the size of each sample are all important considerations. In general, corpora are designed following principles for representing demographic characteristics and recognized types of texts (e.g., see Aston & Burnard, 1998; Biber, Johansson, Leech, Conrad, & Finegan, 1999, chapter 1; Granger, 1998, chapter 1; Hennoste, Koit, Roosmaa, & Saluveer, 1998; McCarthy, 1998; and the websites listed in the annotated bibliography). Advances in computer technology have made increasingly large corpora possible, but there has unfortunately been relatively little empirical investigation of the size and sampling that are reliable yet efficient for representing all the variation in a language. In the early 1990s, Biber (1990, 1993) found 1,000 word samples reliable for representing many grammatical features and 10 texts reliable for representing the genre categories in the Lancaster-Oslo/Bergen (LOB) Corpus (e.g., press reportage, official documents, academic prose), and also called for the representation of empirically determined "text types." However, there has been little work published recently that empirically investigates corpus design issues. (One exception is Kilgarriff, in press, on design issues with reference to word frequencies.)

Use of Computer-assisted Analysis Techniques

Corpus linguistics relies on computer-assisted techniques in order to handle the large amount of data in a corpus. Early publications often emphasized concordancing (e.g., Sinclair, 1991). In addition to showing words in context, concordancers calculate frequencies, analyze collocates (words that occur together) and often calculate statistical measures of the strength of word associations. However, much work in discourse analysis requires other kinds of analyses. With knowledge of computer programming, researchers can write specialized programs to analyze more complex aspects of texts, as described in subsequent sections. Such programs can be interactive, asking the researcher to make judgments about ambiguous forms as they are identified and coded. For an example of an interactive program, see Biber, Conrad, and Reppen, 1998, chapter 5.

Emphasis on Empirical Analysis of Patterns in Language Use

Corpus linguistic studies often develop from research questions that grow out of intuition or casual observations about language, and interpretations of corpus findings often include intuitive impressions about the impact of particular language choices. Nevertheless, the primary focus of analysis is empirical, based on what is observed in the corpus. Researchers are concerned with the patterns in language, determining what is typical and unusual in given circumstances. As McCarthy (1998) states, "The particular strength of computerised corpora is that they offer

the researcher the potential to check whether something observed in everyday language is a one-off occurrence or a feature that is widespread across a broad sample of speakers" (p. 151).

Use of Quantitative and Qualitative/Interpretive Techniques

Some corpus studies emphasize either the quantitative or the qualitative aspects of analysis. A recent issue of *TESOL Quarterly* highlights this contrast in methodologies (Biber & Conrad, 2001; McCarthy & Carter, 2001). However, all studies include both aspects of analysis to some extent. Recognizing patterns of language use necessarily entails assessing whether a phenomenon is common or unusual—a quantitative assessment. At the same time, numbers alone give little insight about language. Even the most sophisticated quantitative analyses must be tied to functional interpretations of the language patterns.

With these four characteristics, corpus linguistics has certain analytical strengths. Primary among them is the capacity for analyzing many more variables and data than were previously possible. Corpus linguistics is thus particularly helpful in providing "big picture" perspectives on discourse—determining patterns of language behavior across many texts, identifying typical and unusual choices by users, and describing the interactions among multiple variables. It provides a complementary perspective to more intensive approaches to discourse analysis, such as in conversation analysis (Schegloff, Koshik, Jacoby, & Olsher, this volume).

Four Approaches in Corpus-Based Research

Investigating Characteristics Associated with the Use of a Language Feature

One of the most common approaches taken in corpus studies is to focus on a particular language feature—a word, phrase or grammatical structure—and investigate factors associated with its use. Such investigations offer insight into the factors that shape the choices that language users make for different discourse conditions.

For an example of this type of approach, consider the choice between omission and retention of the complementizer *that* in clausal complements to verbs and adjectives in English, such as

I think [that] he might have gotten false teeth. Or
It was real clear [that] he did that.[1]

Native and nonnative speakers alike usually recognize that, as one popular ESL grammar book puts it, "Frequently [*that*] is omitted, as in [the example], especially

in speaking" (Azar, 1999, p. 248). However, such a statement leaves many questions unanswered: To what extent is *that* omitted in conversation and in other registers? If omission is common, why is *that* ever retained?

Analyzing about 20 million words in the Longman Spoken and Written English Corpus,[2] Biber et al. (1999) show how common *that* omission is across registers and under what conditions *that* is retained. (Results reported here are summarized in slightly different form from analyses conducted for Biber et al., 1999, chapter 9.) The study finds that the great majority of the complement clauses in conversation do omit *that*. The percentage of *that* omission is striking when compared with three written registers, even a nonexpository written register such as fiction, which often includes dialogue and other informal language (see Figure 1).

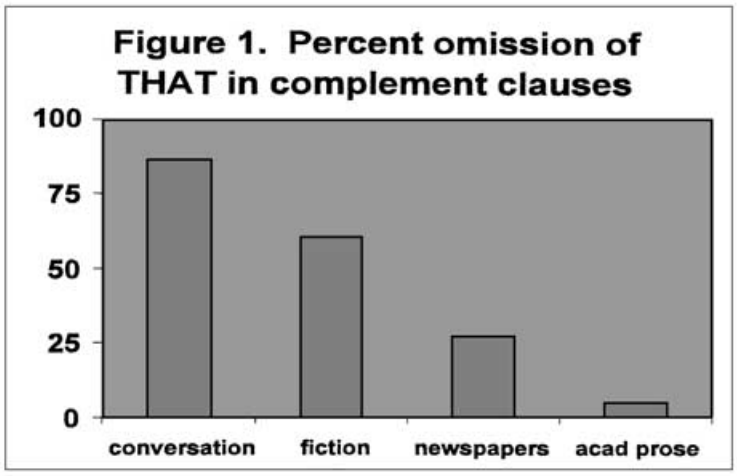

Figure 1: Percent omission of THAT in complement clauses.

The study then goes on to analyze conditions associated with the retention of *that* in conversation. Three conditions are particularly noteworthy:

- virtually all coordinated *that*-clauses retain *that*, for example:

And he said that I looked very good on paper, but that during the interview they felt. . .

- virtually all *that*-clauses that occur with intervening noun phrases retain *that*, for example:

I think I'll tell him that it's not a good idea right now.

- about 75% of the clauses that have passive voice main verbs retain *that*, for example:

I was told that it was too expensive.

The full study compares the influence of these conditions across registers and covers many other aspects of *that*-clause complementation. Yet even this brief piece of the study contributes to our understanding of the characteristics of conversation. Though observers have noticed that *that* is often omitted, the corpus-based study, with over 3,000 conversations included, shows how widespread *that* omission is and how strong a relationship exists between particular conditions and retention of *that*.

In other corpus studies, grammatical features have also been associated with discourse management functions, such as the foregrounding and backgrounding of information. For example, Hughes and McCarthy (1998) analyze the use of past perfect verb forms in cases where other choices of tense or aspect are possible, for example:

> Well yeah I mean Christmas was really for us this time. I mean we'd done a lot of pre-planning for it. . . (Hughes & McCarthy, 1998, p. 269).

With evidence from informal conversation in one million words of the Cambridge and Nottingham Corpus of Discourse in English (CANCODE; see McCarthy, 1998), they argue that, across a wide range of speakers, the past perfect has a broader and more complex function than is often maintained—i.e., to show that events happened in a past time before another past time. They find that speakers use it in conveying relationships among narrated events, manipulating foregrounding and backgrounding, often as part of explanations and justifications. In other corpus studies, Breivik (1999) finds that factors such as end focus and the grounding of main clauses have important associations with the use of relative clauses in existential sentences.

The expression of ideologies and attitudes have also been addressed with corpus-based techniques, especially in studies of word associations. Partington (1998) uses corpus analyses to argue that attitude is created in a text through the use of semantic prosody—not the connotations of a word, but the word's association with positive or negative contexts (e.g., the verb *commit* is usually used with negative objects such as *foul*, *crime*, *suicide*). Stubbs (1996) uses corpus-based techniques to analyze historical speeches to the Boy Scouts and Girl Guides. Looking at word collocations and grammatical structures, he shows how the speaker's language choices reflect his beliefs about men's and women's roles in society. Even the occurrences of a word such as *happy* are found to be intricately

tied to ideology and sexism. Such a study also demonstrates the applicability of corpus linguistics techniques to critical discourse analysis (Luke, this volume).

Corpus linguistic techniques have also been important for discerning the strong associations that exist between the lexicon and grammatical structures. For example, Biber et al. (1999, chapter 9) also analyze the lexico-grammar of verb complement clauses, comparing different registers. In conversation, *that*-clauses are strongly associated with the verbs *think*, *say* and *know*, corresponding to the typical functions in conversation of reporting the thoughts and ideas of oneself or others (e.g., *I think you need a rest*). In academic prose, on the other hand, there are stronger associations between *that*-clauses and the verbs *show* and *suggest*. These constructions are often used with inanimate subjects, expressing an argument without a personal agent (e.g., *Other models show that...*). Hunston and Francis (2000, especially chapter 8) argue the importance of lexico-grammatical patterns further, based on analyses of the COBUILD corpus (see Sinclair, 1987). They argue that the senses of a word are associated with particular patterns, that discourse is made up of sequences of these patterns, and that different patterns are associated with different genres.

The types of characteristics included in corpus-based studies of language features cover a wide gamut. In addition to variation across registers, studies have included subregisters; for example, Ferguson (2001) analyzes how conditionals are used differently across different types of medical discourse—research articles vs. journal editorials vs. doctor-patient consultations. Other studies have focused on personal characteristics of speakers: Stenström (1999), for instance, considers associations between gender and the choice of intensifiers in a corpus of London teenager language (e.g., *really bad day* vs. *bloody right cow*). Other studies primarily focus on variation between national dialects, such as British and American varieties of English; for example, Leech (1999) compares the use of vocatives. Still others consider change over time as the variable of interest (e.g., see Hundt & Mair, 1999, on the historical development of "agile" vs. "uptight" prose). Of course, many studies include more than one characteristic; Botley and McEnery (2001) cover associations with six characteristics in their study of demonstrative pronoun use. Furthermore, all of these characteristics can be studied in other languages besides English (e.g., Butler, 1998, investigates collocations across spoken and written Spanish; Verhagen, 2000, examines language change in verb use in Dutch).

Examining the Realizations of a Particular Function of Language

In the previous section, studies focused on individual language features, such as words or grammatical constructions. A second approach within corpus linguistics is to focus on a function of language and to determine how it is realized in discourse. Thus, within the investigation of stance, a study of an individual

language feature focuses on just stance adverbials (Conrad & Biber, 1999); in contrast, Hunston & Sinclair (1999) address the entire functional system: They use corpus analyses to investigate all the lexico-grammatical patterns whose primary purpose it is to express evaluation, finding patterns that begin with verbs, dummy subject *there*, pseudoclefts, and nouns.

The advantage of a corpus linguistic approach in studying a function of language is that many interacting characteristics can be examined simultaneously. Biber, Conrad, and Reppen (1998, chapter 5) illustrate this advantage in a study of reference. Six characteristics are included in the study: register (conversation, speeches, newspaper writing, academic prose); pronoun vs. noun forms; given vs. new information status; type of reference (anaphoric, exophoric, or inferable); type of expression for anaphoric reference (synonyms vs. full nouns); and the distance between the referring expression and its antecedent. In the results, there are clear relationships among the characteristics. For example, register, average distance, type of referring expression and given/new information status are interconnected: newspaper writing and academic prose are found to have many full-noun referring expressions that present new information, with a large number of expressions intervening between coreferential nouns. Conversation, on the other hand, shows a preference for pronominal reference, with repeated reference to already-mentioned (given) entities, and as a result, has a much lower average distance between the referring expression and its antecedent. Even with the assistance of a computer, such a study can be very time-consuming because disambiguating references requires a good deal of interactive computer coding before the analysis can take place—but by hand, keeping track of the numerous variables in numerous texts would soon become completely unmanageable.

Corpus linguistics has also been used for the analysis of themes in texts. This area of functional-systemic linguistics has traditionally been studied with intensive analysis of a small number of texts. In contrast, Gómez-González (1998) uses corpus linguistic techniques to include over 4,000 themes in her study. She uses multivariate techniques to investigate the typical ordering of parts within 1,660 extended multiple themes (e.g., those that contain textual or interpersonal items as well as topical themes, e.g., *but of course I*...). Green, Christopher, and Kam Mei (2000) apply corpus-based theme analysis to ESL student writing. They conduct a study of theme choices in native speaker and Chinese-ESL writers' academic texts, showing that differences—and especially the greater frequency of connectors as theme choice (e.g., *besides*, *furthermore*)—have a negative effect on the coherence of the ESL writers' texts.

Anping and Kennedy (1999) demonstrate that corpus linguistics techniques can be applied even to an interactive discourse function: turn-bidding in conversation. They examine the linguistic devices (e.g., initial words) and pragmatic devices (e.g., types of repetition) that speakers use when they

successfully bid for a turn, as well as the prosodic, grammatical, and lexical markers of turn-unit boundaries. Their study also looks at associations with the formality of the situation, the familiarity of the interlocutors, and the speaker's status and gender. With 193 speakers in 83 conversations, Anping and Kennedy have a much wider base from which to draw conclusions than in most previous studies of turn-taking.

Characterizing a Variety of Language

Many studies mentioned in the prior two sections have included register or dialect as a variable. However, a third approach within corpus linguistics makes the language variety the primary focus; these studies have as a goal the characterization of a variety of discourse.

One area in which corpus linguistics has proven useful is in characterizing the vocabulary of a specific domain. The desire to identify an English academic vocabulary—and thereby improve the effectiveness of instruction in English for Academic Purposes—has received attention for many decades (e.g., Ghadessy, 1979; Xue & Nation, 1984). Without corpus linguistics techniques, however, the amount of language that could be included was quite limited. For example, Ghadessy (1979) developed a word list using 20 textbooks from three disciplines, with a total of slightly under 500,000 words. With corpus-based techniques, Coxhead (2001) bases an academic word list on a corpus that includes over 400 texts and 28 subject areas with a total of about 3.5 million words. Because Coxhead's goal is a word list representing widespread academic use, the broader coverage of texts and subject areas is a great advantage. (Smaller corpora have also been used to characterize the vocabulary or lexico-grammatical patterns of specialized domains; for example, see Williams, 1998, on plant biology and Gledhill, 2000, on cancer research articles.)

Within corpus linguistics, one methodology was developed specifically for characterizing language varieties: multi-dimensional (MD) analysis, developed by Biber (1988). This method uses factor analysis to analyze the co-occurrence patterns of numerous linguistic features; for example, over 60 linguistic features have been used in the analysis of English, including grammatical, lexical, and some semantic features. The analysis quantitatively determines several continua along which texts vary, and these are then given functional interpretations, such as the expression of overt argumentation or narrative vs. nonnarrative concerns.

One of the strengths of the MD methodology is that it allows complex comparisons to be made among language varieties. To take just two registers and two dimensions, consider conversation and academic prose. As expected, they are very different when compared in their use of features related to involvement and real-time processing (e.g., first and second person pronouns, main verb *be*,

contractions, etc.) vs. features of densely packed information (nouns, prepositions, attributive adjectives). However, they are also similar in the frequency of use of features of overt argumentation; neither has a particularly high frequency as newspaper editorials do, nor a particular absence, as broadcasts do. Analyses can also be extended to investigate the amount of variability across texts within a variety, and to compare more specialized subvarieties. (See Conrad & Biber, 2001, for a more complete description of the MD methodology and relationships among registers in English.)

The MD methodology has proved a useful method of analysis for diverse contexts. For example, a recent collection of MD studies covers among other topics: the historical development of scientific discourse (Atkinson, 2001); oral proficiency test validation (Connor-Linton & Shohamy, 2001); the discourse of different academic disciplines (Conrad, 2001); registers of children's speech and writing (Reppen, 2001); gendered language use in *Star Trek* (Rey, 2001); and comparisons of British and American spoken registers (Helt, 2001). In addition, Biber, Conrad, Reppen, Byrd, & Helt (in press) use MD analysis to compare several spoken and written registers used at American universities, and Hoogesteger (cited in de Haan, 1999) adapts the technique in a study of English writing by Dutch students.

Finally, for characterizing a language variety, it is also possible to consider the body of work that has analyzed a specialized corpus, even if independent studies have focused on individual features. For example, spoken academic language is the focus of studies using the Michigan Corpus of Academic Spoken English (e.g., Mauranen, 2001; Swales, 2001; Swales & Malczewski, 2001). Results from this corpus can also be compared with results of studies from the spoken section of the TOEFL 2000 Spoken and Written Academic Language Corpus (Biber, Reppen, Clark, & Walter, 2001).

Mapping the Occurrence of a Language Feature through a Text

One additional approach within corpus linguistics is particularly applicable to discourse analysis, though it thus far has received little attention. In this approach, one or more features are tracked through an entire text to determine how the features contribute to some aspect of the discourse development, such as its rhetorical organization, topic progression, or the author's construction of authority. A "map" of the feature through the text is produced, giving a visual representation of its use. Multiple texts are then compared to determine consistent patterns.

Working with institutional memos, Burges (1996) maps writers' reference to participants (e.g., *I, you, faculty*), comparing the patterns for memos written to groups with superior, inferior, and equal hierarchical standing. She finds that the selection of the noun or pronoun and the level of prominence (theme or rheme

positioning) give insight into how writers manipulate their writing to construct their authority for writing the memo. A similar technique is exemplified in Biber, Conrad, and Reppen (1998, chapter 5) for the analysis of verb tense and voice throughout science research articles. Areas with numerous shifts—e.g., when occurrences of verbs alternative between active and passive, and past and present—correspond to transition zones that are of particular rhetorical interest. Intensive analysis of the communicative purposes of these transition zones is then possible.

Another mapping technique has been used to track the introduction of new vocabulary into a text, producing a visual display of "vocabulary management profiles" (Youmans, 1991). Csomay (2000) adapts this technique, showing how the profiles of vocabulary correspond to topical and functional structure within oral academic discourse in university classrooms.

Though few studies have used a mapping approach, its potential to contribute to our understanding of discourse is great. In the past, it was usually too time-consuming to track a feature through more than a few complete texts. With corpus linguistics, once a program is written, and especially if it can be run without interactive analyses, it is a simple matter to use it on numerous texts. Studies of patterns in discourse thus have a much firmer basis for generalization than in the past.

Additional Foci Within Corpus Linguistics

In a chapter of this size it is impossible to cover all facets of corpus linguistics, and several other areas may be of interest to discourse analysts. In particular, I have neglected much work with cross-linguistic comparisons, parallel corpora (designed to represent the same varieties in two languages), and the application of corpus linguistics to translation (see, e.g., Botley & Wilson, 2000; Hasselgård & Oksefjell, 1999, section 3; Johansson, 1997; Partington, 1998, chapter 3). Corpus-based studies of recurring word sequences also deserve mention since they have implications for discourse production and processing. Researchers are finding that fixed sequences account for a notable percentage of discourse, and are arguing that these prefabricated units appear to be basic building blocks of discourse (e.g., Barlow, 2000; Biber & Conrad, 1999; McCarthy, 1998). Although their importance was noted in the past, only corpus-based studies have provided quantitative support to show just how common these recurring sequences are over a wide range of texts.

Several studies mentioned above have included comparisons of native speakers with second language learners, but many more such studies of second language learners exist (e.g., de Cock, 1998, and the studies collected in Granger, 1998). Studies of language change also cover diverse topics (see, e.g., the papers

in Curzan & Meyer, 2000). Recently, a good deal of attention has also been paid to profitable connections between corpus linguistics and natural language processing, machine translation, and lexical knowledge databases (e.g., Dini & Di Tomaso, 1998; Hoard, 1998; and the collection edited by Nerbonne, 1998).

The role of corpus linguistics in language teaching has become an increasingly popular topic. Several recent publications discuss the role of corpora in classroom pedagogy (see, e.g., Fox, 1998; Gavioli & Aston, 2001; Granger & Tribble, 1998; Thurston & Candlin, 1998; and numerous papers collected in Aston, 2001), and the role corpus-based research should play for developing, adapting, and assessing pedagogical materials (e.g., Conrad, 1999; Kaszubski, 1998; McCarthy, 1998; for a less positive orientation, see Owen, 1996). Within grammar pedagogy, corpus linguists have argued that their work demonstrates the need for new perspectives on grammar, especially related to stronger discourse and lexico-grammar orientations, register comparisons, and considerations of grammar in terms of probabilities and appropriate choices, rather than deterministic rules and notions of correctness (see, e.g., discussions in Biber & Conrad, 2001; Conrad, 2000; Hughes & McCarthy, 1998; Hunston & Francis, 1998; McCarthy, 1998).

Issues concerning the annotation of corpora—such as identifying the grammatical classes of words or functional categories of expressions—are beyond the scope of this chapter. Interested readers will find useful information in Aarts and Oostdyk (1997); Assi and Abdlhosseini (2000, for experience with a language other than English); Hockey (1998); and Meunier (1998, for work with a learner corpus). Powell and Simpson (2001) provide information about developing a web interface for public access to a corpus.

The Future for Corpus Linguistics and Discourse Analysis

As 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.

In the past several years, corpora have become increasingly numerous and accessible (see annotated bibliography) and corpus-based studies have begun to appear more regularly in mainstream venues. It thus seems likely that corpus linguistics will continue to expand. Nevertheless, I believe there are two major challenges that will affect the popularity and acceptance of corpus-based discourse studies in the future.

The first challenge concerns the availability of computer tools that allow discourse-level studies of corpora. Currently, most researchers have, at most, access to a concordancer, which allows only limited investigation of discourse-level features. More software, more adaptable to discourse concerns, needs to become available to more researchers. More computer programming classes specifically for corpus linguistics need to be offered so that researchers can write their own programs.

The second challenge concerns corpus design. Corpora are meant to capture the variation that exists in a language, and our present corpora are clearly capturing a great deal. However, as mentioned above, empirical evidence about how best to represent all language variation is scanty. Further investigations into corpus sizes and sampling techniques are needed, as well as further research into the kinds of variation that exist in language so that we can make sure to capture all kinds of variation in new corpora. In fact, corpus design itself is interwoven with the process of discourse analysis. A better understanding of discourse leads to improvements in corpus design, and research into corpus design increases our understanding of variation in discourse.

Notes

1. Samples of *that*-clauses have been taken from the Longman Spoken American Corpus.

2. The brevity of this chapter and its purpose—to introduce the discourse research approaches within corpus linguistics—prohibit a full description of the corpora used in each study that I mention. Details can be found in the publications cited.

ANNOTATED BIBLIOGRAPHY

Anping, H., & Kennedy, G. (1999). Successful turn-bidding in English conversation. *International Journal of Corpus Linguistics, 4*, 1–27.

This article is particularly noteworthy because it applies corpus linguistic techniques to a complex area of interaction—turn-taking. The study not only considers the linguistic and pragmatic devices that turn-bidders use, but also examines the influence of different speech domains, different participant relationships, social status, and gender. The preparation of the corpus and analytical techniques for such a study are clearly described.

Conrad, S., & Biber, S. (Eds.). (2001). *Variation in English: Multi-dimensional studies*. Harlow, Essex: Longman.

This collection begins with an explanation of the technique of Multi-Dimensional Analysis and the model of variation in English developed by Biber (1988), written for readers with no previous background in the methodology. The book then presents diverse studies applying the model in new areas, including the historical evolution of registers, applications to specialized domains and dialect variation. Topics are as diverse as American/Soviet nuclear arms talk, the discourse of different academic disciplines and the language of *Star Trek*. The final section of the book presents three studies that develop new Multi-Dimensional models—for student speech and writing, 18th century discourse, and discourse complexity.

Granger, S. (Ed.). (1998). *Learner English on computer*. London: Longman.

In this edited collection, a number of studies using the International Corpus of Learner English and other learner corpora are presented. Studies cover grammar, lexis, and discourse concerns, and include sub-corpora from learners in a wide variety of (mostly European) countries. The studies present numerous interesting findings about learner language, yet the book is also useful as a resource for the compiling and use of learner corpora generally, with several articles about design and pedagogical applications.

McCarthy, M. (1998). *Spoken language and applied linguistics*. Cambridge: Cambridge University Press.

This book brings together revised versions of papers published by McCarthy over the late 1980s and 1990s, covering much of his work with the CANCODE (Cambridge and Nottingham Corpus of Discourse in English). The first chapter of the book describes the CANCODE project, and subsequent chapters discuss findings about informal spoken English, particularly features which have been overlooked in traditional reference materials and textbooks. McCarthy presents many arguments for the positive effects that corpus linguistics can have on language pedagogy, as well as acknowledging concerns such as the teachability and learnability of features and the appropriate use of corpora in different learner contexts.

Partington, A. (1998). *Patterns and meanings: Using corpora for English language research and teaching*. Amsterdam: John Benjamins.

Each chapter of this book presents a different area of interest within corpus linguistics. Some of these are topics commonly covered, such as lexical studies and collocations, translation, and syntax, but the book is notable for its inclusion of less commonly discussed areas as well, including

chapters on metaphor and creative language use. The explanations are accessible for beginners in corpus linguistics, and many teaching applications are exemplified.

Websites

For researchers with computer access, one of the easiest ways to learn about corpora and try some basic concordancing is online. The following is a very short list of sites to illustrate corpora representing varieties of English. The sites are chosen because they clearly describe their design criteria, provide search software on the site, and cover corpora that are available for public (noncommercial) use. Numerous other useful sites also exist.

http://titania.cobuild.collins.co.uk/boe_info.html—The Bank of English

The Collins-COBUILD project has been growing since the 1980s. Currently, the COBUILD corpus is over 400 million words, and continues to have new texts added. The website describes the corpus and its use for language description and learning. Sample concordance and collocation analyses are easy to conduct from the website.

http://info.ox.ac.uk/bnc/—The British National Corpus

The BNC is a 100 million word corpus of spoken and written British English from a variety of sources. The website offers simple concordancing searches, as well as information on obtaining more sophisticated analysis programs and all or part of the corpus.

http://americannationalcorpus.org—The American National Corpus

The American National Corpus project is currently creating a corpus comparable to the British National Corpus, with about 100 million words representing the same registers. This corpus will be distributed free for noncommercial research purposes. The site describes the design and current status of the project.

http://www.ucl.ac.uk/english-usage/ice/—The International Corpus of English

Approximately 20 varieties of English from around the world will be represented in this corpus. The comparative study of Englishes will be facilitated by the common design for each sub-corpus. Parts of the ICE are now available, as is a demonstration of software to use with the corpus.

http://www.lsa.umich.edu/eli/micase/micase.htm—The Michigan Corpus of Academic Spoken English

The current phrase of the MICASE is near completion, with about 1.5 million words of text recorded in academic contexts across the University of Michigan. Almost 1 million words is available on the web, along with easy-to-use concordancing software that allows users to search the corpus using a variety of specifications.

Additional useful sites:

http://www.ruf.rice.edu/~barlow/corpus.html

Michael Barlow's website about corpus linguistics offers many useful links to other sites, including corpora in over 20 languages.

http://www.ldc.upenn.edu and http://www.hit.uib.no/icame.html

Information and many corpora are also available through these websites of the Linguistic Data Consortium and ICAME (International Computer Archive of Modern and Medieval English)

OTHER REFERENCES

Aarts, J., & Oostdyk, N. (1997). Handling discourse elements in syntax. In U. Fries, V. Müller, & P. Schneider (Eds.), *From Ælfric to the New York Times: Studies in English corpus linguistics* (pp. 107–124). Amsterdam: Rodopi.

Assi, S. M., & Abdlhosseini, M. H. (2000). Grammatical tagging of a Persian corpus. *International Journal of Corpus Linguistics, 5*, 69–81.

Aston, G. (Ed.). (2001). *Learning with corpora*. Houston, TX: Athelstan.

Aston, G., & Burnard, L. (1998). *The BNC handbook: Exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press.

Atkinson, D. (2001). Scientific discourse across history: A combined multi-dimensional/rhetorical analysis of *The Philosophical Transactions of the Royal Society of London*. In S. Conrad & D. Biber (Eds.), *Variation in English: Multi-dimensional studies* (pp. 45–65). Harlow, Essex: Longman.

Azar, B. (1999). *Understanding and using English grammar* (3rd ed.). White Plains, NY: Longman.

Barlow, M. (2000). Usage, blends and grammar. In M. Barlow & S. Kemmer (Eds.), *Usage-based models of language* (pp. 315–346). Stanford, CA: Center for the Study of Language and Information.

Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, D. (1990). Methodological issues regarding corpus-based analyses of linguistic variation. *Literary and Linguistic Computing, 5*, 257–269.

Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing, 8*, 243–257.

Biber, D., & Conrad, S. (1999). Lexical bundles in conversation and academic prose. In H. Hasselgärd & S. Oksefjell (Eds.), *Out of corpora: Studies in honour of Stig Johansson* (pp. 182–190). Amsterdam: Rodopi.

Biber, D., & Conrad, S. (2001). Quantitative corpus-based research: Much more than bean counting. *TESOL Quarterly, 35*, 331–336.

Biber, D., Conrad, S., & Reppen, R. (1996). Corpus-based investigations of language use. *Annual Review of Applied Linguistics, 16*, 115–136.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.

Biber, D., Conrad, S., Reppen, R., Byrd, P., & Helt, M. (in press). Speaking and writing in the university: A multi-dimensional comparison. *TESOL Quarterly*.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *The Longman grammar of spoken and written English*. Harlow, Essex: Pearson Education.

Biber, D., Reppen, R., Clark, V., & Walter, J. (2001). Representing spoken language in university settings: The design and construction of the spoken component of the T2K-SWAL Corpus. In R. Simpson & J. Swales (Eds.), *Corpus linguistics in North America: Selections from the 1999 Symposium* (pp. 48–57). Ann Arbor, MI: University of Michigan Press.

Botley, S., & McEnery, T. (2001). Demonstratives in English: A corpus-based study. *Journal of English Linguistics, 29*, 7–33.

Botley, S., & Wilson, A. (Eds.). (2000). *Multilingual corpora in teaching and research*. Atlanta, GA: Rodopi.

Breivik, L. E. (1999). On the pragmatic function of relative clauses and locative expressions in existential sentences in the LOB corpus. In H. Hasselgärd & S. Oksefjell (Eds.), *Out of corpora: Studies in honour of Stig Johansson* (pp. 121–135). Amsterdam: Rodopi.

Burges, J. (1996). Hierarchical influences on language use in memos. Unpublished doctoral dissertation. Flagstaff, AZ: Northern Arizona University.

Butler, C. (1998). Collocational frameworks in Spanish. *International Journal of Corpus Linguistics, 3*, 1–32.

Connor-Linton, J., & Shohamy, E. (2001). Register variation, oral proficiency sampling, and the promise of multi-dimensional analysis. In S. Conrad & D. Biber (Eds.), *Variation in English: Multi-dimensional studies* (pp. 124–137). Harlow, Essex: Longman.

Conrad, S. (1999). The importance of corpus-based research for language teachers. *System, 27*, 1–18.

Conrad, S. (2000). Will corpus linguistics revolutionize grammar teaching in the 21st century? *TESOL Quarterly, 34*, 548–560.

Conrad, S. (2001). Variation among disciplinary texts: A comparison of textbooks and journal articles in biology and history. In S. Conrad & D. Biber (Eds.), *Variation in English: Multi-dimensional studies* (pp. 94–107). Harlow, Essex: Longman.

Conrad, S., & Biber, D. (1999). Adverbial marking of stance in speech and writing. In S. Hunston & G. Thompson (Eds.), *Evaluation in text: Authorial stance and the construction of discourse* (pp. 56–73). Oxford: Oxford University Press.

Coxhead, A. (2001). A new academic word list. *TESOL Quarterly, 34*, 213–238.

Csomay, E. (2000). Episodes and the vocabulary management profile. Unpublished manuscript. Flagstaff, AZ: Northern Arizona University.

Curzan, A., & Meyer, C. (Eds.). (2000). Special issue on historical corpora. *Journal of English Linguistics, 28*(3).

de Cock, S. (1998). A recurrent word combination approach to the study of formulae in the speech of native and non-native speakers of English. *International Journal of Corpus Linguistics, 3*, 59–80.

de Haan, P. (1999). English writing by Dutch-speaking students. In H. Hasselgärd & S. Oksefjell (Eds.), *Out of corpora: Studies in honour of Stig Johansson* (pp. 203–212). Amsterdam: Rodopi.

Dini, L., & Di Tomaso, V. (1998). Corpus linguistics for application development. *International Journal of Corpus Linguistics, 3*, 305–311.

Ferguson, G. (2001). If you pop over there: A corpus-based study of conditionals in medical discourse. *English for Specific Purposes, 20*, 61–82.

Fox, G. (1998). Using corpus data in the classroom. In B. Tomlinson (Ed.), *Materials development in language teaching* (pp. 25–43). Cambridge: Cambridge University Press.

Gavioli, L., & Aston, G. (2001). Enriching reality: Language corpora in language pedagogy. *ELT Journal, 55*, 238–246.

Ghadessy, P. (1979). Frequency counts, word lists, and materials preparation: A new approach. *English Teaching Forum, 17*, 24–27.

Gledhill, C. (2000). The discourse function of collocation in research article introductions. *English for Specific Purposes, 19*, 115–135.

Gómez-González, M. A. (1998). A corpus-based analysis of extended multiple themes in PresE. *International Journal of Corpus Linguistics, 3*, 81–113.

Granger, S., & Tribble, C. (1998). Learner corpus data in the foreign language classroom: Form-focused instruction and data-driven learning. In S. Granger (Ed.), *Learner English on computer* (pp. 199–209). London: Longman.

Green, C. F., Christopher, E. R., & Kam Mei, J. L. K. (2000). The incidence and effects on coherence of marked themes in interlanguage texts: A corpus-based enquiry. *English for Specific Purposes, 19*, 99–113.

Hasselgård, H., & Oksefjell, S. (Eds.). (1999). *Out of corpora: Studies in honour of Stig Johansson*. Amsterdam: Rodopi.

Helt, M. (2001). A multi-dimensional comparison of British and American spoken English. In S. Conrad & D. Biber (Eds.), *Variation in English: Multi-dimensional studies* (pp. 171–183). Harlow, Essex: Longman.

Hennoste, T., Koit, M., Roosmaa, T., & Saluveer, M. (1998). Structure and usage of the Tartu University corpus of written Estonian. *International Journal of Corpus Linguistics, 3*, 279–304.

Hoard, J. (1998). Language understanding and the emerging alignment of linguistics and natural language processing. In J. Lawler & H. A. Dry (Eds.), *Using computers in linguistics: A practical guide* (pp. 197–230). London: Routledge.

Hockey, S. (1998). Textual databases. In J. Lawler & H. A. Dry (Eds.), *Using computers in linguistics: A practical guide* (pp. 101–133). London: Routledge.

Hughes, R., & McCarthy, M. (1998). From sentence to discourse: Discourse grammar and English language teaching. *TESOL Quarterly, 32*, 263–287.

Hundt, M., & Mair, C. (1999). "Agile" and "uptight" genres: The corpus-based approach to language change in progress. *International Journal of Corpus Linguistics, 4*, 221–42.

Hunston, S., & Francis, G. (1998). Verbs observed: A corpus-driven pedagogic grammar of English. *Applied Linguistics, 19*, 45–72.

Hunston, S., & Francis, G. (2000). *Pattern grammar: A corpus-driven approach to the lexical grammar of English*. Amsterdam: John Benjamins.

Hunston, S., & Sinclair, J. (1999). A local grammar of evaluation. In S. Hunston & G. Thompson (Eds.), *Evaluation in text: Authorial stance and the construction of discourse* (pp. 74–101). Oxford: Oxford University Press.

Johansson, S. (1997). In search of the missing *not*: Some notes on negation in English and Norwegian. In U. Fries, V. Müller, & P. Schneider (Eds.), *From Ælfric to the New York Times: Studies in English corpus linguistics* (pp. 197–214). Amsterdam: Rodopi.

Kaszubski, P. (1998). Enhancing a writing textbook: A national perspective. In S. Granger (Ed.), *Learner English on computer* (pp. 172–185). London: Longman.

Kennedy, G. (1998). *An introduction to corpus linguistics*. London: Longman.

Kilgarriff, A. (in press). Comparing corpora. *International Journal of Corpus Linguistics, 6*, 1–37.

Leech, G. (1999). The distribution and function of vocatives in American and British English conversation. In H. Hasselgärd & S. Oksefjell (Eds.), *Out of corpora: Studies in honour of Stig Johansson* (pp. 107–118). Amsterdam: Rodopi.

Luke, A. (this volume). Beyond science and ideology critique: Developments in critical discourse analysis.

Mauranen, A. (2001). Reflexive academic talk: Observations from MICASE. In R. Simpson & J. Swales (Eds.), *Corpus linguistics in North America: Selections from the 1999 Symposium* (pp. 165–178). Ann Arbor, MI: University of Michigan Press.

McCarthy, M., & Carter, R. (2001). Size isn't everything: Spoken English, corpus, and the classroom. *TESOL Quarterly, 35*, 337–340.

McEnery, T., & Wilson, A. (1996). *Corpus linguistics*. Edinburgh: Edinburgh University Press.

Meunier, F. (1998). Computer tools for the analysis of learner corpora. In S. Granger (Ed.), *Learner English on computer* (pp. 19–37). London: Longman.

Nerbonne, J. (Ed.). (1998). *Linguistic databases*. Stanford, CA: Center for the Study of Language and Information.

Oakes, M. (1998). *Statistics for corpus linguistics*. Edinburgh: Edinburgh University Press.

Owen, C. (1996). Do concordances require to be consulted? *ELT Journal, 50*, 219–224.

Powell, C., & Simpson, R. (2001). Collaboration between corpus linguists and digital librarians for the MICASE web search interface. In R. Simpson & J. Swales (Eds.), *Corpus linguistics in North America: Selections from the 1999 Symposium* (pp. 32–47). Ann Arbor, MI: University of Michigan Press.

Reppen, R. (2001). Register variation in student and adult speech and writing. In S. Conrad & D. Biber (Eds.), *Variation in English: Multi-dimensional studies* (pp. 187–199). Harlow, Essex: Longman.

Rey, J. (2001). Changing gender roles in popular culture: Dialog in *Star Trek* episodes from 1966 to 1993. In S. Conrad & D. Biber (Eds.), *Variation in English: Multi-dimensional studies* (pp. 138–156). Harlow, Essex: Longman.

Schegloff, E. A., Koshik, I., Jacoby, S., & Olsher, D. (this volume). Conversation analysis and applied linguistics.

Sinclair, J. (1987). *Looking up: An account of the COBUILD project in lexical computing*. London: Collins ELT.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Stenström, A. (1999). He was really gormless—She's bloody crap: Girls, boys, and intensifiers. In H. Hasselgård & S. Oksefjell (Eds.), *Out of corpora: Studies in honour of Stig Johansson* (pp. 69–78). Amsterdam: Rodopi.

Stubbs, M. (1996). *Text and corpus analysis*. Oxford: Blackwell.

Svartvik, J. (Ed.). (1992). *Directions in corpus linguistics: Proceedings of the Nobel symposium*. Berlin: Mouton de Gruyter.

Swales, J. (2001). Metatalk in American academic talk: The cases of *point* and *thing*. *Journal of English Linguistics, 29*, 34–54.

Swales, J., & Malczewski, B. (2001). Discourse management and new-episode flags in MICASE. In R. Simpson & J. Swales (Eds.), *Corpus linguistics in North America: Selections from the 1999 Symposium* (pp. 145–64). Ann Arbor, MI: University of Michigan Press.

Thomas, J., & Short, M. (Eds.). (1996). *Using corpora for language research.* London: Longman.

Thurston, J., & Candlin, C. (1998). Concordancing and the teaching of the vocabulary of academic English. *English for Specific Purposes, 17,* 267–80.

Verhagen, A. (2000). Interpreting usage: Construing the history of Dutch causal verbs. In M. Barlow & S. Kemmer (Eds.), *Usage-based models of language* (pp. 261–286). Stanford, CA: Center for the Study of Language and Information.

Williams, G. C. (1998). Collocational networks: Interlocking patterns of lexis in a corpus of plant biology research articles. *International Journal of Corpus Linguistics, 3,* 151–171.

Youmans, G. (1991). A new tool for discourse analysis: The vocabulary management profile. *Language, 67,* 763–89.

Xue, G., & Nation, I.S.P. (1984). A university word list. *Language Learning and Communication, 3,* 215–229.