# Corpus linguistics and the study of literature

## Back to the future?

Douglas Biber
Northern Arizona University, Flagstaff

The present paper introduces corpus-based analytical techniques and surveys some of the specific ways in which corpus analysis has been applied to the study of literature. In recent years, those research efforts have been mostly carried out under the umbrella of 'corpus stylistics'. Most of these studies focus on the distribution of words (analyzing keywords, extended lexical phrases, or collocations) to identify textual features that are especially characteristic of an author or particular text. Corpus-based grammatical and pragmatic analyses of literary language are also briefly considered. Then, in the concluding part of the paper, I briefly survey earlier computational and statistical research on authorship attribution and literary style. While that research tradition is in some ways the precursor to more recent work in corpus stylistics, it is also complementary to recent research in its application of sophisticated statistical and computational methods.

**Keywords:** corpus stylistics, keyword analysis, collocations, lexical phrases, key clusters, semantic prosody

Corpus linguistics is a research approach that facilitates empirical descriptions of language use. Corpus linguistic research is based on analysis of a 'corpus': a large and principled collection of texts stored on computer. A corpus is a sample, designed to represent a textual domain in a language, such as everyday conversation in English, newspaper editorials, personal email messages, or the novels of Charles Dickens. Just like any sample, a corpus can be evaluated for the extent to which it represents a 'population' — in this case, the target textual domain (see Biber 1993/2004). Thus, research carried out on a corpus has the goal of describing the patterns of language use in the target textual domain.

In fact, it could be argued that a corpus provides the best way to represent a textual domain, and corpus analysis is the most powerful empirical approach for analyzing the patterns of language use in that domain. Such analyses are applicable

*Scientific Study of Literature* 1:1 (2011), **15–23**. DOI 10.1075/ssol.1.1.02bib
ISSN 2210–4372 / E-ISSN 2210–4380 

in any sub-discipline of linguistics that includes consideration of language use, including the study of lexical and grammatical variation, discourse patterns, spoken and written register variation, historical change, etc.

I see the study of literary language as no exception here. One of the major research questions within the 'scientific study of language' concerns the nature of the words, lexical expressions, and grammatical forms used in different literary texts and varieties. Corpus analysis is ideally suited to research questions of this type.

There are numerous specific methodological techniques that can be applied to corpus investigations. However, corpus analysis generally shares four characteristics (see Biber, Conrad, & Reppen 1998, p. 4):

1. it is empirical, analyzing the actual patterns of language use in natural texts;
2. it utilizes a large and principled collection of natural texts, known as a "corpus," as the basis for analysis;
3. it makes extensive use of computers for analysis, using both automatic and interactive techniques;
4. it depends on both quantitative and qualitative analytical techniques.

Over the past several decades, there have been numerous stylistic studies of literary language with these characteristics, employing computational/quantitative analyses of words in literary texts. In recent years, such studies have come to be explicitly identified with corpus linguistics, carried out mostly under the umbrella of 'corpus stylistics' (see Mahlberg 2007a, to appear; Wynne 2006).

Most corpus-stylistic studies focus on the distribution of words to identify textual features that are especially characteristic of an author, particular text, or even a single character within a play or novel. Three major methodological approaches have been used to study the stylistic relevance of such word distributions: 'keyword' analysis, identifying typical extended lexical phrases, and collocational analysis.

All three types of analysis require the use of two corpora: a target corpus, consisting of the literary works in question, and a larger, more general comparison corpus. The comparison corpus is used to represent 'typical' patterns of use, making it possible to empirically identify distinctive linguistic patterns in the target corpus that depart from those typical patterns. These distinctive patterns of use can then be interpreted for their stylistic influence.

For example, a keyword analysis attempts to identify the individual words that characterize a literary work: words that are statistically more likely to occur in the target corpus than in the comparison corpus. The chi-squared statistic or log-likelihood statistic is used to identify such words. This statistic is often combined with a frequency cut-off, to avoid identification of very rare words. (See Culpeper 2009; Scott, & Tribble 2006 for useful introductions to keyword analysis.)

The choice of comparison corpus is probably the most important decision that must be made when doing a keyword analysis. One common practice is to choose a very general corpus, such as the *British National Corpus* (BNC). The problem with such an approach is that it likely confounds high-level register differences with the much more specific stylistic differences that are the intended focus of investigation. For this reason, many stylistic studies use a much more restricted comparison corpus. For example, Scott and Tribble (2006) do a keyword analysis of the distinctive words used in Shakespeare's play *Romeo and Juliet*; the comparison corpus in this case consisted of all other Shakespeare plays. Fischer-Starcke (2009) employs a similar approach, describing the distinctive keywords of Jane Austen's *Northanger Abbey* contrasted with two different comparison corpora: one consisting of Austen's six novels, and a second consisting of other fictional texts that are contemporary with Austen. This approach thus enables identification of the distinctive words used in a particular literary work (rather than more general words associated with a high level register).

Culpeper takes this type of analysis one step further, analyzing the distinctive keywords used by each individual character in *Romeo and Juliet*. To identify keywords at this level of specificity, Culpeper constructed comparison corpora consisting of the utterances produced by all other characters in the play. This approach resulted in both expected findings (e.g., Romeo uses the words *beauty* and *love* more than other characters) as well as some surprising findings. Thus for Juliet, the analysis shows that the most distinctive word (i.e., the word with the highest 'keyness' score) is the conditional subordinator *if*. A more detailed consideration of these occurrences shows that these conditional clauses support Juliet's preoccupation with different possible events that might occur. While the corpus analysis does not discover anything new about Juliet's mental state, it does help us to understand the linguistic devices used to achieve this effect.

A related major approach used for corpus-stylistic analyses focuses on the extended lexical phrases that are characteristic of particular authors or literary works, referred to as lexical 'clusters' or 'bundles' (see, e.g., Biber, Conrad, & Cortes 2004; Partington, & Morley 2004). Two scholars have been especially productive applying this approach to literary works: Fischer-Starcke (2009, 2010) studying the novels of Jane Austen, and Mahlberg (2007b, to appear) studying the novels of Dickens. Similar to keyword analysis, a 'key cluster' analysis can be used to identify the lexical sequences that are especially characteristic of a particular text when compared to a more general corpus. For example, Mahlberg (2007) shows that lexical sequences like *his hands in his pockets* and *as if he would have* are especially prevalent in Dickens' novels when compared to a more general corpus of 19th c. fiction. The stylistic importance of such sequences, however, is shown by more detailed consideration of the clusters in particular textual contexts. For

this purpose, Mahlberg (2007, to appear) considers the local textual functions of clusters, which can be grouped into the following major categories: characterizing people, places and things, expressing interaction between characters, describing looks and movements, creating textual worlds by comparison and contrast (*as if* clusters), and locating and relating actions in time and place.

Finally, a third major approach used for corpus-stylistic analyses of words relies on the notion of 'collocation': the way in which particular words are associated with each other, as shown empirically by their tendency to co-occur more frequently than would be expected by chance (see Adolphs 2006; Partington 1998). In some cases, a target word co-occurs with a large set of collocates that all have the same 'semantic prosody': an underlying evaluative meaning, which is usually categorized simply as whether the object or event is considered to be good or bad (see Louw 1993/2004; Partington 2003, 2004). For example, Sinclair (1987, p. 155ff) discusses how the intransitive phrasal verb *set in* has a negative underlying evaluation, being almost always associated with undesirable events. One reflection of this evaluation is the set of nouns that co-occur with this verb as grammatical subject. For example, in an early version of the COBUILD Corpus, Sinclair found nouns like *rot*, *decay*, *ill-will*, *decadence*, *impoverishment*, *infection*, *prejudice*, *rigor mortis*, *numbness*, and *bitterness* occurring as subject with the main verb *set in*. In this case, none of these words is individually especially frequent as a collocate, but the whole set of words has an underlying evaluative meaning.

The notion of semantic prosody turns out to be very useful for explaining the stylistic effect of particular expressions in a literary work. For example, Louw (1993.157) discusses the following sentence from the novel *Small World* by David Lodge:

> The modern conference resembles the pilgrimage of medieval Christendom in that it allows the participants to indulge themselves in all the pleasures and diversions of travel while apparently bent on self-improvement.

Corpus analysis helps to explain the reader's perception that this sentence is somehow humorous or ironic. In particular, Louw focuses on the verb *bent on*, investigating its use in an early version of the Bank of English corpus. In many instances, *bent on* is associated with negatively evaluated actions, such as *destroying*, *harrying*, *demanding*, *mayhem*, *mischief*, *revenge*. Lodge builds on this expectation of negative actions to create an ironic effect, by combining *bent on* with the noun *self-improvement*.

Several studies have used corpus-based investigations of semantic prosody to help explain the stylistic effect produced by particular passages in novels or poems (see Adolphs 2006). For example, Adolphs and Carter (2002) investigate the use of semantic prosodies in Virginia Woolf's *To the Lighthouse*, and O'Halloran (2007)

describes how semantic prosodies help to create the 'disturbing' literary effect of Fleur Adcock's poem *Street Song*. These studies have quite different research goals from the keyword investigations surveyed above, focusing on the stylistic effect of a single expression or sentence rather than the overall stylistic characteristics of a text or body of texts. However, in both types of study, large-scale corpus analysis is used to provide empirical evidence to support stylistic judgments.

All of the corpus-stylistic studies surveyed above focus on words. It is much less common to analyze grammatical characteristics in such studies. There are two major reasons for this bias. First, it is difficult to analyze grammatical characteristics in a corpus, requiring the use of specialized computer programs (a grammatical 'tagger'). But additionally, it is easier to notice the stylistic importance of word choice, while grammatical characteristics are much less salient. However, as Biber and Conrad (2010) show, registers and individual styles are distinguished by the use of pervasive grammatical features (see also Biber, & Finegan 1994). These are often not salient to the human observer, but quantitative corpus analysis is ideally suited for the identification of such patterns. Further, grammatical patterns can be interpreted functionally (see, e.g., Biber, & Conrad 2010, pp. 144–157). These interpretations are somewhat more difficult and subtle than the interpretation of lexical characteristics, reflecting communicative purpose, mode, and relations among participants, rather than directly reflecting topic. Nevertheless, grammatical features are functional in addition to being purely indexical of particular authors or periods.

One good example of a detailed corpus-based grammatical investigation is the study of syntactic complements in Flannery O'Connor's fiction (Hardy, & Durian 2000; see also Hardy 2007). This study compares the use of syntactic complements in O'Connor's fiction to the grammatical patterns observed generally in the Brown Corpus, and the distinctive grammatical characteristics of O'Connor's fiction are interpreted relative to the perspectives of physical versus cognitive perceptions and the foreground/background cline.

All of the difficulties inherent in corpus-based analyses of grammatical characteristics are multiplied many times over in analyses of pragmatic characteristics. Such analyses entail detailed hand-coding, which require countless hours of effort when applied to an entire corpus. However, the benefits of such an approach for stylistic descriptions are dramatic. The most impressive example of such a study is the investigation of variability in the forms and functions of linguistic devices used for speech, writing, and thought presentation, carried out by Semino and Short (2004). The analysis is based on a corpus of 120 narrative texts, from fiction, newspaper reportage, and (auto)biographies. Each instance of a language or thought report was manually coded for its grammatical and pragmatic characteristics (e.g. direct speech, indirect speech, free indirect thought, etc.). Then the overall patterns were analyzed in each register, including discussion of instances that did

not fit previous theoretical models. That is, the approach forced the analysts to consider all instances of this pragmatic phenomenon, rather than just presenting examples of 'interesting' instances to illustrate pre-conceived theoretical points.

Corpus-stylistic research is widely considered to be a recent innovation: an extension of earlier linguistic corpus-based studies, which focused on more traditional issues of lexical and grammatical variation and use. In fact, as recently as 2005, Wynne could write that 'apart from some important studies described here, there is little use of language corpora, or the techniques of corpus linguistics, in the study of literary style'.

However, this spin on the historical development of corpus-stylistic research disregards the long tradition of computational and statistical research on authorship attribution and literary style, published mostly in journals like *Literary and Linguistic Computing* and *Computers in the Humanities*. While these studies have somewhat different research objectives, they certainly have made extensive use of "language corpora" and "the techniques of corpus linguistics" for the study of literary style. Exemplary studies in this tradition include Youmans (1990, 1991) on the 'vocabulary management profile' (used to analyze the progression of discourse within literary works); Burrows (1987; cf 1989, 2003) on the analysis of vocabulary and grammatical function words in novels by Jane Austen; McKinnon's (1989) study of 250 common nouns in 34 of Kierkegaard's books; and Hoover's (2002) investigation of extended lexical sequences in a corpus of c. 30 text extracts from novels written by various authors. But even as early as the 1960s, scholars in this tradition were applying essentially the same research techniques as present-day corpus-stylisticians to investigate issues of disputed authorship. Probably the most notable example of this type is Mosteller and Wallace's (1964) analysis of the distribution of function words (e.g., prepositions, conjunctions, articles) in a corpus of the 85 'Federalist Papers'. (See Grieve 2007 for a survey of previous computational research on authorship attribution, and an empirical comparison of the effectiveness of different methodological approaches.)

So, does it make sense for recent research under the umbrella of 'corpus stylistics' to disregard the research of previous decades, carried out under the umbrella of computational analyses of style? While there are important differences, I would argue that the answer to this question is 'no'. That is, despite the differences, these are clearly related research traditions that could benefit from greater integration.

What are the differences? First, there is a difference in primary research goals: to analyze the linguistic bases of literary style versus attribution of authorship. However, earlier computational research often had both of these goals, so this is not a clear-cut distinction between the two traditions.

Second, it might be argued that there is a difference in the textual characteristics under consideration. In particular, some earlier studies were only marginally

linguistic in focus, considering instead variables like grapheme frequency or the frequency of different punctuation marks. However, other early studies focused on the same kinds of linguistic features as present-day corpus-stylistic studies, such as the distinctive vocabulary of texts or the use of distinctive multi-word sequences.

Third, and most importantly, most early studies treated linguistic variables as purely indexical. Thus, the ultimate goal was simply to identify the linguistic characteristics of a text or author that can reliably distinguish that text/author from other texts and authors. These studies usually gave no attention to functional/stylistic interpretation. In contrast, corpus-stylistic studies are situated within the broader subdiscipline of literary stylistics (see, e.g., Leech, & Short 1981), which focuses on the functional and aesthetic associations of linguistic patterns. Thus, corpus-based analyses of linguistic form are only the first step of a corpus-stylistic study, providing the foundation for subsequent functional interpretation.

However, even this difference is not absolute. Thus, researchers like Burrows (1987, 1992), Craig (1999), and van Peer (1989) all argued early on for the importance of literary interpretation following the quantitative identification of distinctive linguistic features.

Conversely, there is one way in which the earlier studies are actually more innovative than most recent corpus-stylistic research: in the application of sophisticated quantitative/statistical analytical techniques. Such research methods are generally avoided by both literary scholars and corpus linguists. However, a hallmark of nearly all corpus research is the reliance on quantitative findings. Early research in the computational analysis of style focused heavily on the development of appropriate, and often quite sophisticated, statistical methods for the analysis of linguistic distributional patterns; present-day research in corpus-stylistics might benefit from a greater awareness of the benefits of such analyses.

In sum, it can be argued that corpus-based research methods have been used to study literature for the past 50 years or more. Over the past decade, this research has become increasingly popular, carried out under the umbrella of 'corpus stylistics'. These recent studies are explicit in their goals of integrating both the methods of corpus linguistics and the goals and methods of traditional stylistic research. However, there seems to be great potential for new lines of research that integrate the statistical methods of earlier research with the more rhetorical concerns of recent studies.

## References

Adolphs, S. (2006). *Introducing electronic text analysis. A practical guide for language and literary studies*. London: Routledge.

Adolphs, S., & Carter, R. (2002). Point of view and semantic prosodies in Virginia Woolf's *To the Lighthouse*. *Poetica*, *58*, 7–20.

Biber, D. (1993/2004). Representativeness in corpus design. *Literary and Linguistic Computing, 8*, 243–257. (Reprinted in Sampson, G., & McCarthy, D. (Eds.). (2004). *Corpus linguistics: Readings in a widening perspective* (pp. 174–97). London: Continuum).

Biber, D., & Conrad, S. (2009). *Register, genre, and style*. Cambridge: Cambridge University Press.

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

Biber, D., Conrad, S., & Cortes, V. (2004). *If you look at…*: Lexical bundles in university teaching and textbooks. *Applied Linguistics, 25*, 371–405.

Biber, D., & Finegan, E. (1994). Multi-dimensional analyses of authors' styles: Some case studies from the eigtheenth century. *Research in Humanities Computing 3*, D. Ross & D. Brink (Eds.), 3–17. Oxford: Oxford University Press.

Burrows, J. F. (1987). *Computation into criticism. A study of Jane Austen's novels and an experiment in method*. Oxford: Clarendon.

Burrows, J. F. (1989). 'An ocean where each kind..': Statistical analysis and some major determinants of literary style. *Computers and the Humanities*, *23*, 309–21.

Burrows, J. F. (1992). 'Not unless you ask nicely': The interpretive nexus between analysis and information. *Literary and Linguistic Computing 7*, 91–110.

Burrows, J. F. (2003). Questions of authorship: Attribution and beyond. *Computers and the Humanities*, *37*, 5–32.

Craig, H. (1999). Authorial attribution and computational stylistics: If you can tell authors apart, have you learned anything about them? *Literary and Linguistic Computing, 14*, 103–13.

Culpeper, J. (2009). Keyness: Words, parts-of-speech and semantic categories in the character-talk of Shakespeare's *Romeo and Juliet. International Journal of Corpus Linguistics, 14*, 29–59.

Fischer-Starcke, B. (2009). Keywords and frequent phrases of Jane Austen's *Pride and Prejudice:* A corpus-stylistic analysis. *International Journal of Corpus Linguistics*, *14*: 492–523.

Fischer-Starcke, B. (2010). *Corpus linguistics in literary analysis: Jane Austen and her contemporaries*. London: Continuum.

Grieve, J. (2007). Quantitative authorship attribution: an evaluation of techniques. *Literary and Linguistic Computing, 22*, 251–270.

Hardy, D. E. (2007). *The body in Flannery O'Connor's fiction: Computational technique and linguistic voice*. Columbia, SC: University of South Carolina Press.

Hardy, D. E., & Durian, D. (2000). The stylistics of syntactic complements: Grammar and seeing in Flannery O'Connor's fiction. *Style*, *34*, 92–116.

Hoover, D. (2002). Frequent word sequences and statistical stylistics and authorship attribution: An empirical investigation. *Literary and Linguistic Computing, 17*, 157–180 .

Leech, G. N., & Short M. H. (1981). *Style in fiction: A linguistic introduction to English fictional prose*. London: Longman.

Louw, W. (1993/2004). Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies. In M. Baker, G. Francis, & E. Tognini-Bonelli (Eds.), *Text and technology* (pp. 157–176). Amsterdam: John Benjamins. [Reprinted in Sampson, G., & McCarthy, D. (Eds.). (2004). *Corpus linguistics: Readings in a widening discipline* (pp. 229–241). London: Continuum].

Mahlberg, M. (2007a). Corpus stylistics: Bridging the gap between linguistic and literary studies. In M. Hoey, M. Mahlberg, M. Stubbs, & W. Teubert (Eds.), *Text, discourse and corpora: Theory and analysis* (pp. 219–246). London: Continuum.

Mahlberg, M. (2007b). Clusters, key clusters and local textual functions in Dickens. *Corpora, 2,* 1–31.

Mahlberg, M. (to appear). *Corpus stylistics and Dickens's fiction*. London: Routledge.

McKinnon, A. (1989). Mapping the dimensions of a literary corpus. *Literary and Linguistic Computing, 4,* 73–84.

Mosteller, F., & Wallace, D. (1964). *Inference and disputed authorship: The federalist*. Reading, MA: Addison-Wesley.

O'Halloran, K. (2007). Corpus-assisted literary evaluation. *Corpora, 2*, 33–63.

Partington, A. (1998). *Patterns and meanings*. Amsterdam: John Benjamins.

Partington, A. (2003). *The linguistics of political argument: The spin-doctor and the wolf-pack at the white house*. London: Routledge.

Partington, A. (2004). 'Utterly content in each other's company': Semantic prosody and semantic preference. *International Journal of Corpus Linguistics, 9*, 131–156.

Partington, A., & Morley, J. (2004). At the heart of ideology: Word and cluster/bundle frequency in political debate. In B. Lewandowska-Tomaszczyk, (Ed.), *PALC 2003: Practical applications in language corpora* (pp. 179–192). Frankfurt: Peter Lang.

van Peer, W. (1989). Quantitative studies of style: A critique and an outlook. *Computers and the Humanities, 23*, 301–307.

Scott, M., & Tribble, C. (2006). *Textual patterns. Key words and corpus analysis in language education*. Amsterdam: John Benjamins.

Semino, E., & Short, M. (2004). *Corpus stylistics. Speech, writing and thought presentation in a corpus of English writing*. London: Routledge.

Sinclair, J. (1987). The nature of the evidence. In J. Sinclair (Ed.), *Looking up* (pp. 150–159). Glasgow: Collins.

Wynne, M. (2005). Stylistics: Corpus approaches. Available from: www.pala.ac.uk/resources/sigs/corpus-style/Corpora_stylistics.pdf

Wynne, M. (2006). Stylistics and language corpora. In K. Brown (Ed.), *Encyclopedia of language and linguistics*. Oxford: Elsevier.

Youmans, G. (1990). Measuring lexical style and competence: The type-token vocabulary curve. *Style, 24*, 584–599.

Youmans, G. (1991). A new tool for discourse analysis: The vocabulary-management profile. *Language, 67*, 763–789.

## *Author's address*

Douglas Biber
Northern Arizona University
English Department
Box 6032
Flagstaff AZ 86011-6032
USA

Douglas.Biber@NAU.EDU