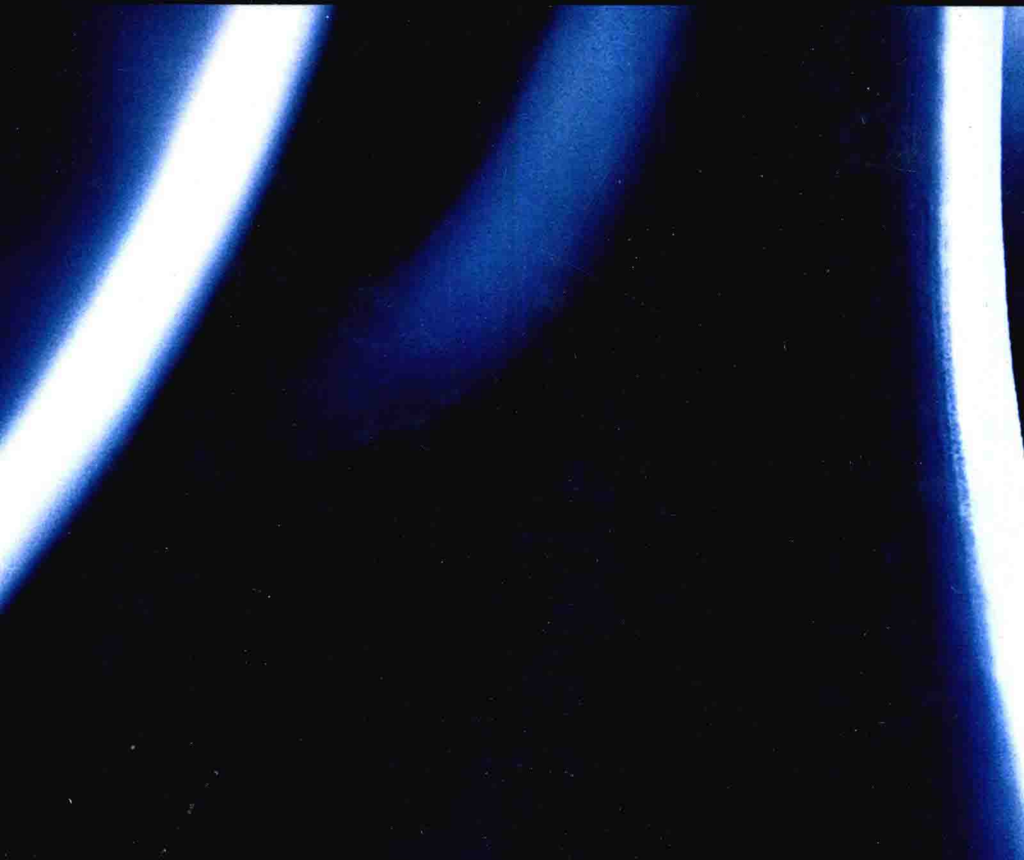
ROUTLEDGE ADVANCES IN CORPUS LINGUISTICS



Digital Literary Studies

Corpus Approaches to Poetry, Prose, and Drama

David L. Hoover, Jonathan Culpeper,
and Kieran O'Halloran

ROUTLEDGE

Digital Literary Studies

Corpus Approaches to Poetry, Prose, and Drama

David L. Hoover, Jonathan Culpeper, and Kieran O'Halloran

Routledge
Taylor & Francis Group
NEW YORK AND LONDON

First published 2014
by Routledge
711 Third Avenue, New York, NY 10017

and by Routledge
2 Park Square, Milton Park, Abingdon, Oxon OX14 4RN

*Routledge is an imprint of the Taylor & Francis Group,
an informa business*

Library of Congress Cataloging-in-Publication Data

A catalog record has been requested for this book.

ISBN: 978-0-415-35230-7 (hbk)
ISBN: 978-0-203-69891-4 (ebk)

Typeset in Sabon
by Apex CoVantage, LLC

Digital Literary Studies

Digital Literary Studies presents a broad and varied picture of the promise and potential of methods and approaches that are crucially dependent upon the digital nature of the literary texts it studies and the texts and collections of texts with which they are compared. It focuses on style, diction, characterization, and interpretation of single works and across larger groups of texts, using both huge natural language corpora and smaller, more specialized collections of texts created for specific tasks, and applies statistical techniques used in the narrower confines of authorship attribution to broader stylistic questions. It addresses important issues in each of the three major literary genres, and intentionally applies different techniques and concepts to poetry, prose, and drama. It aims to present a provocative and suggestive sample intended to encourage the application of these and other methods to literary studies.

Hoover, Culpeper, and O'Halloran push the methods, techniques, and concepts in new directions, apply them to new groups of texts or to new questions, modify their nature or method of application, and combine them in innovative ways.

David L. Hoover is Professor of English at New York University, USA. His publications in stylistics and digital humanities include three books—*A New Theory of Old English Meter*, *Stylistics: Prospect and Retrospect*, and *Language and Style in "The Inheritors"*—and numerous articles on authorship attribution and corpus and computational stylistics.

Jonathan Culpeper is Professor of English Language and Linguistics in the Department of Linguistics and English Language at Lancaster University, UK. His major publications include *Language and Characterisation in Plays and Other Texts* (2001) and *Early Modern English Dialogues: Spoken Interaction as Writing* (2010; coauthored with Merja Kytö).

Kieran O'Halloran is a Reader in Applied Linguistics at King's College, University of London, UK. Publications include *Critical Discourse Analysis and Language Cognition* (2003), *The Art of English: Literary Creativity* (2006 with Goodman), and *Applied Linguistics Methods* (Routledge, 2010 with Coffin and Lillis).

Routledge Advances in Corpus Linguistics

Edited by Tony McEnery, *Lancaster University, UK*

Michael Hoey, *Liverpool University, UK*

1 **Swearing in English**
Bad Language, Purity and Power from 1586 to the Present
Tony McEnery

2 **Antonymy**
A Corpus-Based Perspective
Steven Jones

3 **Modelling Variation in Spoken and Written English**
David Y. W. Lee

4 **The Linguistics of Political Argument**
The Spin-Doctor and the Wolf-Pack at the White House
Alan Partington

5 **Corpus Stylistics**
Speech, Writing and Thought Presentation in a Corpus of English Writing
Elena Semino and Mick Short

6 **Discourse Markers Across Languages**
A Contrastive Study of Second-Level Discourse Markers in Native and Non-Native Text with Implications for General and Pedagogic Lexicography
Dirk Siepmann

7 **Grammaticalization and English Complex Prepositions**
A Corpus-Based Study
Sebastian Hoffmann

8 **Public Discourses of Gay Men**
Paul Baker

9 **Semantic Prosody**
A Critical Evaluation
Dominic Stewart

10 **Corpus Assisted Discourse Studies on the Iraq Conflict**
Wording the War
Edited by John Morley and Paul Bayley

11 **Corpus-Based Contrastive Studies of English and Chinese**
Richard Xiao and Tony McEnery

12 **The Discourse of Teaching Practice Feedback**
A Corpus-Based Investigation of Spoken and Written Modes
Fiona Farr

13 **Corpus Approaches to Evaluation**
Susan Hunston

14 **Corpus Stylistics and Dickens’s Fiction**
Michaela Mahlberg

15 **Spoken Corpus Linguistics**
From Monomodal to Multimodal
Svenja Adolphs and Ronald Carter

16 **Digital Literary Studies**
Corpus Approaches to Poetry, Prose, and Drama
David L. Hoover, Jonathan Culpeper, and Kieran O’Halloran

Figures

Tables

Acknowledgments

The authors are grateful to the publishers below for permission to reproduce copyright material.

Bloodaxe Books kindly gave permission to use Fleur Adcock's "Street Song," from Fleur Adcock, 2000. *Poems 1960–2000*. Tarset, Northumberland: Bloodaxe Books.

Chapter 2 is based on Culpeper (2002), and chapter 3 is based on Jonathan Culpeper, 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* (1): 29–56, by kind permission of John Benjamins Publishing Company, Amsterdam/Philadelphia. www.benjamins.com

Chapter 6 is a revised version of Kieran O'Halloran, "Corpus-Assisted Literary Evaluation," *Corpora* 2 (1) (2007): 33–63, by kind permission of Edinburgh University Press, www.euppublishing.com.

Chapter 7 is a revised version of Kieran O'Halloran, 2012, "Performance Stylistics: Deleuze and Guattari, Poetry, and (Corpus) Linguistics," *International Journal of English Studies* 12(2) (2012): 171–199. It is provided by kind permission of the *International Journal of English Studies*.

The authors would also like to thank the series editors, especially Michael Hoey, for reading and commenting on chapters at various points.

Contents

1 Introduction

David L. Hoover, Jonathan Culpeper, and Kieran O'Halloran

1.1 AIMS

The phrase "digital literary studies," a recent coinage that owes much to the only slightly less recent "digital humanities," has caught fire in academic circles in the last few years. Both phrases usefully focus our attention on the ways that the methods and approaches we present here, among others, are crucially dependent upon the digital nature of our objects of study. In this book, our primary objects of study are literary texts, but not one of them was born digital. In most of our analyses below, these objects themselves are studied in their digitized forms, and in all of them digital texts and collections of texts are crucial to the analysis. Culpeper and Hoover exploit digital forms of (groups of) novels and plays directly by analyzing them computationally. O'Halloran, who focuses on lyric poems that are too short for nearly all methods of computational analysis, instead brings to bear information collected from huge digital collections of natural language or "corpora" ("corpus" being the singular form).

The book presents a broad and varied picture of the possibilities of digital literary studies. We study style, diction, and characterization, both within a single work and across the work of an author, a group of authors, or across an author's career. We use both huge natural language corpora and smaller, more specialized collections of texts created for specific tasks. We apply statistical techniques normally used in the narrower confines of authorship attribution to broader stylistic questions, such as style variation. Finally, we examine important issues in each of the three major literary genres. We have intentionally applied a different selection of techniques and concepts to poetry, prose, and drama, in ways that reflect some of our own current major concerns. Yet, in spite of this variety, we make no attempt to cover all of the possible approaches that are currently being used in this fast-growing field. Rather, we have aimed at a provocative and suggestive sample that will encourage other researchers to apply these and other methods across all genres, periods, and styles of literature. Some of the specific approaches represented here, such as multivariate analysis and text-markup/annotation, have long traditions, others, such as the use of huge corpora, are of more recent origin. We have tried, throughout, however, to push the methods, techniques, and concepts in new directions, to apply them to new groups of texts or to new questions, to modify their nature or method of application, and to combine them in innovative ways.

1.2 DIGITAL LITERARY STUDIES AND CORPUS LINGUISTICS

Most of what we present in this book belongs squarely within the tradition of what might be called rather broadly, "textual analysis." This tradition exploits the digital nature of the texts by analyzing them computationally. At the other end of the spectrum lie studies that focus on the nature of the medium itself and how new media impact literary studies more generally; such studies tend to be more theoretical than practical, and need not make significant use of computation. And there is plenty of room left in the spectrum for many other kinds of digital literary studies. The fact that we have nothing to say about the Text Encoding Initiative (TEI), electronic literature, scholarly digital editions, digital archives and portals, blogging and social media, new media theory, or other areas of digital literary study is a result of our own central interests and not a comment on the significance or value of other approaches.

As reflected in the title of the book, a crucial purpose of ours is to flag the value of corpus linguistics (and corpora in general) for digital literary studies. This is a method that engages in the building and exploitation of corpora, the latter involving software programs that extract particular kinds of linguistic features from the corpus and undertake statistical analyses. It is not as though corpus linguistics is unheard of in the digital humanities. But, it is often seen as one branch of digital humanities that you can take or leave rather than something that has implications for language study generally in the humanities, not to mention the social sciences. This is perhaps, in part, because of "linguistics" in "corpus linguistics"—it looks specialized and forbidding. There are plenty of analytical frameworks in linguistics that are technically sophisticated, demanding to learn, and challenging to apply successfully. But, this is not the case for corpus linguistics. Compared to many other approaches in linguistics, it is accessible, and light on terminology. This is because it is much more of a set of methods and principles for the analysis of electronic language data than a complex theoretical perspective on language. Terminology used in corpus linguistics, such as "collocation," "keywords," "phraseology," "semantic preference," and "semantic prosody," crops up in the book. To promote corpus linguistic methods for use in the digital humanities, we have produced a glossary where these corpus linguistic terms and others are explained.

For our purposes, a corpus can be defined simply as any structured collection of digital texts, and the structuring principles can be quite various. Giant natural language corpora, for example, are typically balanced by genre, containing the same amount of text from genres like spoken language, literature, news reporting, and so forth. Historical corpora are usually balanced by date. Specialized corpora are often specifically designed and created so as to provide a norm against which an author, text, or part of a text can be compared. We use all of these kinds of corpora below. In chapters 2 and 3, for example, Culpeper studies the character parts in *Romeo and Juliet* by comparing the speech of each important character with a specially created corpus consisting of the speech of the remaining characters. In chapters 4 and 5, Hoover uses specially created corpora of novels by contemporaries to investigate the styles of Wilkie Collins, Hannah Webster Foster, and Henry James, and treats early, intermediate, and late James texts as subcorpora to analyze changes in his style over his long career. In chapters 6 and 7, O'Halloran compares the ways words, phrases, and collocations are used in lyric poems and in giant natural language corpora. In all of these uses, the corpora are used in the service of comparison.

1.3 STYLISTICS AND CORPUS STYLISTICS

We cannot neglect mention of a field in which textual analysis engages, typically, literary texts; that is, stylistics—how linguistic analysis can account for readers' interpretations, including their impressions of style and experience of aesthetic effects (for example, Carter and Stockwell 2008; Cook 1994; Jeffries and McIntyre 2010; Leech and Short 2007, Simpson 2004; Verdonk 2002; Wales 2011; Widdowson 1992). Stylistics has been a thriving discipline since the 1960s, but its roots go back further. It can be viewed as the logical outcome of literary criticism in the first half of the twentieth century, which placed emphasis on studying texts rather than authors. Russian formalism (especially the work of Roman Jakobson) and the related Prague school (especially the work of Jan Mukařovský) are key influences on the development of stylistics. Stylistic analysis of literature is a technique we draw on to different degrees in the chapters. The glossary also includes definitions of linguistic concepts we employ.

Another purpose of our book is to broadcast the value of corpus linguistics for stylisticians. The combination of stylistic analysis and corpus linguistic method is relatively recent, but an area of increasing popularity (for example, Semino and Short 2004; Hoover 2010a; Hori 2004; O'Halloran 2007; Stubbs 2005; Mahlberg 2013; Fischer-Starcke 2010). The label "corpus stylistics" perhaps first appears as the main title of Semino and Short (2004), though there are a number of precursors doing corpus stylistic work without using that label (for example, Stubbs 1996, 81–100; Louw 1997; Hoover 1999). Corpus linguistics may seem, at least to some, an alien enterprise in the world of stylistics, but this is an erroneous perception on several counts. Stylisticians, and indeed literary critics, often discuss matters of frequency. Any pattern they point to or tendency they highlight is a matter of frequency, of statistics by the back door. Moreover, key works in stylistics emphasize the need for quantification (for example, Leech and Short 2007, chapter 2), and the empirical study of literature—mostly revolving around informant testing—is an important subfield. Indeed, deviation from statistical norms is one way of characterizing an aspect of foregrounding theory (for example, Mukařovský 1970), a theory that has been a keystone in stylistics for decades. We will touch on some aspects of foregrounding theory in our work, notably in chapter 2.

Corpus linguists do not argue that the corpus-related approach should be all-consuming, but rather that it "should be seen as a complementary approach to more traditional approaches" (Biber, Conrad, and Reppen 1998, 7–8). Moreover, the corpus-related approach itself deploys a mixture of methods; it is not—and this is particularly true of more recent work—a purely quantitative methodology, all computers and numbers, as the following quotations from popular textbooks in corpus linguistics argue:

> both qualitative and quantitative analyses have something to contribute to corpus study. Qualitative analysis can provide greater richness and precision, whereas quantitative analysis can provide statistically reliable and generalisable results. There has recently been a move in social science research towards multi-method approaches which largely reject the narrow analytical paradigms in favour of the breadth of information which the use of more than one method may provide. Corpus linguistics could . . . benefit as much as any field from such multi-method research, combining both qualitative and quantitative perspectives on the same phenomena. (McEnery and Wilson 2001, 76–77)

> it is important to note that corpus-based analyses must go beyond simple counts of linguistic features. That is, it is essential to include qualitative, functional interpretations of quantitative patterns. In each chapter of this book, you will find that a great deal of space is devoted to explanation, exemplification, and interpretation of the patterns found in quantitative analyses. The goal of corpus-based investigations is not simply to report quantitative findings, but to explore the importance of those findings for learning about the patterns of language use. (Biber et al. 1999, 5)

Throughout this book we will be combining qualitative and quantitative analyses.

Still, despite the overlaps between these fields, one might reasonably ask what particular value corpus linguistics brings to stylistics. We would argue that corpus-related techniques allow one's analysis to (1) encompass more than one could possibly encompass with reasonable human labor, and (2) be systematic, and (3) be fine-grained. These are the very features we aim to demonstrate in this book. Through being able to encompass the *whole* of, for example, an author's work, their entire canon, or a large body of