Corpus Linguistics in Literary Analysis

Bettina Fischer-Starcke

Jane Austen and her Contemporaries

Studies in Corpus and Discourse
Series Editors: Wolfgang Teubert and Michaela Mahlberg

continuum

Corpus Linguistics in Literary Analysis

Corpus and Discourse

Corpus linguistics provides the methodology to extract meaning from texts. Taking as its starting point the fact that language is not a mirror of reality but lets us share what we know, believe and think about reality, it focuses on language as a social phenomenon, and makes visible the attitudes and beliefs expressed by the members of a discourse community.

Consisting of both spoken and written language, discourse always has historical, social, functional, and regional dimensions. Discourse can be monolingual or multilingual, interconnected by translations. Discourse is where language and social studies meet.

The *Corpus and Discourse* series consists of two strands. The first, *Research in Corpus and Discourse*, features innovative contributions to various aspects of corpus linguistics and a wide range of applications, from language technology via the teaching of a second language to a history of mentalities. The second strand, *Studies in Corpus and Discourse,* is comprised of key texts bridging the gap between social studies and linguistics. Although equally academically rigorous, this strand will be aimed at a wider audience of academics and postgraduate students working in both disciplines.

Research in Corpus and Discourse

Conversation in Context
A Corpus-driven Approach
With a preface by Michael McCarthy
Christoph Rühlemann

Corpus-Based Approaches to English Language Teaching
Edited by Mari Carmen Campoy, Begona Bellés-Fortuno and Mª Lluïsa Gea-Valor

Corpus Linguistics and World Englishes
An Analysis of Xhosa English
Vivian de Klerk

Evaluation and Stance in War News
A Linguistic Analysis of American, British and Italian television news reporting of the 2003 Iraqi war
Edited by Louann Haarman and Linda Lombardo

Evaluation in Media Discourse
Analysis of a Newspaper Corpus
Monika Bednarek

Historical Corpus Stylistics
Media, Technology and Change
Patrick Studer

Idioms and Collocations
Corpus-based Linguistic and Lexicographic Studies
Edited by Christiane Fellbaum

Meaningful Texts
The Extraction of Semantic Information from Monolingual and Multilingual Corpora
Edited by Geoff Barnbrook, Pernilla Danielsson and Michaela Mahlberg

Rethinking Idiomaticity
A Usage-based Approach
Stefanie Wulff

Working with Spanish Corpora
Edited by Giovanni Parodi

Studies in Corpus and Discourse

Corpus Linguistics in Literary Analysis
Jane Austen and her Contemporaries
Bettina Fischer-Starcke

English Collocation Studies
The OSTI Report
With an introduction by Wolfgang Teubert
John Sinclair, Susan Jones and Robert Daley
Edited by Ramesh Krishnamurthy

Text, Discourse, and Corpora. Theory and Analysis
With an introduction by John Sinclair
Michael Hoey, Michaela Mahlberg, Michael Stubbs and Wolfgang Teubert

This page intentionally left blank

Corpus Linguistics in Literary Analysis

Jane Austen and her Contemporaries

Bettina Fischer-Starcke

continuum

Continuum International Publishing Group

The Tower Building	80 Maiden Lane, Suite 704
11 York Road	New York
London SE1 7NX	NY 10038

www.continuumbooks.com

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library.

ISBN: 978-1-8470-6437-0 (hardcover)

Library of Congress Cataloging-in-Publication Data
A catalog record for this book is available from the Library of Congress.

Typeset by Newgen Imaging Systems Pvt Ltd, Chennai, India
Printed and bound in India by Replika Press Pvt Ltd

For my family

This page intentionally left blank

Contents

List of Tables

List of Figures

Preface

Stylistics has been a source of interest to me since writing my MA thesis on the language of Janet Frame, a New Zealand author. So when choosing a topic for my PhD thesis, corpus stylistics was my immediate choice, and I decided to look at Jane Austen's novels to see whether using corpus linguistic techniques in the analysis of literature could give new insight into already thoroughly analysed texts. I selected three main analytic techniques and found that they were indeed highly successful in revealing new literary meanings of the data.

This book is an adaptation of my PhD thesis which I submitted at the University of Trier, Germany, in 2007. The book presents both the theoretical contexts in which the analyses of my thesis are embedded and the analyses themselves. However, the original appendix of my thesis with its more than 100 pages of data could not be reproduced in this book. This would have included the complete lists of keywords, the complete sets of concordance lines, all CBDF-values and so on, that are mentioned in this book, and I would be happy to provide this data to anyone interested in it. Also, the corpora that I use for the analyses do not come with this book.

Neither the book nor the thesis could have been written without the moral and practical support of a number of people. I am grateful to Michael Stubbs for helpful comments and conversations both on my PhD thesis and on earlier versions of this book. I am also grateful to Katrin Oltmann who gave me permission to use her corpus of literature contemporary to Austen, *ContempLit* in this book, and to Isabel Barth who gave me permission to use her software *Word-Distribution*. Thank you also to Christian Fischer who helped me with the statistics and provided moral support whenever I needed it, to Anna Maria Duplang, Bernd Elzer, Clare Fielder, Kieran O'Halloran, Sabine Starcke and Kurt Ubelhoer for proof reading and helpful comments.

The meaning of a word is its use in the language.

Ludwig Wittgenstein

Chapter 1

Introduction

Stylistics is the linguistic analysis of literary texts. Corpus linguistics is the electronic analysis of language data. The combination of both disciplines is corpus stylistics, the linguistic analysis of electronically stored literary texts.

Corpus stylistics pursues two goals:

1. to study how meaning is encoded in language and to develop appropriate working techniques to decode those meanings, and
2. to study the literary meanings of texts.

The first goal is a traditional goal in linguistics, which includes gaining knowledge of analytic techniques. The second goal is a traditional goal in literary studies, which includes gaining knowledge of the meanings of a specific text or body of texts.

In corpus stylistics, the use of corpus linguistic techniques and the goals of stylistics complement each other as both disciplines decode linguistic patterns and their meanings in texts. In both disciplines, the knowledge gained is used to generate a more general understanding of, for instance, literary meanings or the organization of language. The two disciplines therefore complement each other when they are combined to form corpus stylistics.

This combination of the two disciplines is the reason for the great analytic potential of corpus stylistics. It allows for decoding meanings of literary texts that cannot be detected either by intuitive techniques as in literary studies or with the necessary restriction to short texts or text extracts as in traditional stylistics. Corpus linguistic techniques allow (1) a systematic and detailed analysis of large quantities of language data for lexical and/or grammatical patterns and (2) to subsequently decode the meanings of these patterns. These patterns are not intuitively recognizable because of the very size of the data. In the analyses in Chapters 5 to 7 in this book, the data is comprised of between 77,000 and 4,370,000 tokens.

1.1 Stylistics and style

Widdowson defines stylistics as the 'study of literary discourse from a linguistic orientation' (1975: 3) which 'treats literature as *discourse*' (6). Toolan supports this view by saying that stylistics is 'the study of language *in* literature' (1998: viii) and that it is therefore part of linguistics. By analysing the linguistic patterns of a text, it gives answers to questions such as how literary effects are encoded in language. And Weber is even more pointed by demanding that it answers questions such as 'what is literature? How does literary discourse differ from other discourse types? How do we read and interpret literary texts?' (1996: 1).

The definition of stylistics in this book is closely related to the views quoted above. Here, stylistics is defined as the linguistic analysis of literary texts and therefore as a linguistic discipline. Its goal is to decode literary meanings and structural features of literary texts by identifying linguistic patterns and their functions in the texts. Consequently, the term *style* means lexical and grammatical patterns in a text that contribute to its meaning. This ties in with Fowler (ed.) who says that

> [w]e must assume that all texts manifest style, for style is a standard feature of all language (. . .). [S]tyle is a manner of expression, describable in linguistic terms, justifiable and valuable in respect of non-linguistic factors. (. . .) it is a facet of language. (1987: 236)

This rather general definition of style is further explained and expanded in Chapter 3 *Language and meaning*.

The definition of style above emphasizes that stylistics is a linguistic discipline. However, apart from the goals of linguistics, namely to gain knowledge of how meaning is encoded in language and of the meaning itself, it also pursues the goal of literary studies, namely to gain knowledge of the literary meanings of a specific text. In the following, the relationship between style, meaning in language and the analytic techniques of both linguistics and literary studies are discussed, in order to show how the goals of linguistics and literary studies can be combined, and to explain how their goals are pursued and related to each other in stylistics.

Linguistics analyses language systematically to gain knowledge of language patterns either in a specific text or in language in general. Depending on the question set, the textual basis of an analysis is either a text or a corpus, that is, a compilation of texts or text fragments in electronic form. In most linguistic analyses, the textual data is non-fiction and non-literary.

One basic assumption of linguistic analyses is that the linguistic form of the data indicates its meaning. Corpus linguistics further assumes a correlation between the frequency of a pattern and its significance in the data. Frequent linguistic patterns have significance for either the content of the data or its structural organization (Teubert 2005). The frequency with which a feature occurs therefore influences its qualitative analysis. This is further explained in Chapter 2 *Goals, techniques, principles.*

Literary studies analyse the meanings of literary texts by looking at their language and at extratextual features. This is subsumed under 'criticism' which is '[t]he conscious evaluation or appreciation of a work of art, either according to the critic's personal taste or according to some accepted aesthetic ideas' (Shipley ed. 1970: 66). Bressler, quoting the nineteenth-century critic Matthew Arnold as evidence for his proposition, further defines literary criticism as 'a disciplined activity that attempts to describe, study, analyze, justify, interpret, and evaluate a work of art' (2003: 4f.). And he goes on to say that 'this discipline attempts to formulate aesthetic and methodological principles on which the critic can evaluate a text.' Eagleton goes still further in his definition of 'literary theory' and says that it 'is less an object of intellectual enquiry in its own right than a particular perspective in which to view the history of our times' (1983: 195). This is because

> any body of theory concerned with human meaning, value, language, feeling and experience will inevitably engage with broader, deeper beliefs about the nature of human individuals and society, problems of power and sexuality, interpretations of past history, versions of the present and hope for the future. (195)

The definition and the exact nature of literary criticism therefore changes in the course of time as different social and political conditions prevail.

The uniqueness of the linguistic style of a text, as manifest in its lexical, phraseological and grammatical patterns, is of less importance for the analysis of a text in literary studies than in linguistics. In literary studies, language is mainly relevant as a criterion for literariness and the literary meanings of the text. This means that linguistic deviations from language norms are one criterion for culturally valued literature, and it occasionally distracts attention from a rather banal content of a text (Cook 1986: 150). The analysis of texts frequently conforms to the principles of classical rhetoric and is intuitive by often following a particular school of thought, for example Reception Theory or the New Criticism.

Basic questions that are addressed in literary criticism are concerned with the philosophical, psychological, functional and descriptive nature of the text itself:

- Does the text have only one correct meaning?
- Is a text always didactic; that is, must a reader learn something from every text?
- Does a text affect each reader in the same way?
- How is a text influenced by the culture of its author and the culture in which it is written?
- Can a text become a catalyst for change in a given culture? (Bressler 2003: 5)

Linguistic features that are not prominent in the text are frequently not recognized and are therefore only rarely analysed in literary studies.

Both linguistics and literary studies analyse texts and their meanings, but they differ in their methods of analysis and in their choice of texts. Literary studies are restricted to literary texts and the analysis of their meanings. The question what 'literature' actually is, has not been answered definitively.

Shipley, for example, says that literature is '[w]ritten productions as a collective body. The total preserved writings belonging to a given language or people; that part which is notable for literary form or expression, "belles lettres" (. . .)' (ed. 1970: 183f.). Eagleton (1983), however, does not offer a single definition of 'literature' in his chapter entitled 'Introduction: What is Literature', but instead shows that its definition has changed in the course of time and depends on the school of thought proposing the definition. He emphasizes that linguistic, philosophical, social and political factors influence whether a piece of writing is accepted as literature or not.

In this book, the term 'literature' follows the rather general definition of the *Oxford English Dictionary* (1989) as a 'literary work or production (. . .) the realm of letters' or

> [l]iterary productions as a whole; the body of writings produced in a particular country or period, or in the world in general. Now also in a more restricted sense, applied to writing which has claim to consideration on the ground of beauty of form or emotional effect

in order to cover the diverse views on the concept.

Linguistics usually analyses non-fiction texts, text fragments and collections of texts and the functions of patterns in the language of the data. The criteria for the selection of the data in linguistics are often functional and are based on the research question. In literary studies, it is often the literary value of a text which functions as a criterion for it to be selected for an analysis. Consequently, literary studies gain knowledge mainly of a specific text or a literary period. Linguistics, on the other hand, gains knowledge of the specific text or corpus and of the language system with its mechanisms for encoding meaning.

The two disciplines also differ in their definitions of style. While linguistics perceives 'style as choice' (de Beaugrande 1993: n.p.) of an author, literary studies perceive 'style as ornamentation' (n.p.) of a text. In literary studies, style is an aesthetic choice which makes a text either literary or non-literary and which 'serves to mark the critic's approval or disapproval of the quality of a writing' (Shipley ed. 1970: 314).

> A style is a manner of expression, describable in linguistic terms, justifiable and valuable in respect of non-linguistic factors. (. . .) it is a facet of language (. . .) that is given significance by personal or cultural, rather than verbal, qualities. (Fowler 1987: 236f.)

It is an exclusive criterion.

In linguistics, and therefore also in stylistics, a text 'represents the results of a complicated selection process, and each selection has meaning by virtue of all other selections which might have been made, but have been rejected' (Sinclair 1965: 76f.). The individual style of a text is the author's or speaker's choice and its meaning derives precisely from the fact that it was the sender's choice. This means that style is not an exclusive, but a describable criterion in linguistics which allows for determining the degree to which it is specific to a sender or to which it conforms to conventions. The sender's choice of language is not evaluated; it is merely described.

This goal of a linguistic analysis is modified when the objects of an analysis are large and representative corpora. A sender's individual choices are no longer of interest, but instead patterns in the general usage of language of a large number of senders are identified and described. A comparison between the data of various senders leads to the identification of intertextual patterns, the functions of which are subsequently decoded. The main objective of the analysis is to decode the significance of these patterns for the content and the structure of the data. This is the same objective as in the analysis of a text. However, the analyses of a text and a corpus differ in the quantities of the data.

Stylistics combines the data of literary studies, that is, literary texts, with the analytic techniques and objectives of linguistics. It thereby fills a gap within linguistics, since stylistics is the only linguistic discipline which allows the analysis of literary texts and their literary meanings by way of linguistic techniques. It holds this singular position despite the fact that literature is made of language. But a study of language which is unable to analyse one text type, especially a culturally significant one, is incomplete (Sinclair 1975, 1982). Thus, a further goal of linguistics should be to be able to draw conclusions about the language, the structure and the meanings of literary texts by means of linguistic analytic techniques.

Stylistics is based on the assumption that meaning in language is a linguistic phenomenon which can be decoded by way of linguistic analyses. Corpus stylistics specifies this assumption by choosing corpus linguistic techniques for the analyses. Unlike traditional stylistics, which can analyse only short texts or extracts from longer texts, corpus stylistics also permits the analysis of longer works such as novels. It utilizes software to aid in identifying language patterns which are objectively in the data. This provides the linguist with detailed and neutral insights into the data, which are independent of, for example, previous knowledge of the reception of the work or of genre conventions. The analysis is text-internal and gives a new perspective on the data, so that the researcher can detect new meanings even in a widely discussed text. The detailed linguistic analysis permits detecting meanings, which are virtually invisible in an intuitive approach to the data as in literary studies.

The corpus stylistic focus on the most frequent linguistic features, however, precludes it from detecting infrequent features. This is the case even though these features may be foregrounded and often contribute to a text's meaning, especially in literary texts. They are identified in literary critical or in traditional stylistic analyses.

In corpus stylistics, conclusions about the meanings of the data are based on the assumption that form and meaning correlate. However, this correlation is neither obvious nor stable. One language pattern can have different meanings in different texts and contexts so that generalizations about the meanings of language patterns are valid only for the data analysed or similar data. Furthermore, different linguists might interpret the same pattern differently, since an interpretation is always a subjective process (cf. Chapter 2 *Goals, techniques, principles*). The patterns, however, are objective features of the data. The aim of stylistics is therefore to make the connection between language structures and their meanings explicit and, by doing so, to reveal the linguistic basis of a literary interpretation.

Using a corpus stylistic approach to the analysis of literary texts is an explicitly quantitative approach. It makes the quantitative element of many stylistic studies, which is frequently implicit (Fowler 1987: 237f.), explicit.

There are also critics who oppose the use of corpus linguistic techniques in the analysis of literature. One of them is Miall (1995) who argues that analysing a text electronically results in a loss of an analyst's individual perception of the text, since this perception is based on personal experience and individual knowledge of the text. The use of software counteracts the simulation of the reading process and therefore, hampers the understanding of a text. Yet, it is precisely this loss of individuality, that is, a reader's personal textual competence and experiences, that corpus stylistics aims for in the generation of the data that is analysed, as this is what contributes to the intersubjectivity of an analysis. The generation of frequency data as a basis of the analysis of literary meanings is as much stripped off an analyst's individual choices and perceptions as possible. The ensuing interpretative process of the data, however, is necessarily subjective (cf. Chapter 2 *Goals, techniques, principles*).

A further point of criticism of corpus stylistics is that it disregards literary elements of texts, such as metaphors, in its analyses (van Peer 1989). However, unlike van Peer (1989), corpus stylisticians do not perceive lexis and grammar to be 'on the lower levels of linguistic organization' in comparison to 'figurative meanings' (301) in literary texts. On the contrary, lexical and grammatical patterns contribute to the literary character of a text and analysing them contributes to decoding meanings in literary texts.

Even though stylistics uses the same data for its analyses as literary studies, namely literary texts, it does not aim at replacing literary studies. A collaboration of the two disciplines would generate deeper insights into the texts than can be gained by strictly separate analyses. The analyses of this book show that the two disciplines provide different insights into the same texts. A co-operation between the two disciplines could therefore result in more detailed and more extensive knowledge of a text than by keeping the disciplines strictly separate. The fact that this co-operation does not exist at present is no reason for doubting its benefit in understanding the various shades of meanings of a text and for ultimately rejecting it.

In the following analyses, reference to findings by literary critics is given only when it seems appropriate. The main emphasis in this book is on the literary findings from the analyses presented here. This does not, however, mean that I am not aware of the comprehensive discussion and seemingly exhaustive research on Austen's novels by literary critics. I use the literary critical studies to complement my own findings and, in turn, complement the literary findings with my corpus stylistic insights.

1.2 The data

The textual bases of the analyses in this book are Jane Austen's novel *Northanger Abbey* (henceforth *NA*) and the corpus *Austen* which consists of Austen's six novels *Emma, Mansfield Park, Northanger Abbey, Persuasion, Pride and Prejudice* and *Sense and Sensibility*. The corpus *ContempLit,* which is analysed in Chapter 6 *Phraseology,* represents the literary language contemporary to Jane Austen (cf. Chapter 2 *Goals, techniques, principles* for information on the compilation of the corpora). *NA* (1818) is both Austen's first completed and also her last, posthumously published novel.

The reasons for choosing Austen's text *NA* and the corpus *Austen* as data for the analyses in this book are practical ones. First, their original publications date back to about 200 years ago. This means that the texts do not fall under the copyright and it is therefore legal to store and analyse them electronically. The same is true for the texts comprising the corpora *ContempLit* and *Gothic,* two of the reference corpora for the analyses in Chapters 5 to 7 in this book. Legal access to electronically stored language data is one of the necessary preconditions for corpus linguistic and corpus stylistic analyses.

Second, *NA* and Austen's other novels have been intensively discussed and analysed over the past about 200 years. Jane Austen is one of the most widely read British classical authors, her novels are known worldwide and are still popular today. Evidence for this includes the various film adaptations of Austen's novels since the 1990s, for instance *Sense and Sensibility* (1995), *Emma* (1996) and *Pride and Prejudice* (2005). The novels' popularity is based on their ironic, humorous and light-hearted tone and on their ostensibly simple and romantic contents.

NA is one of Austen's least discussed works – even though there are still a significant number of writings on the novel. The query *Northanger* and *Abbey* in the database of the *Modern Language Association* (*MLA*) results in 191 hits (26 June 2009). Austen's most popular and most widely discussed novel, *Pride and Prejudice,* only results in more than twice the hits in the *MLA* database (459, query of *Pride* and *Prejudice,* 26 June 2009). In addition, there are numerous writings on Austen's novels in publications on the author's complete oeuvre, so that a query of *Austen* and *novels* produces 2588 hits in the *MLA* database (26 June 2009).

Analyses of Austen's novels are not restricted to literary studies only, but also linguists have examined her language. Chapmann (1933), for example, discusses 'Miss Austen's English', Phillipps (1970) *Jane Austen's English*, Page (1972) *The Language of Jane Austen*, Tave (1973) *Some Words of Jane Austen*, Stokes (1991) also looks at *The Language of Jane Austen*, and Burrows' (1987) study of Austen's characters' idiolects is still influential in linguistics.

Systematic linguistic research on Austen's language, such as Burrows' (1987), is still the exception, however. Most published secondary literature on the author and her language is situated within literary studies and is therefore intuitive in its analyses. Nevertheless, because literary critics have examined Austen's novels closely in the past and still do so today, it might, at first sight, seem unlikely that new insights into their contents and language could be gained. This makes the novels ideal data for evaluating the effectiveness of corpus stylistic analyses, since a comparison of findings from the following analyses with those already published in secondary literature allows one to evaluate the novelty and innovativeness of findings. It is possible to see whether corpus stylistic analyses produce new insights into the novels, or whether they only replicate literary critics' findings.

Corpus stylistic analyses are only successful when they produce new findings on the data (cf. section 1.3. for further explanations of this claim). Consequently, I emphasize new findings on the data analysed in this book, but only rarely previous findings by literary critics. This shows that many observations made in this book do not seem to have been made previously.

The present book takes Burrows (1987) as a model and demonstrates a systematic and comprehensive linguistic analysis of the data (cf. Chapter 3 *Language and meaning* for a discussion of Burrows 1987). The analyses in this book examine the different sets of data, *NA*, *Austen* and in one chapter *ContempLit*, by using different analytic techniques. Consequently, the different analyses investigate linguistic units on different hierarchical levels in language, namely lexis, phraseology, text parts and text. Analysing these different linguistic units creates a comprehensive picture of literary meanings in the data and of the effectiveness of the different analytic techniques for the different sets of data. The analytic techniques that are mainly used in the analyses are extracting keywords and frequent phrases from the data, generating distribution diagrams of lexis and analysing concordance lines. This range of analyses goes beyond that by Burrows' (1987).

1.3 The potential and goals of corpus stylistic analyses

In corpus stylistics, we can use the same methods to analyse both an individual text and a corpus. This allows us

- to develop analytic techniques for investigating various research questions,
- to evaluate the success of different research techniques for different sets of data, and
- to gain new literary and structural insights into the data.

All this is demonstrated in the analyses later in this book.

The present work is broader in its range of analyses than previous corpus stylistic studies which usually have one textual basis examined using one analytic technique (cf. Chapter 3 *Language and meaning*). In this book, the potential of various techniques are explored for various sets of data, thereby filling a gap in corpus stylistics and stylistics in general by systematically examining which analytic techniques generate (1) the most and (2) so far unknown insights into a text and a corpus.

This book also enlarges the scope of data that is analysed in comparison to previous publications. In stylistics, including corpus stylistics, the objects of analyses have mostly been short texts, such as poems or extracts from longer works (e.g. Louw 1993, O'Halloran 2007b). The possibility of analysing longer texts in a corpus stylistic analysis has only recently been put into practice (e.g. Stubbs 2005, Starcke 2006, Fischer-Starcke 2009b) and is further developed in the present volume by systematically analysing a text and a corpus. The analyses result in literary and structural knowledge of *NA* and Jane Austen's oeuvre in general. In addition, the analysis of the corpus *ContempLit* in Chapter 6 *Phraseology*, and its use as a reference corpus in other analyses of this book, gives insight into general literary language contemporary with Austen. This enlarges the scope and the quantity of data of corpus stylistic research.

Apart from gaining literary insights into the data, a second goal of this book is to evaluate the use of corpus linguistic techniques in the analysis of literature. The basis for this evaluation is a comparison between the findings from the analyses in this book and interpretations of the data published as secondary literature. Since Austen's texts are part of the literary canon and have been studied intensively over the centuries (as shown earlier), they are particularly well-suited for this task.