

A Corpus Linguistic Approach
to Literary Language and Characterization

Linguistic Approaches to Literature (LAL)

Linguistic Approaches to Literature (LAL) provides an international forum for researchers who believe that the application of linguistic methods leads to a deeper and more far-reaching understanding of many aspects of literature. The emphasis will be on pragmatic approaches intersecting with areas such as experimental psychology, psycholinguistics, computational linguistics, cognitive linguistics, stylistics, discourse analysis, sociolinguistics, rhetoric, and philosophy.

For an overview of all books published in this series, please see
<http://benjamins.com/catalog/lal>

Editors

Sonia Zyngier
Federal University of Rio de Janeiro

Joanna Gavins
University of Sheffield

Advisory Editorial Board

Douglas Biber
Northern Arizona University

Arthur C. Graesser
University of Memphis

Keith Oatley
University of Toronto

Marisa Bortolussi
University of Alberta

Frank Hakemulder
Utrecht University

Willie van Peer
University of München

Donald C. Freeman
University of Southern
California

Geoff M. Hall
University of Wales, Swansea

Yeshayahu Shen
Tel Aviv University

Richard Gerrig
Stony Brook University

David L. Hoover
New York University

Mick Short
Lancaster University

Raymond W. Gibbs, Jr.
University of California,
Santa Cruz

Don Kuiken
University of Alberta

Michael Toolan
University of Birmingham

Rachel Giora
Tel Aviv University

Geoffrey N. Leech
Lancaster University

Reuven Tsur
Tel Aviv University

Paisley Livingston
University of Copenhagen

Peter Verdonk
University of Amsterdam

Volume 18

A Corpus Linguistic Approach to Literary Language and Characterization
Virginia Woolf's *The Waves*
by Giuseppina Balossi

A Corpus Linguistic Approach to Literary Language and Characterization

Virginia Woolf's *The Waves*

Giuseppina Balossi

John Benjamins Publishing Company
Amsterdam / Philadelphia



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

Library of Congress Cataloging-in-Publication Data

Balossi, Giuseppina.

A corpus linguistic approach to literary language and characterization: Virginia Woolf's
The Waves / Giuseppina Balossi.

p. cm. (Linguistic Approaches to Literature, ISSN 1569-3112; v. 18)

Includes bibliographical references and index.

1. Language and languages--Style--Psychological aspects. 2. Cognitive grammar.
3. Discourse analysis, Literary. 4. Woolf, Virginia, 1882-1941. *The Waves*.
5. Psychological fiction, English--History and criticism. I. Title.

P301.5.P75B36 2014

823'.912--dc23

2014001873

ISBN 978 90 272 3407 0 (Hb; alk. paper)

ISBN 978 90 272 7042 9 (Eb)

© 2014 – John Benjamins B.V.

No part of this book may be reproduced in any form, by print, photoprint, microfilm, or any other means, without written permission from the publisher.

John Benjamins Publishing Co. · P.O. Box 36224 · 1020 ME Amsterdam · The Netherlands
John Benjamins North America · P.O. Box 27519 · Philadelphia PA 19118-0519 · USA

Cui, nisi tibi?

*This book is dedicated to Antonio Nogara,
whose academic example, teaching and deep sensitivity
made me a different person*

Table of contents

Acknowledgements	XI
Foreword	XIII
List of conventions	XV
List of figures and tables	XVII
List of appendixes	XIX
List of concordances	XXI
CHAPTER 1	
Introduction	1
1.1 The issues at stake	1
1.2 The design of this book	2
CHAPTER 2	
Virginia Woolf's <i>The Waves</i>	5
2.1 Virginia Woolf and the modernist character	5
2.2 Introduction to <i>The Waves</i>	6
2.2.1 Conception	7
2.2.2 Structure	8
2.2.3 Method of characterization	9
2.2.4 Debate over characterization	11
2.2.5 Studies on lexical patterns	13
2.2.6 Psychoanalytic approaches to character	15
CHAPTER 3	
Literature review	19
3.1 The state of affairs in literary characterization	19
3.1.1 Characters as people and as textual constructs	19
3.1.2 Character typologies	22
3.2 Psychological concepts in the perception of personality	24
3.2.1 Prior knowledge and schema theory	25
3.2.2 History of schema theory	26

3.3	Stylistic approaches to characterization	28
3.3.1	Culpeper's model for characterization: Top-down and bottom-up	29
3.3.2	Cognitive metaphor, mind style and characterization	33
3.4	Studies of language and personality	35
3.5	Conclusion	39
CHAPTER 4		
	Corpus approaches to the study of language and literature	41
4.1	Introduction	41
4.2	Corpus annotation	44
4.3	Techniques employed in corpus analyses	45
4.4	Corpus annotation of literary texts	49
4.4.1	Quantitative stylistic approaches to literary texts	51
4.4.2	Authorial style	52
4.4.3	Characterization	55
4.4.3.1	Characterization in <i>The Waves</i>	56
CHAPTER 5		
	Methodology	59
5.1	Introduction	59
5.2	The narrative framework of <i>The Waves</i>	59
5.2.1	Synchronic and diachronic structure	60
5.2.2	<i>The Waves</i> as e-text: Text annotation and text division	64
5.3	Wmatrix: A software tool for corpus analysis and comparison	66
5.3.1	The CLAWS and USAS taggers and their tagsets	67
5.4	<i>The Waves</i> through Wmatrix	70
5.4.1	Adaptation and revision of POS and USAS tagsets	72
5.4.2	Data layout in Microsoft Excel	76
5.4.3	Post-editing of the soliloquy text	78
5.4.4	Token count of the soliloquy text	81
CHAPTER 6		
	Character differentiation through word-classes	83
6.1	Introduction	83
6.2	Word-class and characters' language	83
6.2.1	Word-classes in the whole soliloquy text and in each character	85
6.2.2	Data comparison and treatment of the LL statistics	89
6.2.3	Characters' word-classes in each phase of life: Significant differences relative to each character	93
a.	Susan	96
b.	Jinny	96

- c. Rhoda 97
- d. Louis 97
- e. Neville 98
- f. Bernard 98
- 6.2.4 Bernard's word-classes in the final soliloquy 99
- 6.3 Summary 102

CHAPTER 7

Character differentiation: Semantic fields 107

- 7.1 Introduction 107
- 7.2 Data comparison and treatment of the LL statistics 108
- 7.3 The characters' schema of gender dichotomy 112
- 7.4 The female characters 114
 - 7.4.1 Susan 115
 - a. The Natural World 116
 - b. The House 120
 - c. Home-making activities 121
 - d. Family and Motherhood 122
 - e. Time 123
 - f. The Senses 125
 - 7.4.2 Jinny 127
 - a. The Body 127
 - b. Colours 130
 - c. Movement 131
 - 7.4.3 Rhoda 133
 - a. The Natural World 134
 - b. Movement 136
 - c. Emotions 139
 - d. Shapes 140
 - e. Colours 142
- 7.5 The female characters' semantic fields and personality traits 146
- 7.6 The male characters 150
 - 7.6.1 Louis 150
 - a. Geographical names 151
 - b. People and Relationships 152
 - c. The Natural World 158
 - d. History and Literature 159
 - 7.6.2 Neville 161
 - a. Literature and Philosophy 161
 - b. Academic Interests 165

7.6.3	Bernard	166
	a. Cognition and Abstraction	167
	b. Intellectual Interests	172
7.7	The male characters' semantic fields and personality traits	175
7.8	Bernard's final summing-up and his dramatic role	178
7.9	Conclusion: The characters' semantic fields and their personality traits	182
7.10	Percival: The other-presented character	183
CHAPTER 8		
	Conclusion	185
8.1	Achievements	185
8.2	Characterization in this study	186
8.3	Reading through the computer	187
8.4	Authorial style <i>versus</i> character voice	188
8.5	Further directions	190
	References	193
	Websites	211
	Appendixes	213
	Author index	273
	Subject index	275

Acknowledgements

I should like to express my gratitude to a number of people who have contributed to the ‘birth’ of this book. Special thanks are due to Olivia Fialho, whom I met at the annual conference of the Poetics and Linguistics Association (PALA) in Genoa in 2010. It was through her that my research reached Sonia Zyngier. And it is thanks to Sonia’s support and patience that I had the time to bring this work to its completion. I am especially grateful to Elena Semino, who supervised both my MA and Ph.D. theses on characterization at Lancaster University. I would also like to express my gratitude to Jonathan Culpeper, who supervised me while Elena was on sabbatical. Without their help and encouragement, I could not have gone this far.

I would also like to acknowledge the help of several academic colleagues I met at Lancaster University who have now taken different academic paths. Nicholas Smith and Amy Wang accompanied me throughout, Nick helping with the statistics and Amy giving much valuable advice. Michael Oakes also gave me kind, helpful support with the statistical part of the work, always ready to settle my doubts. Izumi Tanaka very kindly helped with the tagging of the text and Paul Rayson with detailed information about Wmatrix whenever I needed it.

I am also grateful to friends from other walks of life, especially Mike Cardwell, who, besides giving moral support, also spent a huge amount of time straightening out the layout of the data and providing practical advice on how to deal with the very large datasets. I am also grateful to Adrian Paul and Christina Chow, Andrew Wilson, Emily Bianchi-Bazzi, Giangio Gilardi, Massimo Vaghi, and my student Loris Molinari, who have either helped with some parts of the book or provided general and specific comments.

Last but not least, I am deeply grateful to John Heywood, who first proof-read the Ph.D. thesis from which this book developed and then supported me throughout the writing process from the first draft to the very last. Not only did he read and re-read it through carefully, but he also made several important editorial suggestions, especially regarding metaphor analysis. He commented thoroughly and patiently on the style and presentation of each chapter, and helped with the bibliography and index. This long collaboration has also contributed to our deep affective bond. It has been an extraordinary experience to work with him.

This book began while my father was still with us and I am aware of how proud he would have been of its completion, as my mother is today. I would also like to thank the assistance I had from my sisters, Mari and Tina, who did more than their fair share during my withdrawal from family life.

Finally, I owe my deepest gratitude to my husband Mario and my daughter Fabiola, who have kept me on track throughout the project and had often had to do without me. To them I offer the seeds I have sown.

Foreword

“Language use is one of the most mysterious products of the human mind, and literature is probably the most enigmatic form of language use” (van Peer 2011: 1). Before the arrival of structuralism, the study of literature tended to be based on free, intuitive interpretation. Today, with the advent of the computer, more systematic studies can be carried out. Will they offer possibilities for solving the enigma of language in literature?

While computer-assisted methods of enquiry have been familiar for some time now across a range of disciplines, such as psychology and other social sciences (Murphy 2010; Pennebaker 2011), the field of literary criticism has resisted the lure of the computer (Gottschall 2008). Corns (1991: 129) observed that computer-aided literary analyses were not being paid sufficient attention and unfairly associated with poor achievements, whereas stylometric studies (Burrows 1987) formed a distinct strand kept quite separate from the main literary critical tradition. Louw (1993: 152–176; 1997: 240–251) was one of the first scholars to appreciate the potential value and effectiveness of corpus linguistics in the analysis of literary texts. Now computer-aided studies have been accepted as a major research paradigm (Stockwell 2008: 351–363) that can contribute a great deal to developing fresh critical insights into the stylistic qualities of texts, including literary ones (Deignan 2005; Zyngier 2008: 169–190; Lüdeling and Kytö 2008; 2009; Starcke 2010; Viana *et al.* 2011).

While we may all agree that the computer will never be able to understand and appreciate a literary text in the ways the human mind does (Graesser *et al.* 2011: 24–33), we must acknowledge that corpus methodologies open up, in Leech’s words (1992: 106), “a new way of thinking about language.” Using the computer leads to discoveries that cannot be seen with the naked eye. A computer-aided approach can also offer new starting points for investigation. It has proven to be a strong research method anchored firmly in theoretical frameworks which can be combined with other qualitative methodologies (Leech and Fligelstone 1992: 115–140; Biber *et al.* 1998; Biber 2011: 15–23; Mahlberg and McIntyre 2011: 204–227; McIntyre 2012: 409–415).

This book attempts to demonstrate how a computer-aided approach can guide us through the investigation of language and characterization in Virginia Woolf’s *The Waves* (1931). The focus is on whether characters can be differentiated

through examining the word-classes and semantic categories associated with them. Although these corpus-based findings may sometimes be problematic, it is argued that they produce new and detailed insights into both the diachronic development of the characters and the novel as a whole. While the investigation of the word classes conducted to assess the presence of individuality in the characters' voices or idiolects (Stewart 2003: 129–138) is purely quantitative, the semantic analysis is more finely articulated. By applying Culpeper's (2001) cognitive model of characterization, the semantic fields distinguishing each character to a statistically significant extent are also subjected to qualitative analysis, allowing inferences to be drawn about probable differences in personality traits and idiosyncratic mind styles (Fowler 1977: 103–113; Semino 2007: 153–203).

The book is conceived in such a way as to bring both theory and practice together. Background information is provided on the theoretical aspects of the debate over characterization in *The Waves*, followed by a broad survey of studies of characterization from their outset to the present state of the art, and of proposals for viable ways to study fictional people's language and their personalities (Eder *et al.* 2010). The study also provides and puts into practice a detailed methodology for the analysis of character. The corpus-based approach, as suggested here, is also a methodology applicable to the contrastive investigation of language, and of language and personality, in different types of discourses encompassing both fictional *personae* and real people.

List of conventions

ACAMRIT	Automatic Content Analysis of Market Research Interview Transcripts
ACASD	Automatic Content Analysis of Spoken Discourse
BNC	British National Corpus
CAMET	Computer Archive of Modern English Texts
CLAWS	Constituent Likelihood Automatic Word-tagging System
CMT	Cognitive Metaphor Theory
LL	log-likelihood
LOB	Lancaster/Oslo-Bergen Corpus
MDA	multi-dimensional analysis
MVA	multi-variate analysis
MWE	multi-word expression
OTA	Oxford Text Archive
POS	part-of-speech
REVERE	REVerse Engineering of Requirements project
SEMTAG	Semantic Tagger
SGML	Standard Generalized Mark-up Language
SP	semantic prosody
UCREL	[Lancaster] University Centre for Computer Corpus Research on Language
USAS	UCREL Semantic Analysis System
WD	<i>The Diaries of Virginia Woolf</i>

List of figures and tables

Tables

Table 5.1	The characters' life stages and major events	62
Table 5.2	Text files used in the analysis of the soliloquy part of <i>The Waves</i>	65
Table 5.3	The top semantic levels in the USAS system	69
Table 5.4	The semantic subfields in USAS level 'B: The body and the individual'	70
Table 5.5	The top semantic levels in the modified USAS system	73
Table 5.6	The data layout taken from Susan's speaking section in ALLCHRS POS and USAS dataset	77
Table 5.7	Token size in the soliloquy part of <i>The Waves</i>	82
Table 6.1	Word-class frequency distribution in the soliloquy part of <i>The Waves</i>	85
Table 6.2	Comparison between word-class frequencies in the soliloquy part of each character's spoken text	87
Table 6.3	Sample output of the statistically significant LL values in Susan's word-classes compared with Jinny's word-classes in Speaking Section 2	91
Table 6.4	Sample output of the statistically significant LL values in Susan's word-classes compared with Jinny's word-classes in each phase of life	92
Table 6.5	Bernard's significant word-classes in the final summing-up	99
Table 7.1	Sample analysis of the LL statistics comparing Susan's semantic fields with the other characters' semantic fields	110
Table 7.2	Sample analysis of the frequency distributions of Susan's most overused semantic fields in each phase of life	111
Table 7.3	Susan's most overused semantic fields relative to the other characters	115
Table 7.4	Jinny's most overused semantic fields relative to the other characters	127
Table 7.5	Rhoda's most overused semantic fields relative to the other characters	134
Table 7.6	Louis' most overused semantic fields relative to the other characters	151
Table 7.7	Frequencies of occurrence of names in Louis' semantic field tag Z1.2: Minor characters' personal names	156
Table 7.8	Neville's most overused semantic fields relative to the other characters	161
Table 7.9	Neville – frequencies of occurrence of names in the semantic field tag Z1: Personal names	162
Table 7.10	Bernard's most overused semantic fields relative to the other characters	167
Table 7.11	Bernard's most significant semantic fields in the summing-up	179
Table 7.12	Frequency of occurrence of the most frequent items in L1: Life and living things in Bernard's Section 18 compared to Sections 2–16	180

Figures

Figure 6.1 Each character's corpus size in all speaking sections

86

Note. All Figures and Tables that are too large to fit into the body of the book are listed as Appendixes.

List of appendixes

- Appendix A. Log-likelihood calculator
- Appendix B. UCREL semantic tagset and UCREL modified semantic tagset
- Appendix C. ALLCHRS POS and USAS
- Appendix D. Each character's statistically significant LL values for the word-classes in each speaking section
- Appendix E. Number of statistically significant differences in each character's life stage
- Appendix F. Susan's significant differences relative to each character
- Appendix G. Jinny's significant differences relative to each character
- Appendix H. Rhoda's significant differences relative to each character
- Appendix I. Louis' significant differences relative to each character
- Appendix J. Neville's significant differences relative to each character
- Appendix K. Bernard's significant differences relative to each character
- Appendix L. Each character's semantic fields in all speaking sections
- Appendix M. Each character's overused semantic fields
- Appendix N. Each character's frequency distribution for the most overused semantic fields in each phase of life

