Combining statistical modeling and archival study, *English and Empire* investigates how African diasporic, Chinese, and Indian characters have been voiced in British fiction and drama produced between 1768 and 1929. The analysis connects patterns of linguistic representation to changes in the imperial political economy, to evolving language ideologies that circulate in the Anglophone world, and to shifts in sociocultural anxieties that crosscut race and empire. In carrying out his investigation, David West Brown makes the case for a methodological approach that links the distant (quantitative) and close (qualitative) reading of diverse digital artifacts. Taking an interdisciplinary approach, the book will appeal to a variety of scholars and students including sociolinguists interested in historical language variation, as well as literary scholars interested in postcolonial studies and the digital humanities.

**David West Brown** is Assistant Professor of Linguistics at Marymount University, Virginia. His research examines writing as a social practice, its structures, and its history, covering topics ranging from Singaporean identity performance in online discussion boards, to representations of gender in sports reportage, to the stylistic differences between higher- and lower-scoring essays on educational assessments. He is the author of *In Other Words: Lessons on Grammar, Code-Switching, and Academic Writing* (2009).
STUDIES IN ENGLISH LANGUAGE

General Editor
Merja Kytö (Uppsala University)

Editorial Board
Bas Aarts (University College London)
John Algeo (University of Georgia)
Susan Fitzmaurice (University of Sheffield)
Christian Mair (University of Freiburg)
Charles F. Meyer (University of Massachusetts)

The aim of this series is to provide a framework for original studies of English, both present-day and past. All books are based securely on empirical research, and represent theoretical and descriptive contributions to our knowledge of national and international varieties of English, both written and spoken. The series covers a broad range of topics and approaches, including syntax, phonology, grammar, vocabulary, discourse, pragmatics and sociolinguistics, and is aimed at an international readership.

Already Published in This Series

Christiane Meierkord: Interactions across Englishes: Linguistic Choices in Local and International Contact Situations
Haruko Momma: From Philology to English Studies: Language and Culture in the Nineteenth Century
Raymond Hickey (ed.): Standards of English: Codified Varieties around the World
Benedikt Szmarcsanyi: Grammatical Variation in British English Dialects: A Study in Corpus-Based Dialectometry
Daniel Schreier and Marianne Hundt (eds.): English as a Contact Language
Bas Aarts, Joanne Close, Geoffrey Leech and Sean Wallis (eds.): The Verb Phrase in English: Investigating Recent Language Change with Corpora
Martin Hilpert: Constructional Change in English: Developments in Allomorphy, Word Formation, and Syntax
Jakob R. E. Leimgruber: Singapore English: Structure, Variation, and Usage
Christoph Rühlemann: Narrative in English Conversation: A corpus Analysis of Storytelling
Dagmar Deuber: English in the Caribbean: Variation, Style and Standards in Jamaica and Trinidad
Eva Berlage: Noun Phrase Complexity in English
Nicole Dehé: Parentheticals in Spoken English: The Syntax–Prosody Relation
Jock O. Wong: The Culture of Singapore English
Marianne Hundt: *Late Modern English Syntax*
Irma Taavitsainen, Merja Kytö, Claudia Claridge and Jeremy Smith: *Developments in English: Expanding Electronic Evidence*
Arne Lohmann: *English Co-ordinate Constructions: A Processing Perspective on Constituent Order*
Nuria Yáñez-Bouza: *Grammar, Rhetoric and Usage in English: Preposition Placement 1500–1900*
Anita Auer, Daniel Schreier and Richard J. Watts: *Letter Writing and Language Change*
John Flowerdew and Richard W. Forest: *Signalling Nouns in English: A Corpus-Based Discourse Approach*
Jeffrey P. Williams, Edgar W. Schneider, Peter Trudgill and Daniel Schreier: *Further Studies in the Lesser-Known Varieties of English*
Jack Grieve: *Regional Variation in Written American English*
Douglas Biber and Bethany Gray: *Grammatical Complexity in Academic English: Linguistic Change in Writing*
Gjertrud Flermoen Stenbrenden: *Long-Vowel Shifts in English, c. 1050–1700: Evidence from Spelling*
Zoya G. Proshina and Anna A. Eddy: *Russian English: History, Functions, and Features*
Raymond Hickey (ed.): *Listening to the Past: Audio Records of Accents of English*
Phillip Wallage: *Negation in Early English: Grammatical and Functional Change*
Marianne Hundt, Sandra Mollin and Simone E. Pfenninger: *The Changing English Language: Psycholinguistic Perspectives*
Joanna Kopaczyk and Hans Sauer (eds.): *Binomials in the History of English: Fixed and Flexible*
Alexander Haselow: *Spontaneous Spoken English*
Christina Sanchez-Stockhammer: *English Compounds and Their Spelling*

Earlier titles not listed are also available.
ENGLISH AND EMPIRE

Literary History, Dialect, and the Digital Archive

DAVID WEST BROWN

Marymount University, Virginia
## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>List of Figures</td>
<td>ix</td>
</tr>
<tr>
<td>List of Tables</td>
<td>xv</td>
</tr>
<tr>
<td>Acknowledgments</td>
<td>xvii</td>
</tr>
<tr>
<td>1 Introduction</td>
<td>1</td>
</tr>
<tr>
<td>A Note about Terminology and a Word of Caution</td>
<td>12</td>
</tr>
<tr>
<td>2 Literary Dialect, Race, and Empire</td>
<td>15</td>
</tr>
<tr>
<td>The Distant Reading of Structure</td>
<td>18</td>
</tr>
<tr>
<td>The Close Reading of the “Imperial Archive”</td>
<td>21</td>
</tr>
<tr>
<td>A Few Comments about the Statistical Analysis</td>
<td>28</td>
</tr>
<tr>
<td>Patterns, Their Propagation, and Their Meaning</td>
<td>29</td>
</tr>
<tr>
<td>3 Corpus Design</td>
<td>31</td>
</tr>
<tr>
<td>Data Collection</td>
<td>33</td>
</tr>
<tr>
<td>Data Preparation</td>
<td>36</td>
</tr>
<tr>
<td>The Voicing Imperial Subjects in British Literature (VSIbL) Corpus</td>
<td>40</td>
</tr>
<tr>
<td>Data Coding</td>
<td>42</td>
</tr>
<tr>
<td>Conclusion</td>
<td>51</td>
</tr>
<tr>
<td>4 An Overview of the Data and the Digital Toolkit</td>
<td>53</td>
</tr>
<tr>
<td>Feature Variation</td>
<td>54</td>
</tr>
<tr>
<td>Speaker Variation</td>
<td>72</td>
</tr>
<tr>
<td>Conclusion</td>
<td>103</td>
</tr>
<tr>
<td>5 Case #1: African Diasporic Dialogue</td>
<td>106</td>
</tr>
<tr>
<td>Constituents of African Diasporic Dialogue</td>
<td>108</td>
</tr>
<tr>
<td>Diachronic Trends in African Diasporic Dialogue</td>
<td>121</td>
</tr>
<tr>
<td>Resemblances in African Diasporic Dialogue</td>
<td>132</td>
</tr>
<tr>
<td>Conclusion</td>
<td>142</td>
</tr>
<tr>
<td>6 Case #2: Indian Dialogue</td>
<td>146</td>
</tr>
<tr>
<td>Constituents of Indian Dialogue</td>
<td>149</td>
</tr>
<tr>
<td>Diachronic Trends in Indian Dialogue</td>
<td>157</td>
</tr>
</tbody>
</table>
## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resemblances in Indian Dialogue</td>
<td>161</td>
</tr>
<tr>
<td>Conclusion</td>
<td>187</td>
</tr>
<tr>
<td><strong>7 Case #3: Chinese Dialogue</strong></td>
<td>189</td>
</tr>
<tr>
<td>Constituents of Chinese Dialogue</td>
<td>191</td>
</tr>
<tr>
<td>Diachronic Trends in Chinese Dialogue</td>
<td>201</td>
</tr>
<tr>
<td>Resemblances in Chinese Dialogue</td>
<td>208</td>
</tr>
<tr>
<td>Conclusion</td>
<td>223</td>
</tr>
<tr>
<td><strong>8 The Enduring Power of Mimicry and the Politics of Measurement</strong></td>
<td>226</td>
</tr>
</tbody>
</table>

### Annotated Glossary

- **Appendix A**  Corpus Composition  262
- **Appendix B**  Coding Taxonomy  268
- **Appendix C**  Coding Category Descriptions  275
- **Appendix D**  Feature Tables: All Dialogue  288
- **Appendix E**  Feature Tables: African Diasporic Dialogue  296
- **Appendix F**  Feature Tables: Indian Dialogue  304
- **Appendix G**  Feature Tables: Chinese Dialogue  312

### References

- 320

### Index

- 343
Figures

1.1 A panel from the comic book *Western Thrillers* (Fox 1949) page 6
2.1 Close-ups of “The Nabob Rumbled or A Lord Advocates Amusement” by James Gillray with the full print center (image courtesy of the British Museum, ©Trustees of the British Museum)
2.2 Frequencies (normalized per million words) of lemmatized *massa* in the Google Books data tables from 1770 to 1929
2.3 Frequencies (normalized per million words) of lemmatized *sahib* in the Google Books data tables from 1770 to 1929
2.4 A letter circulated both in North American and British newspapers that purports to be written by a slave in Herring Bay, Maryland (Domestic News 1747)
2.5 A letter printed in *The Spectator* purportedly written by a “Bengali Baboo” (An Anglo-Indian 1907)
2.6 Excerpts from a letter printed in the *British Bee Journal* supposedly written by a Chinese American beekeeper (Lung 1893)
3.1 A notice advertising *Don Juan* from *The Morning Chronicle*, November 6, 1837
3.2 Percentages of word counts by period (1768–1829, 1830–1879, and 1880–1929) and controlling for speaker
3.3 A diagram showing the taxonomy of the verb phrase subcategory
4.1 Bar plot showing the deviation of proportions for features in all dialogue with DP ≤ 0.80 and coded by category
4.2 Bar plot showing the deviation of proportions for morphosyntactic features in all dialogue with DP ≤ 0.80 and coded by subcategory
List of Figures

4.3 Base 10 logarithms of frequencies (x-axis) plotted against deviation of proportions (y-axis) for the 25 most dispersed features 66

4.4 Base 10 logarithms of frequencies (x-axis) plotted against deviation of proportions (y-axis) for the 25 most frequent features 67

4.5 A scene from Isabella Banks’s (1882) *Through the Night* showing the discovery of Archie Corbyn, who is mistaken for the ghost of his father 71

4.6 Effect sizes for features with Kruskal-Wallis test-statistics that are significantly attributable to speaker ($p < 0.001$) 74

4.7 Correlation matrix including features where DP ≤ 0.80 for any speaker group 76

4.8 Correlation matrix including features with the highest coefficients and characteristic of African diasporic dialogue 78

4.9 Euler diagrams showing the intersections of word-final -r insertion with other selected features 78

4.10 Euler diagrams showing the intersections of feature pairs with the highest Jaccard similarity coefficients 79

4.11 A scatter plot showing the normalized composite frequencies (the y-axis) over time (the x-axis) of dialect features for texts with a minimum of 95 words 80

4.12 Box plots of composite feature frequencies by speaker 81

4.13 A plot showing the trend lines for composite frequencies 83

4.14 A scatter plot showing the diversity indices for texts with a minimum of 95 words 84

4.15 Box plots of diversity indices by speaker 85

4.16 A plot showing the trend lines for diversity indices 87

4.17 Base 10 logarithms of composite frequencies (x-axis) plotted against diversity indices (y-axis) 88

4.18 A dendrogram showing the hierarchical clusters for texts with a minimum of 95 words 92

4.19 A dendrogram cut into three clusters and coded by speaker 95

4.20 A heat map showing the weighted mean frequencies for three clusters, with features determined by the Kruskal-Wallis test and arranged by effect size 97

4.21 A heat map showing the recalibrated mean frequencies for nine clusters, with features determined by the Kruskal-Wallis test and arranged by effect size 98
List of Figures

4.22 Box plots showing the spread in publication dates in the subgroupings of cluster 1
4.23 The dendrogram with clusters identified by their AU p-values
4.24 A trifoliate grouping from cluster 1A
4.25 A pentafoilate grouping from cluster 1C
5.1 Deviation of proportions for features in African diasporic dialogue with DP ≤ 0.80 and coded by category
5.2 Scatter plots showing linear trends in frequency for the lexical, morphosyntactic, and phonological categories for African diasporic dialogue
5.3 Box plots for the frequencies of lexical, morphosyntactic, and phonological categories for African diasporic dialogue
5.4 Scatter plots showing the linear trends in diversity for the morphosyntactic and phonological categories for African diasporic dialogue
5.5 Stacked area chart showing the nineteenth-century trends (using a generalized additive model) for frequencies of the four superordinate categories in African diasporic dialogue
5.6 Effect sizes for comparisons between the early (pre-1830), middle (1830–1880), and late (1880–1929) periods for the morphosyntactic, orthographic, and phonological categories in African diasporic dialogue
5.7 Stacked area chart showing the nineteenth-century trends (using a generalized additive model) for selected phonological features in African diasporic dialogue
5.8 A dendrogram zoomed and highlighted for African diasporic dialogue
5.9 Box plots for the frequencies of phonological features in the sub-clusters of 1
5.10 Effect sizes for comparisons between cluster 1C and the combined clusters 3B and 3C for the four superordinate categories in African diasporic dialogue
5.11 Cluster containing the African diasporic dialogue from *Cupid in Africa* (1920) and *The Forest* (1924)
5.12 Cluster containing the African diasporic dialogue from *Americans Abroad* (1824) and *No Followers* (1837)
6.1 A bar plot showing the feature frequencies for the four superordinate categories and coded for speaker
List of Figures

6.2 Effect sizes for comparisons between Indian and African diasporic dialogue and between Indian and Chinese dialogue for lexical-type features

6.3 Effect sizes for comparisons between Indian and African diasporic dialogue and between Indian and Chinese dialogue for morphosyntactic-type subcategories

6.4 Effect sizes for comparisons between Indian and African diasporic dialogue and between Indian and Chinese dialogue for phonological-type subcategories

6.5 Effect sizes for comparisons between Indian and African diasporic dialogue and between Indian and Chinese dialogue for the features that are more significantly frequent in Indian dialogue

6.6 Scatter plots showing linear trends in frequency for the lexical, morphosyntactic, and phonological categories for Indian dialogue

6.7 Box plots for the lexical, morphosyntactic, and phonological categories for Indian dialogue

6.8 Scatter plot showing linear trends over time for the morphosyntactic category for Indian dialogue using quantile regression ($\tau = 0.1$–0.9)

6.9 A plot showing goodness-of-fit ($R(\tau)$) for quantiles where $\tau = 0.1$–0.9

6.10 Scatter plot showing linear trends over time for the lexical category for Indian dialogue using two different regression models

6.11 A dendrogram zoomed for Indian dialogue

6.12 A subsection of cluster 1C, which includes three early examples of Indian dialogue

6.13 A subsection of cluster 3C, which includes two early examples of Indian dialogue

6.14 Clusters containing Indian dialogue (left) and African diasporic dialogue (right) from Lutchmee and Dilloo (1877)

6.15 Effect sizes for comparisons between Indian and African diasporic dialogue from Lutchmee and Dilloo (1877) for the four main categories

6.16 Effect sizes for comparisons between Indian and African diasporic dialogue from With a Stout Heart (1874) for the four main categories
List of Figures

6.17 Effect sizes for comparisons between Indian and African diasporic dialogue from *With a Stout Heart* (1874) for the lexical subcategories

6.18 Scatter plot showing linear trends over time for code-mixing for Indian dialogue using two different regression models

7.1 Effect sizes for comparisons between Chinese and African diasporic dialogue and between Chinese and Indian dialogue for the four superordinate categories and total composite frequency

7.2 Scatter plots showing linear trends in frequency for the lexical, morphosyntactic, and phonological categories for Chinese dialogue

7.3 Frequencies (normalized per million words) of lemmatized *chinaman* in the Google Books data tables from 1770 to 1929

7.4 A dendrogram zoomed for Chinese dialogue

7.5 Effect sizes for comparisons between Chinese and African diasporic dialogue from *The Hero of Panama* (1912) for features, where \( p < 0.01 \)

7.6 Effect sizes for comparisons between Chinese from *Under the Waves* (1876) and African diasporic dialogue from *Middy and the Moors* (1888) for features, where \( p < 0.01 \)

7.7 Pentafoliate grouping from cluster 3C in the full dendrogram, which contains the Chinese dialogue from *East of Suez* (1922), *Under the Dragon Throne* (1897), and *Limehouse Nights* (1916)

7.8 Frequencies (normalized per million words) of lemmatized *chinatown* in the Google Books data tables from 1770 to 1929

G.1 A histogram of composite frequencies for all dialex files

G.2 A q-q plot of composite frequencies

G.3 A q-q plot assessing the multivariate normality of all coded features

G.4 A q-q plot of the plot of the residuals after the data has undergone a \( \log + 1 \) transformation

G.5 A q-q plot of the residuals from nontransformed residuals

G.6 Histograms for the frequencies of invariant present tense in the dialogue of each speaker group
List of Figures

G.7 Scatter plot showing the number of features (the y-axis) appearing within a given range of dialex files (the x-axis) 245
G.8 Scatter plot of log-likelihood and log ratio values for all tokens in a comparison of VISiBL and standard dialogue 249
G.9 Sample dendrogram with numbered clades 252
G.10 Chi-squared probability curve for two degrees of freedom, with the area under the curve shaded and marked at the points where it equals 5%, 1%, and 0.1% of the total area 256
G.11 Chi-squared probability curves for 2, 4, and 8 degrees of freedom 257
G.12 Chi-squared probability curve for 1 degree of freedom, with the area under the curve shaded and marked at the points where it equals 5%, 1%, and 0.1% of the total area 257
B.1 The coding taxonomy for the lexical category 269
B.2 The coding taxonomy for the morphosyntactic category 270
B.3 The coding taxonomy for the phonological category 272
B.4 The coding taxonomy for the orthographic category 274
Tables

1.1 Concordance lines from COHA showing adjectives collocating with redskin one token to the left page 5

3.1 Composition of the VISiBL corpus 41
4.1 Superordinate feature types for all categories of speakers 55
4.2 Lexical features for all speakers 56
4.3 Morphosyntactic subcategories for all speakers 59
4.4 The ten most dispersed morphosyntactic features for all speakers 60
4.5 Orthographic subcategories for all speakers 62
4.6 Phonological subcategories for all speakers 63
4.7 Phonological features for all speakers, where DP ≤ 0.80 64
4.8 The top portion of the table shows the top ten keywords (by log-likelihood) where the frequency in standard dialogue is greater than the frequency in literary dialect. The bottom portion shows the top ten keywords where the frequency in literary dialect is greater than the frequency in standard dialogue 69

5.1 Four superordinate categories in African diasporic dialogue 108
5.2 Lexical features in African diasporic dialogue, where DP ≤ 0.80 111
5.3 Morphosyntactic subcategories in African diasporic dialogue 113
5.4 Pronoun-type features in African diasporic dialogue, where DP ≤ 0.80 116
5.5 Phonological subcategories in African diasporic dialogue 119
5.6 Consonant-substitution-type features in African diasporic dialogue, where DP ≤ 0.80 119
5.7 The ten most dispersed phonological features in African diasporic dialogue 120
6.1 Four superordinate categories in Indian dialogue 149
6.2 Lexical features in Indian dialogue 151
## List of Tables

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Table Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.3</td>
<td>Morphosyntactic subcategories in Indian dialogue</td>
<td>153</td>
</tr>
<tr>
<td>6.4</td>
<td>Phonological features in Indian dialogue, where DP &lt; 0.80</td>
<td>155</td>
</tr>
<tr>
<td>7.1</td>
<td>Four superordinate categories in Chinese dialogue</td>
<td>191</td>
</tr>
<tr>
<td>7.2</td>
<td>Lexical features in Chinese dialogue</td>
<td>192</td>
</tr>
<tr>
<td>7.3</td>
<td>Morphosyntactic subcategories in Chinese dialogue</td>
<td>195</td>
</tr>
<tr>
<td>7.4</td>
<td>The 10 most dispersed morphosyntactic features in Chinese dialogue</td>
<td>196</td>
</tr>
<tr>
<td>7.5</td>
<td>Phonological subcategories in Chinese dialogue</td>
<td>198</td>
</tr>
<tr>
<td>7.6</td>
<td>The 10 most dispersed phonological features in Chinese dialogue</td>
<td>199</td>
</tr>
</tbody>
</table>
Acknowledgments

This project began as a purely qualitative, archival study with an inchoate sense that there were important patterns in the ways social variation in English has historically been represented on the page. Those patterns, it seemed to me at the time, continued to assert themselves into the present, though not without revisioning, adaptation, and challenge. Yet, I didn’t think that I had the vocabulary or the tools to effectively map those patterns. It wasn’t until after I was introduced to corpus techniques by my mentors at the University of Michigan – Richard Bailey, John Swales, and Anne Curzan – that I began to imagine a way forward. Still, it took a few more years and the encouragement of Paul Nerney, my colleague at the National University of Singapore, to finally pick up the project again. Since then, I have continued to be supported by generous colleagues at a number of institutions. While I was at UCLA, I had innumerable conversations with John Schumann. There are few people I would rather share a beer with and talk shop. Stefan Gries, whom I met one afternoon after making the short drive up to the University of California, Santa Barbara, took the time to point me in productive directions, when I was just getting my feet wet with R. The book was finished after I eventually landed at Marymount University. Because there are so many people at Marymount to name, I fear I would leave someone out, so let me just say that I cannot imagine a more perceptive and supportive group with whom to work and share ideas. Finally, my readers in the UK were instrumental as the project took shape. These include Tony McEnery, Andrew Hardie, Claire Hardaker, Christopher Hart, and Jonathan Culpeper at Lancaster University, as well as Michaela Mahlberg at the University of Birmingham. There are, however, two people most responsible for the project’s completion. The first of these is Paul Baker. The book would not be what it is without his help and insight. The other is Jasmine Trice – an incredible scholar, friend, and partner. And thanks, too, to all of my friends and family for the myriad ways you have encouraged and sustained me over the years.