Variation within the English language is a vast research area, of which dialectology, the study of geographic variation, is a significant part. This book explores grammatical differences between British English dialects, drawing on authentic speech data collected in over 30 counties. In doing so it presents a new approach known as “corpus-based dialectometry,” which focuses on the joint quantitative measurement of dozens of grammatical features to gauge regional differences. These features include, for example, multiple negation (e.g. don’t you make no damn mistake), non-standard verbal -s (e.g. so I says, What have you to do?), or non-standard weak past tense and past participle forms (e.g. they knewed all about these things). Utilizing state-of-the-art dialectometrical analysis and visualization techniques, the book is original both in terms of its fundamental research question (“What are the large-scale patterns of grammatical variability in British English dialects?”) and in terms of its methodology.

BENEDIKT SZMRECSEANYI is a Fellow at the Freiburg Institute for Advanced Studies, Germany.
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:

Raymond Hickey (ed.): Legacies of Colonial English
Merja Kyto, Mats Rydén and Erik Smitterberg (eds.): Nineteenth Century English: Stability and Change
John Algeo: British or American English? A Handbook of Word and Grammar Patterns
Christian Mair: Twentieth-Century English: History, Variation and Standardization
Evelien Keizer: The English Noun Phrase: The Nature of Linguistic Categorization
Raymond Hickey: Irish English: History and Present-Day Forms
Günter Rohdenburg and Julia Schlüter (eds.): One Language, Two Grammars?: Differences between British and American English
Laurel J. Brinton: The Comment Clause in English
Geoffrey Leech, Marianne Hundt, Christian Mair and Nicholas Smith: Change in Contemporary English: A Grammatical Study
Jonathan Culpeper and Merja Kyto: Early Modern English Dialogues: Spoken Interaction as Writing
Daniel Schreier, Peter Trudgill, Edgar Schneider and Jeffrey Williams: The Lesser-Known Varieties of English: An Introduction
Hilde Hasselgård: Adjunct Adverbials in English
Raymond Hickey: Eighteenth-Century English: Ideology and Change
Charles Boberg: The English Language in Canada: Status, History and Comparative Analysis
Thomas Hoffmann: Preposition Placement in English: A Usage-Based Approach
Claudia Claridge: Hyperbole in English: A Corpus-Based Study of Exaggeration
Paivi Pahta and Andreas H. Jucker (eds.): Communicating Early English Manuscripts
Irma Taavitsainen and Paivi Pahta (eds.): Medical Writing in Early Modern English
Colette Moore: Quoting Speech in Early English
David Denison, Ricardo Bermúdez-Otero, Chris McCully and Emma Moore (eds.): Analysing Older English: Evidence, Methods and Solutions
Jim Feist: Premodifiers in English: Their Structure and Significance
Steven Jones, M. Lynne Murphy, Carita Paradis and Caroline Willners: Antonyms in English: Construals, Constructions and Canonicity
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 Szmrecsanyi: Grammatical Variation in British English Dialects: A Study in Corpus-Based Dialectometry

Earlier titles not listed are also available
Grammatical Variation in British English Dialects

A Study in Corpus-Based Dialectometry

BENEDIKT SZMRECZANYI

Freiburg Institute for Advanced Studies
Contents

List of figures xi
List of maps xiii
List of tables xv
Preface and acknowledgments xvii

1 Introduction 1
1.1 Rationale, method, and objectives 1
1.2 Previous big-picture accounts 6
  1.2.1 On the Survey of English Dialects 7
  1.2.2 Nineteenth-century accent differences: Alexander Ellis’ survey of English dialects (1889) 7
  1.2.3 SED-based analyses I: Trudgill’s (1990) division of traditional dialects 8
  1.2.4 SED-based analyses II: The Salzburg School of Dialectometry 9
  1.2.5 SED-based analyses III: Bamberg-type dialectometry 9
  1.2.6 SED-based analyses IV: Shackleton (2007) 10
  1.2.7 The network perspective: McMahon et al. (2007) 11
  1.2.8 The perceptual dialectology perspective: Inoue (1996) 12
  1.2.9 Synopsis 12
1.3 Structure 13

2 Data and methods 15
2.1 Data 15
  2.1.1 The Freiburg Corpus of English Dialects 15
  2.1.2 Standard English text corpora 23
2.2 Method: corpus-based dialectometry 24
  2.2.1 The empirical foundation: defining the feature catalogue 24
## Contents

2.2.2 Data mining: extracting feature frequencies and creating a frequency matrix 24
2.2.3 Aggregation: obtaining a distance matrix 27
2.2.4 Cartographic projections to geography 28
2.2.5 Software tools 30

3 The feature catalogue 32
  3.1 The feature catalogue: an overview 32
  3.2 Feature selection 35
    3.2.1 Frequency 36
    3.2.2 Feasibility 37
  3.3 Feature extraction 38
    3.3.1 Fully automatic extraction 38
    3.3.2 Semi-automatic extraction 38
  3.4 The feature catalogue in detail 40
    3.4.1 Pronouns and determiners 41
    3.4.2 The noun phrase 44
    3.4.3 Primary verbs 48
    3.4.4 Tense, mood, and aspect 50
    3.4.5 Modality 52
    3.4.6 Verb morphology 53
    3.4.7 Negation 56
    3.4.8 Agreement 60
    3.4.9 Relativization 63
    3.4.10 Complementation 64
    3.4.11 Word order and discourse phenomena 66
  3.5 Chapter summary 69

4 Surveying the forest: on aggregate morphosyntactic distances and similarities 71
  4.1 Aggregate distances and similarities: the big picture 71
    4.1.1 On the network nature of morphosyntax relationships 73
    4.1.2 Interpoint relationships 74
    4.1.3 Two similarity profiles 75
  4.2 Two regionally selective analyses 77
    4.2.1 Aggregate morphosyntactic distances and similarities in England 77
    4.2.2 Aggregate morphosyntactic distances and similarities in Scotland 80
  4.3 On linguistic compromise areas and dialect kernels 81
  4.4 Dialectal similarities and distances to Standard English 83
  4.5 Chapter summary 87
5 Is morphosyntactic variability gradient? Exploring dialect continua 89
  5.1 On dialect continua 89
  5.2 The cartographic take on dialect continua 91
    5.2.1 Multidimensional Scaling 91
    5.2.2 Some MDS plots 92
    5.2.3 Introducing continuum maps 95
    5.2.4 Cartography: interim summary 100
  5.3 Quantification 100
    5.3.1 As-the-crow-flies distance 101
    5.3.2 Least-cost walking distance 103
    5.3.3 Least-cost travel time 103
    5.3.4 Linguistic gravity 105
    5.3.5 Where do continuous predictors fail us? 107
    5.3.6 Quantification: interim summary 109
  5.4 Chapter summary 109

6 Classification: the dialect area scenario 111
  6.1 On the notion of dialect areas 111
  6.2 Quantifying the explanatory power of previous dialect partitions 113
  6.3 Deriving dialect areas inductively 117
    6.3.1 Method: Hierarchical Agglomerative Cluster Analysis 117
    6.3.2 Clustering: a simple example 119
    6.3.3 Analysis: dendrograms, cluster maps, and overlap with previous partitions 121
    6.3.4 Interim summary 126
  6.4 Chapter summary 127

7 Back to the features 128
  7.1 Revisiting the outliers 128
    7.1.1 Banffshire 129
    7.1.2 Denbighshire 130
    7.1.3 Dumfriesshire 131
    7.1.4 Middlesex 133
    7.1.5 Leicestershire 134
    7.1.6 Warwickshire 136
  7.2 The quest for feature bundles: Principal Component Analysis 137
    7.2.1 Principal component 1: the non-standard come component 141
    7.2.2 Principal component 2: the DO and HAVE component 145
    7.2.3 Principal component 3: the BE component 147
x Contents

7.2.4 Principal component 4: the WOULD component 149
7.3 Chapter summary 150
8 Summary and discussion 151
  8.1 Summary: morphosyntactic variability in British English dialects 151
  8.2 Is geography overrated? 157
9 Outlook and concluding remarks 164

Appendices
Appendix A: the feature catalogue–summary statistics 166
Appendix B: component-loading matrix 169
Appendix C: color maps 172

References 190
Index 207
Figures

2.1 Converting a fictional $3 \times 2$ frequency matrix into a $3 \times 3$ distance matrix utilizing Euclidean distance as an aggregation measure  

4.1 Aggregate morphosyntactic distances: histogram, all measuring points in dataset 72
4.2 Aggregate morphosyntactic distances: histogram, measuring points in England only 78
4.3 Connections among nineteenth-century English English dialects 79
4.4 Aggregate morphosyntactic distances: histogram, measuring points in Scotland and the Hebrides only 80
4.5 Aggregate morphosyntactic distances to Standard British English: histogram 84
4.6 Aggregate morphosyntactic distances to Standard American English: histogram 85
5.1 Two-dimensional MDS plot, as-the-crow-flies distances 93
5.2 Two-dimensional MDS plot, morphosyntactic distances plus Standard English 93
5.3 Two-dimensional MDS plot, morphosyntactic distances 93
5.4 Three-dimensional MDS plot, morphosyntactic distances plus Standard English 96
5.5 Morphosyntactic dialect distances versus as-the-crow-flies distances 101
5.6 Morphosyntactic dialect distances versus walking distance estimates 104
5.7 Morphosyntactic dialect distances versus travel time estimates 105
5.8 Morphosyntactic dialect distances versus log-transformed linguistic gravity scores 107
7.1 Principal Component Analysis 138
xii  List of figures

7.2  Text frequencies of feature [33] versus text frequencies of feature [32]  140
7.3  Text frequencies of feature [33] versus text frequencies of feature [14]  141
8.1  Morphosyntactic dialect distances (selected feature sets) versus as-the-crow-flies distances  161
8.2  Morphosyntactic dialect distances (various levels of data reduction) versus as-the-crow-flies distances  162
C.1  Dendrogram, as-the-crow-flies distances (algorithm: WPGMA)  183
C.2  Dendrogram, morphosyntactic distances (algorithm: WPGMA)  184
C.3  Dendrogram, as-the-crow-flies distances (algorithm: Ward’s Method)  185
C.4  Dendrogram, morphosyntactic distances (algorithm: Ward’s Method)  186
Maps

2.1 FRED counties considered in the present study. page 20
2.2 Distance measures matter 27
2.3 A Voronoi tessellation of Great Britain’s map space 30

3.1 Feature [1] (non-standard reflexives) 42
3.2 Feature [9] (the s-genitive) 46
3.3 Feature [14] (TO BE) 49
3.4 Feature [20] (used to) 51
3.5 Feature [26] (GOT TO) 53
3.6 Feature [29] (non-standard past tense done) 55
3.7 Feature [30] (non-standard past tense come) 56
3.8 Feature [31] (the negative suffix -nac) 57
3.9 Feature [32] (ain’t) 58
3.10 Feature [33] (multiple negation) 59
3.11 Feature [40] (don’t with 3rd person singular subjects) 61
3.12 Feature [49] (as what/than what in comparative clauses) 64
3.13 Feature [56] (prepositional datives) 68
3.14 Feature [57] (double object structures) 69

C.1 Network map, all measuring points in dataset 173
C.2 Reverse network map, all measuring points in dataset 173
C.3 Beam map 174
C.4 Honeycomb map 174
C.5 A similarity map for Kincardineshire 175
C.6 A similarity map for Nottinghamshire 175
C.7 Network map, measuring points in England only 176
C.8 Reverse network map, measuring points in England only 176
C.9 Network map, measuring points in Scotland and the Hebrides only 177
C.10 Reverse network map, measuring points in Scotland and the Hebrides only 177
C.11 Skewness map 178
C.12 Kernel map: 178
xiv  List of maps

C.13  A similarity map for Standard British English  179
C.14  A similarity map for Standard American English  179
C.15  A perfect continuum map for reference purposes  180
C.16  Continuum map (including distances to Standard English)  180
C.17  Continuum map  181
C.18  Visualizing regression residuals (full dataset)  182
C.19  Visualizing regression residuals (England only)  182
C.20  Visualizing regression residuals (Scottish Lowlands only)  182
C.21  Cluster map, projecting the dendrogram in Figure C.1 to geography  183
C.22  Cluster map, projecting the dendrogram in Figure C.2 to geography  184
C.23  Cluster map, projecting the dendrogram in Figure C.3 to geography  185
C.24  Cluster map, projecting the dendrogram in Figure C.4 to geography  186
C.25  Principal component 1: mean component score map  187
C.26  Principal component 2: mean component score map  187
C.27  Principal component 3: mean component score map  188
C.28  Principal component 4: mean component score map  188
C.29  A synopsis map combining the continuum scenario and the dialect area scenario  189
# Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>FRED counties considered in the present study</td>
<td>19</td>
</tr>
<tr>
<td>4.1</td>
<td>Aggregate morphosyntactic distances: summary statistics, all measuring points in dataset</td>
<td>72</td>
</tr>
<tr>
<td>4.2</td>
<td>Aggregate morphosyntactic distances: summary statistics, measuring points in England only</td>
<td>78</td>
</tr>
<tr>
<td>4.3</td>
<td>Aggregate morphosyntactic distances: summary statistics, measuring points in Scotland and the Hebrides only</td>
<td>80</td>
</tr>
<tr>
<td>4.4</td>
<td>Aggregate morphosyntactic distances to Standard British English: summary statistics</td>
<td>84</td>
</tr>
<tr>
<td>4.5</td>
<td>Aggregate morphosyntactic distances to Standard American English: summary statistics</td>
<td>85</td>
</tr>
<tr>
<td>4.6</td>
<td>Euclidean distances to Standard British English and Standard American English</td>
<td>86</td>
</tr>
<tr>
<td>6.1</td>
<td>Assigning FRED measuring points to Early English dialect areas according to the customary partitions in Baugh and Cable (1993)</td>
<td>115</td>
</tr>
<tr>
<td>6.2</td>
<td>The explanatory power of previous dialect partitions for the current dataset. Number of measuring points on which the analysis is based (N), and PERMANOVA $R^2$ value (% variance explained)</td>
<td>116</td>
</tr>
<tr>
<td>6.3</td>
<td>Adjusted Rand Indices (ARIs). (Partial) overlap between the WPGMA 6-cluster solution and previous dialect divisions. ARIs closer to 1 indicate more overlap</td>
<td>124</td>
</tr>
<tr>
<td>6.4</td>
<td>Adjusted Rand Indices (ARIs). (Partial) overlap between Ward’s 3-cluster solution and previous dialect divisions. ARIs closer to 1 indicate more overlap</td>
<td>126</td>
</tr>
<tr>
<td>7.1</td>
<td>Bonferroni-significant feature frequency differentials between Dumfriesshire English and a geographically close neighbor, West Lothian English</td>
<td>132</td>
</tr>
<tr>
<td></td>
<td></td>
<td>xv</td>
</tr>
</tbody>
</table>
### List of tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.2</td>
<td>Bonferroni-significant feature frequency differentials between Middlesex English and a geographically close neighbor, London English</td>
<td>134</td>
</tr>
<tr>
<td>7.3</td>
<td>Bonferroni-significant feature frequency differentials between Leicestershire English and a geographically close neighbor, Nottinghamshire English</td>
<td>135</td>
</tr>
<tr>
<td>7.4</td>
<td>Principal component 1: rotated component loadings</td>
<td>142</td>
</tr>
<tr>
<td>7.5</td>
<td>Principal component 2: rotated component loadings</td>
<td>145</td>
</tr>
<tr>
<td>7.6</td>
<td>Principal component 3: rotated component loadings</td>
<td>147</td>
</tr>
<tr>
<td>7.7</td>
<td>Principal component 4: rotated component loadings</td>
<td>149</td>
</tr>
<tr>
<td>8.1</td>
<td>Morphosyntactic features with a geographically significant distribution in FRED.</td>
<td>152</td>
</tr>
<tr>
<td>8.2</td>
<td>Shared variance ($R^2$) between as-the-crow-flies distances and linguistic distances in selected studies.</td>
<td>158</td>
</tr>
</tbody>
</table>
Preface and acknowledgments

This book is a revised version of my Habilitation thesis, which I submitted to the University of Freiburg in 2011. Partial summaries of the research discussed in this book have appeared, or are appearing, as Szmrecsanyi (2008, 2011, forthcoming) and Szmrecsanyi and Wolk (2011). Generous support, intellectual and material, by the Freiburg Institute for Advanced Studies (FRIAS) is gratefully acknowledged.

For assistance in writing this book, I owe gratitude to many individuals, including the following: Bernd Kortmann, for discussion, feedback, and support; Peter Auer, Bernd Kortmann, and Christian Mair, who acted as official reviewers of my Habilitation thesis and provided me with eminently helpful and constructive reports; Hans Goebel, for inviting me to Salzburg in 2009 and teaching me Salzburg dialectometry, and for invaluable, open-minded comments and feedback before, during, and after this visit; Bernhard Castellazzi, Hans Goebel, and Pavel Smecka, for creating a number of Salzburg-style dialectometry maps (Maps C.3–C.6 and C.11–C.12); Jack Grieve, for calculating the global spatial autocorrelation (Moran’s I) scores reported in Chapter 3; Wilbert Heeringa, for a number of excellent comments and suggestions, and for providing me with a convenient software tool to conduct the Mantel test; Paul Kerswill, for discussing some morphosyntactically ill-behaved dialects with me; Peter Kleiweg, for creating and maintaining the RUG/L04 dialectometry software (without which this book would be a lot less colorful); Merja Kytö, the series editor, and two anonymous referees for encouraging and constructive feedback; John Nerbonne, Wilbert Heeringa, and Bart Alewijnse, for having me over in Groningen in 2007 and explaining Groningen dialectometry to me; Sigrid Saou, for a close reading, many helpful comments, and good questions; and Christoph Wolk, for help with some of the statistics in this book. I also wish to thank Elin Arbin, for meticulous proof-reading, and Katharina Ehret, for help in preparing the final typescript.

The usual disclaimers apply. Specifically, none of the above individuals necessarily agrees with any or all of my interpretations and conclusions, and all errors are solely mine.

This book is dedicated to my wife Sigrid and to my daughter Leonore.