Change in Contemporary English

Based on the systematic analysis of large amounts of computer-readable text, this book shows how the English language has been changing in the recent past, often in unexpected and previously undocumented ways. The study is based on a group of matching corpora, known as the 'Brown family' of corpora, supplemented by a range of other corpus materials, both written and spoken, drawn mainly from the later twentieth century. Among the matters receiving particular attention are the influence of American English on British English, the role of the press, the 'colloquialization' of written English, and a wide range of grammatical topics, including the modal auxiliaries, progressive, subjunctive, passive, genitive and relative clauses. These subjects build an overall picture of how English grammar is changing, and the linguistic and social factors that are contributing to this process.

GEOFFREY LEECH is Emeritus Professor of English Linguistics in the Department of Linguistics and English Language at Lancaster University.

MARIANNE HUNDT is Professor of Linguistics in the Department of English at the University of Zürich.

CHRISTIAN MAIR is Professor of English Linguistics in the Department of English at the University of Freiburg.

NICHOLAS SMITH is Lecturer in English Language and Linguistics in the School of English, Sociology, Politics and Contemporary History at the University of Salford.

STUDIES IN ENGLISH LANGUAGE

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:

Christian Mair: *Infinitival Complement Clauses in English: A Study of Syntax in Discourse*
Charles F. Meyer: *Apposition in Contemporary English*
Jan Firbas: *Functional Sentence Perspective in Written and Spoken Communication*
Izchak M. Schlesinger: *Cognitive Space and Linguistic Case*
Katie Wales: *Personal Pronouns in Present-Day English*
Laura Wright: *The Development of Standard English, 1300–1800: Theories, Descriptions, Conflicts*
Charles F. Meyer: *English Corpus Linguistics: Theory and Practice*
Stephen J. Nagle and Sara L. Sanders (eds.): *English in the Southern United States*
Anne Curzan: *Gender Shifts in the History of English*
Kingsley Bolton: *Chinese Englishes*
Irma Taavitsainen and Päivi Pahta (eds.): *Medical and Scientific Writing in Late Medieval English*
Elizabeth Gordon, Lyle Campbell, Jennifer Hay, Margaret Maclagan, Andrea Sudbury and Peter Trudgill: *New Zealand English: Its Origins and Evolution*
Raymond Hickey (ed.): *Legacies of Colonial English*
Merja Kytö, 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*
Lieselotte Anderwald: *The Morphology of English Dialects: Verb Formation in Non-standard English*

Change in Contemporary English

A Grammatical Study

GEOFFREY LEECH
Lancaster University

MARIANNE HUNDT
Universität Zürich

CHRISTIAN MAIR
Albert-Ludwigs-Universität Freiburg, Germany

NICHOLAS SMITH
University of Salford

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore,
São Paulo, Delhi

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK

Published in the United States of America by Cambridge University Press,
New York

www.cambridge.org
Information on this title: www.cambridge.org/9780521867221

First published 2009

Printed in the United Kingdom at the University Press, Cambridge

A catalogue record for this publication is available from the British Library

Library of Congress Cataloguing in Publication data
Leech, Geoffrey N.
Change in contemporary English : a grammatical study / Geoffrey Leech . . . [et al.].
p. cm. – (Studies in English language)
Includes bibliographical references and index.
ISBN 978-0-521-86722-1
1. Corpora (Linguistics) 2. English language – Grammar. 3. Computational linguistics. I. Title. II. Series.
P128.C68L43 2009
425 – dc22 2009019728

ISBN 978-0-521-86722-1 hardback

Contents

Figures

List of Figures in Appendix III

Tables

Preface

This book aims to give an account of how the English language has been changing recently, focusing especially on (a) the late twentieth century, (b) the written standard language, (c) American and British English, (d) grammatical rather than lexical change, and using the empirical evidence of computer corpora.

Corpus linguistics is now a mainstream paradigm in the study of languages, and the study of English in particular has advanced immeasurably through the availability of increasingly rich and varied corpus resources. This applies to both synchronic and diachronic research. However, this book presents, we argue, a new kind of corpus-based historical research, with a narrower, more intense focus than most, revealing through its rather rigorous methodology how the language (more especially the written language) has been developing over a precisely defined period of time in the recent past.

The period on which the book concentrates is the thirty years between the early 1960s and the early 1990s, and the four corpora that it studies in most detail are those which go increasingly by the name of the 'Brown family': the Brown corpus (American English, 1961); the Lancaster–Oslo/Bergen corpus (British English, 1961); the Freiburg–Brown corpus (American English, 1992); and the Freiburg–Lancaster–Oslo/Bergen corpus (British English, 1991).[1]

These corpora, described in more detail in Chapter 2 (section 2.2) and in Appendix II, are reasonably well known, and have been studied as a group, not only by ourselves, but by others, since the completion of this corpus quartet in the mid-1990s. All four corpora are available to researchers around the world, and can be obtained under licence from either ICAME at the Aksis centre, University of Bergen, or the Oxford Text Archive, University of Oxford.[2] However, we venture to claim that as authors of this book we have been more intimately engaged with these corpora than any other research group: in their compilation, their annotation and their analysis. Indeed, this

[1] An informative manual of information for the Brown family of corpora, including their POS tagging, is provided by Hinrichs *et al.* (forthcoming).

[2] The web addresses of these two corpus resource agencies are as follows: http://icame.uib.no/ and http://ota.ahds.ac.uk/.

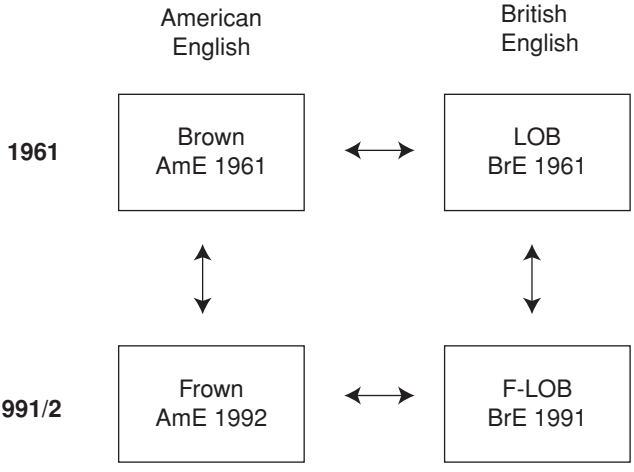

Figure 0.1 The four matching corpora on which this book focuses

intimacy entitles us to feel a certain familial affection for these textual time-capsules, and almost invariably (like many others) we refer to them by their acronymic nicknames: Brown, LOB, Frown and F-LOB.[3]

The strength of these four corpora lies in their comparability: the fact that they are constructed according to the same design, having virtually the same size and the same selection of texts and genres represented by 500 matching text samples of c. 2,000 words. This means that we can use the Brown family as a precision tool for tracking the differences between written English in 1961 and in 1991/2. How has the English language changed, in these two leading regional varieties, over this thirty-year generation gap? The findings brought to light by this comparison between matching corpora are fascinating: they reveal, for the first time, or at least with a new sense of accuracy, how significant are the changes in a language that take place over even such a short timespan of thirty years. Even though these changes, as we report them, are almost entirely matters of changing frequency of use, they often show a high degree of statistical significance.[4]

The affection we feel for this corpus family does not blind us to their considerable limitations (see section 2.1), notably their restriction to the standard written language. We have therefore taken care to supplement the evidence they provide with analyses of other corpora relating to the later twentieth century, so as to enlarge and corroborate our findings on how the language has recently been changing. In extending our range in this way, most

[3] The explanations of these names for corpora, as well as other abbreviations, are found in the list of 'Abbreviations and symbolic conventions', pp. xxvii–xxx.
[4] Significance levels are shown, where appropriate, by asterisks: *, **, *** in the quantitative tables – see the table of Abbreviations and symbolic conventions.

important have been the corpora that record indications of what has been happening to the spoken language. The Diachronic Corpus of Present-Day Spoken English (DCPSE),[5] released in 2006, has made it possible to study, over the sample period of time, changes in the spoken language, though not under the same rigorous conditions of comparability that apply to the Brown family. In addition, the British National Corpus (BNC), though it has no reliable diachronic dimension, gives us a large (ten-million-word) well-sampled subcorpus of spoken English from the early 1990s. Both of these corpora are limited to British English: but we have been able to consult the CIC (Cambridge International Corpus) and LCSAE (Longman Corpus of Spoken American English, comparable in date and method of collection to the spoken demographic subcorpus of the BNC) to see how that presumably most trail-blazing variety of the language – spoken American English – compares with others. Again, there is only an indirect diachronic dimension here, through the study of 'apparent time' by comparison of different age groups of speakers. But at least we are able to speculate on tangible evidence about how the spoken American variety has been moving in the period under review.[6]

Apart from these (necessarily imperfect and incomplete) comparisons between corpora of speech and writing, we have also been able to extend our range, when need arises, along the diachronic dimension. In the months preceding the publication of this book, we were able to make limited use of the newest member of the Brown family – though oldest in date – the Lancaster 1931 Corpus[7] (inevitably nicknamed 'B-LOB' for 'before-LOB'), sampled from a seven-year period centring on 1931, and so effectively providing us with three equidistant reference points, 1931 (± 3 years), 1961 and 1991/2, for further diachronic comparison. For even greater historical depth, we have occasionally used the ARCHER corpus and the *OED* citation bank. These valuable resources again lack the strict comparability criterion of the Brown

[5] The DCPSE, consisting of 885,436 words, and compiled by Bas Aarts and associates at the Survey of English Usage, University College London, consists of transcribed British spoken texts originally collected as parts of two different corpora: (a) the Survey of English Usage corpus (of which the spoken part was later largely incorporated into the London–Lund Corpus) collected in 1958–1978; and (b) the ICE-GB corpus collected in 1990–1992. Geoffrey Leech is grateful to Bas Aarts for letting him have an advance copy of DCPSE at a point when it was timely for drafting certain chapters of this book.

[6] It should be mentioned that there are several corpora of present-day spoken English of which we have not made detailed use, since, although admirable for other types of research, they are either two small for our present purposes (e.g. the Santa Barbara Corpus of Spoken American English) or too genre-restricted (e.g. MICASE, Corpus of Spoken Professional American English, the Switchboard corpus).

[7] This corpus, now in a provisional pre-release form, has been compiled by Nicholas Smith, Paul Rayson and Geoffrey Leech with the financial support of the Leverhulme Foundation. With further support from the Leverhulme Foundation, we will shortly have yet another member of the Brown family, with a corpus of BrE at the beginning of the twentieth century (1901 ± 3 years to be precise). However, this corpus, provisionally called Lanc-1901, was not completed in time for its results to be used in this book.

family, but allow corpus-based investigations of trends going back to EModE (in the case of ARCHER) and to OE (in the case of the *OED* citation bank). Turning towards the future: we have not been able to draw on more recent progeny of the Brown family, since none are yet available; but the 'corpus of last resort' these days, the World Wide Web (see a number of contributions to Hundt *et al.* 2007), has sometimes given us persuasive evidence about what has been happening since the early 1990s.[8]

What has become obvious is that the corpus resources available for recent diachronic research do not comprise a static platform for research, but a moving staircase: every year new text resources become available, in increasing numbers and increasing size, enhancing our evidential basis for researching the recent development of the language. In such a situation of continuing advance, it is a reasonable compromise to adopt the position we have taken – to focus on the four tried-and-tested Brown family corpora, while using other corpora where it is particularly rewarding or important (as well as feasible) to do so.

The unavoidable assumption of incompleteness is familiar in many fields of scientific endeavour: if researchers before publication waited until complete results and complete answers were available, there would be no publication. Certainly, it would have been easy for us to engage in further research on the range of topics we have investigated here, collecting or consulting further corpora, carrying out deeper analyses, and so on, without reaching a natural endpoint. We hope that in spite of its existing limitations, this book will be felt to have achieved a valuable conspectus of new or recent findings across a wide variety of grammatical topics. Although we have taken care to achieve a consistent perspective and framework of research throughout the book, readers may notice some lack of consistency in the kinds of coverage of corpus analyses offered in individual chapters. In the 'moving staircase' scenario described above, this is almost inevitable, and there is after all no harm in a book which reflects to some extent the different emphases, interests and strengths of individual chapter authors.

One of the most positive achievements of our collaboration is the uniform part-of-speech annotation (or POS tagging) of all four corpora – all five, if one includes the 1931 corpus. We have used the same software annotation practices (the Lancaster tagger CLAWS, the supplementary tagger Template Tagger and the enriched C8 tagset of grammatical categories – see Appendix II and also the detailed tagging guide in Hinrichs *et al.* forthcoming). This has enabled the corpora to be compared, grammatically, on an equal footing, using equivalent search and retrieval patterns to extract instances of abstract constructions, such as progressives, and in some important instances (e.g.

[8] Paul Baker of Lancaster University has provisionally compiled a twenty-first century web-derived corpus on the Brown model, and this will eventually take its place in the Brown family of corpora.

zero relativization) even grammatical categories not explicitly realized in surface structure. Here again, however, we have not managed to achieve complete consistency of treatment: the three corpora LOB, F-LOB and Frown have all been manually post-edited after automatic tagging, while the Brown corpus, the earliest of all to be compiled and tagged, has not undergone the manual post-edit with the new set of tags. This has meant a lower degree of confidence (with an initial error margin of c. 2 per cent) in the correctness of some results in the American English (AmE) comparison of Brown and Frown, alongside the more accurate British English (BrE) comparison of LOB and F-LOB. However, this margin of error has been minimized by employing a corrective coefficient based on the tagger's error rates observed in the comparison of pre-corrected and post-corrected versions of the Frown Corpus – see further p. 24, footnote 27.[9] The dictum that 'Most corpus findings are approximations' (see section 2.3) is particularly to be taken to heart in interpreting our findings for grammatical constructions and categories in AmE, and this has sometimes led us to give more attention to the results for BrE than those for AmE.

Given that the book focuses on changes in grammar, the POS tagging combined with powerful CQP search software (see section 2.4 C) has enabled us, without aiming at comprehensiveness, to achieve a broad grammatical coverage of the language.[10] After two introductory chapters, the next seven chapters concentrate on topics relating to the verb phrase. They cover the subjunctive (Chapter 3), the modal auxiliaries (Chapter 4), the so-called semi-modals (Chapter 5), the progressive aspect (Chapter 6), the passive (Chapter 7), expanded predicates such as *have/take a look* (Chapter 8) and non-finite constructions (Chapter 9). In Chapter 10 we move on to the noun phrase, enquiring particularly into noun–noun sequences, genitives and relative clauses. In the last chapter, Chapter 11, we seek a synthesis, dealing with social and linguistic determinants of the short-term changes demonstrated in earlier chapters, and extending the book's coverage by illustrating these determinants with a number of additional linguistic trends.

The book abounds with statistical tables and charts, comparing frequencies (often normalized to occurrences per million words) according to period of

[9] Tables and figures relying on approximations based on adjusted automatic tagging counts in this way occur mostly in Chapters 10 and 11, or in the part of Appendix III relating to these chapters. Such tables and figures are indicated by a warning note '(automatic)' or '(AmE automatic)' beneath the relevant table or figure.

[10] A simple and obvious point has to be made here: we have naturally given primary attention to areas of English grammar known or suspected to be undergoing change. (In some cases the 'knowledge' or 'suspicion' comes from our own exploratory study of the corpora.) There are, however, interesting areas of contemporary English grammar that we have not dealt with: for example, we will have nothing to say about corpus findings relating to the choice of singular or plural verb after a collective-noun subject (*The team is/are . . .* – a construction that has, however, been more than adequately studied elsewhere – see Levin 2001, 2006; Depraetere 2003; Hundt 2006). Our failure to treat a particular topic is not a reliable signal of its lack of interest from the present-day diachronic viewpoint.

time (mostly 1961 vs 1991), region, genre, etc. We have aimed to provide sound corpus description, using inferential statistics to generalize beyond corpus observations, looking at single dependent variables at a time, and interpreting the findings in the framework of a reasonable and robust usage-based model of language change. To avoid cluttering up the descriptive chapters (Chapters 3–11) with statistical details that might obscure the main findings and lines of argument, we have consigned many of the statistical tables and diagrams, particularly the more complex ones, to Appendix III.

The four authors are jointly responsible for the whole work in its final form; nevertheless, it may be of interest to know which authors took particular responsibility for which chapters. They are here identified by their initials: GL: 2, 4, 5, 10, 11; MH: 3, 7, 8, also the References; CM: 1, 9; NS: 6, Appendices. It should be added, however, that the relative input of individual authors can by no means be measured in this way.

This is the appropriate point to acknowledge gratefully our debt to those who helped us in various ways; to Merja Kytö as series General Editor, and to Helen Barton, editor at Cambridge University Press, we owe a great deal for their encouragement, support and forbearance. We also owe much to the research assistants who helped us in the processing of textual data: Lars Hinrichs, Barbara Klein, Luminiţa-Irinel Traşcă and Birgit Waibel in Freiburg; and Martin Schendzielorz in Heidelberg. We are grateful, too, to Paul Rayson and Sebastian Hoffmann, colleagues at Lancaster; to Gunnel Tottie, for expert guidance on American and British English; and to Chris Williams for comments on Chapter 6; also to the funding agencies without whose support our research reported here would not have been possible. Thanks are due, on this score, to the Deutsche Forschungsgemeinschaft (DFG) for grant MA 1652/3 to Christian Mair and the University of Freiburg, to the Arts and Humanities Research Board (AHRB; subsequently changed to AHRC), the British Academy, and the Leverhulme Trust for research grants awarded to Geoffrey Leech at Lancaster University. We also record our gratitude to Cambridge University Press for making available to us relevant sections of the Cambridge International Corpus (CIC), and to Pearson/Longman for allowing us to consult the Longman Corpus of Spoken American English (LCSAE).

Abbreviations and symbolic conventions

A. Abbreviations for corpora, corpus collections and subcorpora
(listed approximately in order of importance for this book)

1	Brown	the Brown (University) Corpus (see 2.1, Appendix I)
2	LOB	the Lancaster–Oslo/Bergen Corpus
3	Frown	the Freiburg–Brown Corpus
4	F-LOB	the Freiburg–Lancaster–Oslo/Bergen Corpus
5	The Brown family	the four corpora above, regarded as a group
6	Lanc-31	the Lancaster 1931 Corpus, matching the four corpora above
7	B-LOB	a nickname for Lanc-31, meaning 'before LOB'
8	Press,	Four subcorpora into which the corpora of
9	General Prose,	the Brown family are divided. For the
10	Learned,	composition of the Brown corpus (and hence
11	Fiction	of the other corpora of the Brown family), see Appendix I.
12	BNC	the British National Corpus[1]
13	the BNC demographic subcorpus (BNCdemog)	a part of the BNC, consisting of largely spontaneous spoken English discourse by 153 individuals and their interlocutors, sampled from the population of the UK on demographic principles
14	BNC Sampler	A subcorpus of the BNC, consisting of c. one million words of writing and c. one million words of speech. The POS tags are more refined than for the whole BNC, and have been post-edited for correctness.
15	ICE	the International Corpus of English

[1] We have used the World Edition of the British National Corpus.

16	ICE-GB	the International Corpus of English (Great Britain) – one of the constituent corpora of ICE
17	DCPSE	the Diachronic Corpus of Present-Day Spoken English
18	DSEU	a mini-corpus consisting of an early part of the DCPSE
19	DICE	a mini-corpus consisting of a later matching part of the DCPSE
20	ANC	the American National Corpus
21	ARCHER	A Representative Corpus of Historical English Registers
22	LCSAE	the Longman Corpus of Spoken American English
23	CIC	the Cambridge International Corpus
24	ACE	the Australian Corpus of English
25	CONCE	Corpus of Nineteenth-Century English
26	MICASE	the Michigan Corpus of Academic Spoken English

Alphabetical index to the above list:

ACE 24	Brown family 5	Fiction 11	Lanc-31 6
ANC 20	CIC 23	F-LOB 4	LCSAE 22
ARCHER 21	CONCE 25	Frown 3	Learned 10
B-LOB 7	DCPSE 17	General Prose 9	LOB 2
BNC 12, 13, 14	DICE 19	ICE 15	MICASE 26
Brown 1	DSEU 18	ICE-GB 16	Press 8

B. Abbreviations for Geographical and Historical Subdivisions of English

AmE	American English	LModE	Late Modern English
BrE	British English	ME	Middle English
ModE	Modern English	OE	Old English
EModE	Early Modern English	PDE	Present-Day English

C. Other Abbreviations

C8	The C8 tagset: a set of part-of-speech tags used for annotating the Brown family of corpora (the C8 tags are listed in Appendix II)
CLAWS	Constituent-Likelihood Automatic Word Tagging System (a POS tagger)
CLAWS4	The newest version of CLAWS

CQP	Corpus Query Processor (software: a tool for interpreting corpus queries)
LL	Log likelihood (a measure of statistical significance)
NP	Noun phrase
N+N	Sequence consisting of noun + noun
N+CN	Sequence consisting of noun + common noun
OED	*Oxford English Dictionary*
pmw	Per million words (in statistical tables, frequencies are often normalized to this standard frequency measure)
PN+PN	A sequence of proper noun + proper noun
POS	Part of speech (used especially in the collocation 'POS tagger/tagging')
XML	Extensible Markup Language (an artificial metalanguage used for the encoding and processing of textual material, including corpora)

D. Conventions

[Brown L12], [LOB A09], and the like	These are address labels used to identify the whereabouts, in the Brown family of corpora, of a particular example, sentence, etc. After the corpus name, the letter indicates the text category and the two digits the number of the text sample in that category. Similar address labels are used for examples from the BNC and other corpora.
*	An asterisk before an (invented) example indicates its status as an unacceptable or ungrammatical usage.
?	A question mark before an example (invented or otherwise) indicates its questionable acceptability.
[. . .]	In a corpus example, an ellipsis in square brackets indicates where the example has been simplified by the omission of part of the original corpus sentence.
*, **, ***	Placed next to a numerical quantity in a statistical table or bar chart, these are indicators of increasingly higher statistical significance.
*	* means 'significant at the level $p < 0.05$ (LL > 3.84)'.
**	** means 'significant at the level $p < 0.01$ (LL > 6.63)'.

***	*** means 'significant at the level $p < 0.001$ (LL > 10.83)'.
N* etc.	In referring to POS tags, an asterisk is occasionally used as a 'wildcard symbol', standing for any number (including zero) of characters, excepting a space or other delimiting character. For example, N* will identify any tag beginning with N, which means, in fact, any noun in the C8 tagset.
HAVE got to, *SHE*, *HAVE*, *NEED to*, and the like	In certain chapters, the small capitals indicate that the word cited is understood as a lemma, not as an individual word form. For example, *HAVE to* signifies any form of the verb *HAVE* followed by *to* (i.e. *have to*, *has to*, *had to*, *having to*). The chapters in which this convention chiefly applies are 4 and 5. It is important to avoid confusion in some contexts by using this convention. In other contexts the convention is unnecessary, as the interpretation of a graphical form like *be going to* is clear from the context. Hence we use this convention only in some chapters.