Analysing Variation in English

EDITED BY
Warren Maguire and
April McMahon

CAMBRIDGE

CAMBRIDGE www.cambridge.org/9780521898669

Analysing Variation in English

Analysing Variation in English brings together a range of perspectives on the collection, analysis, and broader relevance of variable language data. In the first half of the book, the focus is firmly on the description and comparison of methods for collecting and analysing examples of variation in language. Novel quantitative and computational methods are introduced and exemplified alongside more traditional approaches. The innovative second half of the book establishes and tests the relevance of language variation to other aspects of linguistics such as language change, and to other disciplines such as law and education. Each chapter concludes with a 'Where next?' section, providing guidance on further reading, but also pointers to under-researched areas, designed to help identify good topics for projects and dissertations. Designed to be used by students as well as researchers, the book will be welcomed by those working in English language and linguistics, sociolinguistics or language change.

WARREN MAGUIRE is a lecturer in English Language in the Department of Linguistics and English Language at the University of Edinburgh.

APRIL MCMAHON is Forbes Professor of English Language in the Department of Linguistics and English Language at the University of Edinburgh.

Analysing Variation in English

Edited by

Warren Maguire

and

April McMahon

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore,
São Paulo, Delhi, Dubai, Tokyo, Mexico City

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/9780521898669

© Cambridge University Press 2011

This publication is in copyright. Subject to statutory exception
and to the provisions of relevant collective licensing agreements,
no reproduction of any part may take place without the written
permission of Cambridge University Press.

First published 2011

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
Analysing Variation in English / [edited by] Warren Maguire, April McMahon.
p. cm
Includes bibliographical references and index.
ISBN 978-0-521-89866-9
1. English language–Variation. 2. Linguistic change. I. Maguire, Warren
II. McMahon, April (April M. S.) III. Title.
PE1074.7.A53 2011
427.0072–dc22
2010045719

ISBN 978-0-521-89866-9 Hardback

Cambridge University Press has no responsibility for the persistence or
accuracy of URLs for external or third-party internet websites referred to in
this publication, and does not guarantee that any content on such websites is,
or will remain, accurate or appropriate.

Contents

Figures

Tables

Contributors

JOAN BEAL, University of Sheffield

ISABELLE BUCHSTALLER, Newcastle University

KAREN CORRIGAN, Newcastle University

ALEXANDRA D'ARCY, University of Victoria

GREGORY R. GUY, New York University

PATRICK HONEYBONE, University of Edinburgh

WARREN MAGUIRE, University of Edinburgh

APRIL MCMAHON, University of Edinburgh

ROB MCMAHON, University of Edinburgh

HERMANN MOISL, Newcastle University

CHRIS MONTGOMERY, Sheffield Hallam University

EMMA MOORE, University of Sheffield

FRANCES ROCK, Cardiff University

ERIK R. THOMAS, North Carolina State University

GRAEME TROUSDALE, University of Edinburgh

Abbreviations

Corpora and tools for corpus analysis

ACE	Australian Corpus of English *aka* Macquarie Corpus of Written Australian English
AFAM	African American and Gullah Project
ANAE	Atlas of North American English
ANC	American National Corpus
ARCHER	A Representative Corpus of Historical English Registers
BNC	British National Corpus
Brown	Brown University Standard Corpus of Present-Day American English
CELEX	Dutch Centre for Lexical Information
CLAWS	Constituent Likelihood Automatic Word-tagging System
COBUILD	Collins Birmingham University International Language Database
COCA	Corpus of Contemporary American English
CofP	Community of practice
COLT	The Bergen Corpus of London Teenage Language (now part of the BNC)
CONCE	Corpus of Nineteenth-century English
CONTE-pC	Corpus of Early Ontario English, pre-Confederation section
FLOB	Freiburg-Lancaster-Oslo-Bergen Corpus
Frown	Freiburg-Brown Corpus of American English
ICAME	International Computer Archive of Modern and Medieval English
ICE	International Corpus of English
LAEME	Linguistic Atlas of Early Middle English
LAGS	Linguistic Atlas of the Gulf States
LAMSAS	Linguistic Atlas of the Middle and South Atlantic States
LANCS	Linguistic Atlas of the North Central States
LANE	Linguistic Atlas of New England
LAOS	Linguistic Atlas of Older Scots

LAS	Linguistic Atlas of Scotland
LAUM	Linguistic Atlas of the Upper Midwest
LDC	Linguistic Data Consortium
LOB	Lancaster-Oslo-Bergen Corpus
LSWE	Longman Spoken and Written English Corpus
MICASE	Michigan Corpus of Spoken Academic English
NECTE	The Newcastle Electronic Corpus of Tyneside English
ONZE	Origins of New Zealand English
PVC	Phonological Variation and Change in Contemporary Spoken English
SBCSAE	Santa Barbara Corpus of Spoken American English
SCOTS	Scottish Corpus of Texts and Speech
SED	Survey of English Dialects
SEU	Survey of English Usage
SURE	Survey of Regional Englishes
TEI	Text Encoding Initiative
TLS	Tyneside Linguistic Survey

Others

AAE	African American English
BP	Before present
CSD	Coronal Stop Deletion
HGP	Human Genome Project
HPSG	Head-driven phrase structure grammar
IViE	Intonational Variation in English
LPC	Linear predictive coding
NC	National Curriculum (UK)
NORM	Non-mobile older rural male
NSR	Northern Subject Rule
OCP	Obligatory Contour Principle
OT	Optimality theory
POS	Part of speech
SNP	Single nucleotide polymorphism
SR	Surface representation
SVO	Subject–Verb–Object (word order)
ToBI	Tone and Break Index
UR	Underlying representation
VOT	Voice-onset time

Introduction. Analysing variation in English: what we know, what we don't, and why it matters

April McMahon and Warren Maguire

Variation in language is ubiquitous. It is both highly structured and sometimes perplexing; it correlates with external factors, which might be social, or geographical, or something else entirely, but it also follows its own rules and arises for its own, language-internal reasons; it is constant, in the sense that some sort of variation is always there, but it changes its locus within the language across generations, and is a crucial ingredient in language change. Linguists sometimes shy away from variation: it gets in the way when we want to describe straightforwardly 'what happens in English', and meet the response 'not in my dialect'. Sometimes it is used as a default explanation; but at the same time, many recent approaches to linguistic theory see variation 'as a core explanandum' (Adger and Trousdale 2007: 274). These paradoxes can be infuriating and challenging, and linguists may choose to engage more or less with variation and its consequences, but the existence of variation is incontrovertible, and, in our view at least, the collection, analysis and explanation of variable data is one of the most lively and fascinating challenges of current linguistics.

Linguistic variation is also subject to a range of complementary and competing approaches and perspectives. The existence of a range of conferences dedicated specifically to work on variation and its historical corollary, change in progress, provide evidence of the liveliness and popularity of the field: so, we find regular meetings in the series Methods in Dialectology, UK Language Variation and Change, and NWAV; while slots and sessions at the Sociolinguistics Symposium, and the International Conferences on English Historical Linguistics, Historical Linguistics, and the Linguistics of Contemporary English are regularly occupied by papers on variation, change and their intersection. There are workshops, papers and books on analysing variation within theoretical approaches from optimality theory to cognitive grammar to construction grammar; laboratory phonologists debate where variation comes from, while evolutionary linguists place it in a more general context of cultural evolution and diversification. Variation in English (and indeed in other languages) is also an extremely popular area with students, and there are many courses in this area, from general to highly theoretical and specific,

and a host of undergraduate projects and postgraduate dissertations and theses researched and written every year.

There are already many introductory and advanced textbooks, handbooks, monographs and journals on variation and change, and on varieties of English both past and present: for outlines of individual varieties, see Kortmann, Schneider, Burridge *et al*. (2004), and the Dialects of English series from Edinburgh University Press; and for overviews of the history of English and of Scots, see the monumental Cambridge History of the English Language (Hogg 1992–2001), Jones (1997), the more recent 'baby CHEL' (Hogg and Denison 2006) and Mugglestone (2006), for instance. In this book, however, rather than provide descriptions of individual varieties, or accounts of variation within individual theoretical frameworks, we have a different, more general, and dual focus. In Part I, we consider methodological issues on how variable language data can be collected and analysed. In Part II, we turn to the relevance of variation, building on Adger and Trousdale's (2007: 274) view that 'furthering our knowledge of syntactic variation in English dialects is of relevance to a range of different "kinds" of linguists', but extending beyond syntax, and indeed beyond linguistics. In brief, we ask **how** and **why** variation should be studied.

Our aim is also to provide assistance to students, not just by giving overviews and background reading, but also by pointing to areas where work is needed. The current focus on project work and first-hand dissertation research for undergraduate as well as graduate students has led to a need for help in identifying likely projects, and therefore in finding information on under-researched areas. Even quite advanced students may not be familiar with the whole range of methodologies through which language variation can be investigated and, since new methods are emerging rather rapidly, nor may their advisers. Authors of each chapter have therefore made their discussion accessible to students who may have taken only fairly elementary courses on variation, but also write at a level suitable for a colleague who might work in another sub-area of variation, and needs a quick but reliable update. At the end of each chapter, they have also provided some suggestions for the next steps interested readers can take in investigating a topic. These 'Where next?' sections always include ideas for further reading, but they often highlight areas that urgently require further research too.

1 Investigating variation in English: how do we know what we know?

Chapters in this first section focus on methods used to analyse variation, and in each case consider the benefits and limitations of the methods at issue, along with an indication of the situations in which each method has been applied, and

those where it might be helpful but has not yet been used. The central questions here are how we might most reliably gather data demonstrating variation; how those data can then be analysed, stored and presented; and how different methods can be compared and validated.

In the first two chapters, Thomas and Buchstaller & Corrigan discuss methods, both established and emerging, for the collection of data in phonology and in morphosyntax, respectively. D'Arcy considers protocols for the construction, sharing and maintenance of corpora, and asks and answers fundamental questions on what a corpus is, and how corpora should be used; this chapter leads into Moisl's more general discussion of how we decide what questions to ask of our data, or how we generate the hypotheses we aim to test. In Moisl's chapter and our own, we have chosen to focus primarily on more mathematical and computational techniques, partly because there is already plentiful coverage of more standard interview and questionnaire-based methods in the sociolinguistics literature, and partly because so many historical, dialectological and typological projects are now inclining towards methods which involve maps, trees and networks. There is rather little non-technical coverage of such techniques in the literature, especially aimed at students; again, however, we envisage these chapters as providing a helpful overview also for colleagues who may be interested in the possibilities these new methods offer, but may not have the time or inclination to engage immediately with the more technical primary literature. Finally, Montgomery and Beal's chapter provides a helpful and up-to-date overview of developments in perceptual dialectology: increasingly, the viewpoints of speakers are being included in accounts of variation in sociolinguistics, for instance in approaches based around communities of practice (e.g. Eckert 2000), and perceptual dialectology encourages a similar integration in dialectological work.

2 Why does it matter? Variation and other fields

In the second section, we step outside studies of variation per se, to assess the importance of their results for other fields, and vice versa. Each chapter outlines the relevance of linguistic variation for either another area of linguistics, or another discipline, again with some consideration of areas that remain unclear or under-investigated. Authors focus on the ways in which investigations of variation in English can be integrated with research elsewhere, and likewise how results from cognate subject areas can help us understand variation.

Through these six chapters, there is a gradual progression outwards from the relevance of variation to other sub-disciplines within linguistics, notably linguistic theory and historical linguistics in the chapters by Honeybone and Guy, to interfaces between linguistics and other disciplines which are constructed primarily through work on variation. Understanding variation and

describing it fully is clearly vital for forensic linguistic work, as Rock demonstrates; and Moore's chapter shows that variation is also key to new approaches in sociolinguistics, as speakers use their repertoire to construct and express their identity. McMahon proposes that data from genetics can help us evaluate hypotheses on possible sources of past or present linguistic variation by tracing the histories of genetic markers in individuals and local populations. This sort of approach is also reflected in the work of the Centre for the Evolution of Cultural Diversity at UCL in London, and in a range of applications of the 'new synthesis' between archaeology, linguistics and genetics. Finally, the use and analysis of non-standard varieties in schools makes variation a hot topic in educational linguistics, as Trousdale shows.

We see an automatic progression in the book between the first section on methods, and this section on applications. Inevitably (and in our view entirely properly), students tend to ask why they should be interested in particular modes of study, and in particular kinds of data; and what they typically mean here is, what relevance does this kind of research have for questions that might be being asked in the 'real world', or in areas I might wish to move into when I have finished my studies? The chapters in Part II seek to answer these (sometimes implicit) questions, and also allow connections to be made across disciplines and sub-disciplines; this accurately reflects the increasingly interdisciplinary character of work on variation in language. At the same time, however, results from research on variation can only be truly relevant if they are reliable, and hence if the data have been collected, analysed and presented through the methods discussed in Part I. The first set of chapters is therefore a prerequisite for the second, and the second perhaps a series of reasons for getting properly to grips with the methods in the first. Together, these chapters add up to a picture of how we know what we know about variation in English; which methods of investigation are used and how these are likely to change; and why these findings and methods are relevant for disciplines and sub-disciplines sometimes quite distant from our own. We hope the book will encourage students and colleagues to find out more and to fill some of the gaps identified in these chapters.

Part I

Investigating variation in English: how do we know what we know?

1 Collecting data on phonology

Erik R. Thomas

1.1 Introduction

Few problems have engaged the creativity of language variationists to the extent that the collection of phonological data has. In studying phonology, researchers have to discern how phonetic variation fits together to form phonological primitives. The variation may be phonetic in nature, that is, dependent on factors such as rate of speech, degree of stress or other prosodic factors, and elasto-dynamic constraints on articulators. It may also be due to social factors, as with style-shifting and social and class variation. In addition, researchers have to consider how variation interacts with the speech production/speech perception opposition. The means of studying production generally involve impressionistic auditory transcription or acoustic analysis, while analysis of perception usually entails cognitive experiments. Different kinds of variables also require different approaches. As broad categories, consonants, vowels, prosody, and, though it has barely been studied by variationists, voice quality, all require distinct sorts of analyses, and within each category individual variables need their own kinds of analysis.

The shifting sands of theory and technology create more challenges. Theoretical stances in phonology, such as generativism, autosegmental phonology, optimality theory, and exemplar theory, have at times induced variationists to adjust aspects of how they study data. However, variationists have often been content to let phonology work out its own issues without adapting phonological theories to sociolinguistics or vice versa (see Honeybone, this volume). At the same time, changes in the focus of study, from geographical variation to social variation to the behaviour of 'communities of practice', have resulted from theoretical developments in dialectal studies. In addition, technological innovations – statistical packages, digitisation of recordings, spectrographic analysis, speech synthesis, and perhaps soon, brain scanning – continually change how phonological variation is studied. Nevertheless, variationists have proved quite able to adapt to all of these factors and influences.

Variation in phonology and phonetics can serve as a proving ground for hypotheses in those topics, as well as a source of new hypotheses. Docherty

et al. (1997) discuss the tension between 'top-down' approaches to phonology, in which hypotheses are formed on the basis of a small body of evidence and before empirical testing, and 'bottom-up' approaches, in which surveys of speakers are conducted before theories about phonological organisation are constructed. They consider at length one example, the glottalisation of voiceless stops in the accent of Tyneside in northern England. For this example, the bottom-up approach favoured by sociolinguists appears superior to the top-down approach favoured by formal phonologists because surveying sufficient numbers of speakers produces cases that violate expectations of top-down hypotheses. Moreover, the survey produced other, unexpected results, such as a disfavouring of glottalisation before a pause, which differs from patterns found in other dialects.

The remainder of this chapter will survey approaches taken over the years to discerning phonology by means of examining dialectal and sociolectal variation. Dialect geographers generally followed methods that reflected the phonological theories of their time and tended to focus on variation in segmental production. Sociolinguists have also been somewhat constrained by phonological theories. However, they have gradually expanded into new areas of variation, such as sociolectal variation, speaking style, perception, and intonation. They have also integrated acoustic and statistical analysis into the study of linguistic variation. Yet there remain significant areas that are hardly touched, such as voice quality.

1.2 Dialectology

Linguistic geographers traditionally used the method of sending fieldworkers out to local communities with a questionnaire. The questionnaire usually contained a mixture of questions to elicit lexical, phonological/phonetic, and morphological data. For example, a fieldworker might ask 'What would you call two animals worked together?' to elicit the word *oxen*, which was used as an example of the LOT vowel in the American linguistic atlas projects.[1] With regard to phonetics and phonology, the fieldworker had to be proficient at fine-grained impressionistic phonetic transcription in order to record the phonetic variants that distinguish dialects of English within Great Britain and North America. The fieldworkers were required to make transcriptions on the spot because, especially in the earlier projects, the interviews were not taped or otherwise mechanically recorded. The system worked well when fieldworkers were expert transcribers. However, some projects suffered from poor transcriptions by fieldworkers. For example, the *Linguistic Atlas of the North Central States* (LANCS), which covers parts of the American Midwest, employed a range of fieldworkers who varied from experts to novices, and the transcriptions they produced reflect that, creating comparability problems. In addition,

even the best fieldworkers often differed in their transcription norms, leading to 'fieldworker isoglosses' (e.g. Trudgill 1983: 38–41), in which false dialectal boundaries appear that are actually boundaries between territories covered by different fieldworkers.

Nevertheless, this system produced vast amounts of usable and informative data. Important works illustrating the findings of dialect geography for pronunciation include, among others: *A Structural Atlas of the English Dialects* (Anderson 1987), the *Linguistic Atlas of New England* (LANE; Kurath *et al*. 1939–43), Kurath and Lowman (1970), Kurath and McDavid (1961), the *Survey of English Dialects* (SED; Orton, Sanderson and Widdowson 1978), the *Linguistic Atlas of the Gulf States* (LAGS; Pederson *et al*. 1986–92), and the *Linguistic Atlas of the Upper Midwest* (LAUM; Allen 1976). They each show regional phonetic variation, such as [uː~ɔʊ~aʊ~æʊ~ɛʊ~aː] for the MOUTH vowel, extensively. They also show phonological differentiation, such as the FACE/DAY and GOAT/TOW mergers and the TRAP/BATH split in England or the NORTH/FORCE and LOT/THOUGHT mergers in North America. Finally, they put considerable emphasis on the lexical incidence of phonemes, as with whether the FOOT, GOOSE, or STRUT vowel occurs in such words as *room*, *roof*, *root*, and *Cooper*.

The most recent dialect geography projects, most notably LAGS, have tape recorded all interviews. The interviews were transcribed later by trained phoneticians. This procedure allows the transcriptions to be checked for accuracy. The survey of the United States conducted for the *Dictionary of American Regional English* included tape recordings for about half of its subjects that are now available to scholars, and the SED taped excerpts of its interviews.

Editors resorted to numerous methods of processing and presentation of linguistic atlas data. The narrow phonetic transcriptions were themselves a challenge. LANE simply mapped each transcription in a folio-sized publication. That approach soon became too expensive, however. The phonetic transcriptions for SED were published as a multi-volume book (Orton and Dieth 1962–71). The field records from LAGS and two other American projects, the *Linguistic Atlas of the Middle and South Atlantic States* and the *Linguistic Atlas of the North Central States*, were published on microfilm. A more selective approach was used in Kurath and McDavid (1961) and Orton *et al*. (1978). The most important aspects of these publications were maps that showed the distributions of dialectal variations in phonetic forms, or *diaphones* (the term used by Kurath and McDavid). Diaphones were represented as symbols in Kurath and McDavid, while they were shown as zones separated by isophones (phonological or phonetic boundaries analogous to isoglosses) on the maps in Orton *et al*. Kurath and McDavid showed isophones only occasionally. Another selective approach was used when LAUM was published; the volume

that covered pronunciation (Allen 1976) listed variants and showed small interpretive maps for certain keywords.

A different kind of interpretive map is found in Kurath and Lowman (1970) and Anderson (1987), which covered two unrelated surveys of England (Anderson the same one as Orton and Dieth 1962–71 and Orton *et al*. 1978). Both use symbols plotted on maps to summarise data from numerous elicited words with a particular sound. Kurath and McDavid show the number of words with a particular diaphone out of the total number that have the respective phoneme. Anderson shows percentages of words instead. Mergers are shown by Anderson as the percentage of words with the merged pronunciation.

Kurath and McDavid (1961) and Allen (1976) had one additional way of representing vowel variants: idiolect synopses. An idiolect synopsis consists of a table that lists the phonetic transcription for the vowel in each word in a set that were elicited, including two or three for each phoneme. Words were sorted into columns representing each phoneme. The synopsis thus allows readers to see what contrasts a speaker makes. LAGS also employed idiolect synopses, though they were published on microfiche instead of in a book. Figure 1.1 shows an idiolect synopsis assembled from the field records of a linguistic atlas participant and modelled after the synopses in Kurath and McDavid (1961).

Dialectologists usually made just a few general assumptions about phonology – for the most part, the existence of phonemes and contrastiveness, a distinction between phonological and phonetic representations, and primacy of production over perception. The American linguistic atlas projects were somewhat tied to structuralist theories of phonology, particularly those of George Trager and Bernard Bloch. For example, they recognised three levels of phonemic vowel height and three possible types of glides, /h/ (for inglides), /w/, and /y/. As a whole, though, dialectology did not serve as a source of new phonological theories. An exception was *The Linguistic Atlas of Scotland* (LAS; Mather and Speitel 1986). LAS introduced the notion of the 'polyphoneme', in which phones were grouped into ten types, or polyphonemes, based on their phonetic similarity. Contrastive sounds could be subsumed within one polyphoneme. The presentation obscured both contrastiveness and the degree of phonetic variation and was not adopted by any other projects. Nor did it gain a following among phonologists.

The most concerted effort to modernise dialectological data presentation appears in LAGS. LAGS was begun during the late 1960s and emphasised some of the independent variables used in sociolinguistic studies: social class, ethnicity, gender, and age cohort. When LAGS was published, volumes were devoted to those factors, and even the volumes on geographical variation showed geography in conjunction with other independent variables. The treatment of geography differs sharply from other dialect geography publications as well. Whereas earlier works showed the responses of individual speakers, LAGS

	i	ɪ	e	ɛ	ɜ	æ	ɑˑ	ɑ	ai	ɔi	au	ɒ	ɔ	ʌ	ə	o	ʊ	u	
three	i̱i																	ʉ̱ʉ̱	two
grease	j̱i																	ʊʉ	tooth
six		ɪ															ʊ˳		wood
crib		ɪˑ˳															ʊ̱ˑ˳		pull
ear	i̱ə																	ʉ̱ə	poor
beard	ʲəˑ															ɜˑʊ̱			ago
eight			ɛɨ													ɜʊ			coat
April			ɛɨ													ɜʊ			road
ten		ɪˑ˳														ɜʊ			home
egg				ɛ˳												ɜʊ̱			know
head				ɛˑ˳												ọ˚ə			four
Mary			ɛɨ													ọ˚ə			door
stairs						æ̱əˑ										ọ˚ə			hoarse
care						æ̱ə										oˑə			mourn
merry				—															
thirty					ˢəˑ									ʌˢ					sun
sermon					əˑ									ʌˑ˳					brush
furrow					əˑ														
ashes						æˑ							ɔˑ˳						frost
bag						æˑ˳		ɑ̣˚											log
married						æ̱							ɔː						dog
half						æ˳							ɔˑ˳						water
glass						æ˳							ɔˑ						daught
aunt						æˑ˳							ɔˑ						law
father								ɑ̣˚											
palm						æˑə		ɑ̣ːˑ											
barn								ɑ̣ə					ɔˑə						forty
garden								ɑ̣ə					ɔˑə						morning
crop								ɑ̣ːˑ					ɔˑə						corn
John								ɑ̣ːˑ					ɔˑə̣						horse
college								ɑ̣˚											
borrow								ɑ̣˚											
five									ɑˑɨ		æˑə̱								down
twice									ɑˑɨ		æ̱ɤ								out
wire								ɑ̣ːə			ạʊə								flower
										ɔˑɨ									joint
										ọˑɨ									boil
	i	ɪ	e	ɛ	ɜ	æ	ɑˑ	ɑ	ai	ɔi	au	ɒ	ɔ	ʌ	ə	o	ʊ	u	

Figure 1.1 Idiolect synopsis for LAMSAS participant NC 11B, a European American female, born 1897, from Hyde County, North Carolina. The format follows that of the idiolect synopses in Kurath and McDavid (1961). A few symbols have been modernised from the original transcriptions

grouped speakers into regions that were delineated by features of the physical landscape. Results were then shown collectively by region. The physiogeographic features dictated farming practices and industry in the LAGS territory, thus attracting different settlers with differing origins and social classes, which made the divisions relevant to dialectal features. In most volumes, only one or two elicited words from each phoneme were shown, and the number and percentage of speakers in a particular category who have a certain diaphone

are listed. The maps highlight regions with especially high incidences of that diaphone.

The use of telephone surveys has pumped new life into the study of geographical differences in pronunciation. Guy Bailey and his colleagues conducted pioneering work in Texas and Oklahoma (e.g. Bailey, Wikle and Sand 1991; Bailey, Wikle, Tillery and Sand 1991, 1993). They inserted a few questions eliciting particular words into public polls that were primarily used to gauge political opinions, knowledge of health issues, and similar topics and then coded the responses from tape recordings of the interviews. The results provided considerable information about the regional distribution of the variants examined, including some related to phonological mergers, as well as how the variants were correlated with various social factors (Bernstein 1993). An important advantage of the methodology was that it came as close to a random sample as any dialectal survey had ever come. A disadvantage was that they could investigate only a few variables.

Telephone surveying was carried out on a much larger scale for the *Atlas of North American English* (ANAE; Labov, Ash and Boberg 2006). ANAE surveyed urban centres over all of North America, including every state in the United States and every Canadian province. The sample was designed to be random, but with subjects who were not natives of their community of residence screened out. Unlike Bailey and his colleagues' work in Texas and Oklahoma, the interview consisted entirely of linguistic questions, most of which targeted phonetic or phonological variables. As a result, the survey could include dozens of linguistic questions and every vowel was elicited. Respondents took part in two interviews. The first one had a question-and-answer format like those used for linguistic atlases. The second interview consisted of respondents reading a wordlist that the researchers mailed them. One innovation over previous geographical surveys was that the tokens were measured acoustically and the speech production results consisted of interpretations of the acoustic measurements. They devoted considerable attention to certain mergers, such as the LOT/THOUGHT merger.

1.3 Sociolinguistics

1.3.1 Elicitation techniques of sociolinguistics

Consistent use of audio recording has been a mainstay of sociolinguistic practice. The only situation when sociolinguists do not record interviews in studies of speech production is for 'rapid anonymous surveys', in which large numbers of subjects are approached anonymously and asked only a few questions. The best known rapid anonymous survey is the first published sociolinguistic one, the well-known department store survey in New York City (Labov 1966).

Labov went to three department stores in New York, a working-class store, a middle-class store, and an upper-class store. He asked numerous employees at each store where to find something located on the fourth floor, and when the employee said *fourth floor*, he noted whether the employee pronounced the *r* in each word. The survey showed that, as the stores increased in prestige, the rhoticity, or *r*-fulness, of its employees increased.

Whereas dialect geographers relied primarily on a question-and-answer format to collect data and used conversation only secondarily (though many fieldworkers did use conversation a great deal), sociolinguists have made spontaneous speech, usually interview-style conversation, their primary mode of collecting data. The reason for the emphasis on conversation is that, as Labov (1966) demonstrated clearly, speech styles differ considerably and formal variants that are seldom present in more casual speech often predominate in citation-form speech. Many of the important early sociolinguistic studies of phonological/phonetic variation were based largely on spontaneous speech, usually in interviews with the fieldworker(s) (e.g. Labov 1963, 1966; Wolfram 1969; Trudgill 1974). Later studies have perpetuated the use of spontaneous speech as the most important source of data (e.g. Horvath 1985; Labov 1994; Kerswill and Williams 2000; Wolfram and Thomas 2002), most often in interviews but sometimes in dyadic conversations between subjects, even when they used it in conjunction with other kinds of speech, as Horvath (1985) did. Extracting usable tokens from conversation is more time-consuming than doing so from other sources because the researcher has to listen to long stretches of speech in order to find sufficient numbers of tokens. Furthermore, there is no control over the words produced and some sounds may be underrepresented or unrepresented. However, the tokens are more naturalistic than those from any other sort of speech.

Sociolinguists have employed other methods along with eliciting conversation, however. Labov (1966) added reading passages and two types of wordlists to the conversation he collected. Dialect geographers had already been experimenting with reading passages, and the *Dictionary of American Regional English* embarked on a survey of the entire United States that included tape recording readings of a story called 'Arthur the Rat'. Whereas 'Arthur the Rat' was designed to capture as many different variables in a short passage as possible, Labov devised reading passages that targeted particular variables by including many words with that variable in them. Both types of stories have been used subsequently in various studies.

The two kinds of wordlists that Labov used were simple wordlists and minimal pair lists. Simple wordlists require subjects to read a list of isolated words, usually arranged so that the subject cannot guess what variables are being tested. Minimal pair tests, in contrast, require subjects to read pairs of words, most of which differ by only a single sound. Some of the pairs are distractors,

words that are pronounced differently or the same by everyone. Others, though, are words for which people differ in whether they are pronounced alike or differently. Such pairs can be used as one method of investigating phonological mergers. Frequently, subjects are asked for their own judgements about whether members of a minimal pair sound alike or different.

Labov (1966) treated speech style as a linear factor determined by the amount of attention the speaker paid to his or her speech. The level of attention paid to speech was considered to increase with the formality of the speech style. He elicited five speech styles, casual conversation, interview-style conversation, reading passage style, wordlist style, and minimal pair style, in increasing order of formality. He demonstrated clearly that linguistic variables were correlated with his style scale. Subsequent studies, most notably Bell (1984), have demonstrated the inadequacy of viewing style solely in terms of attention paid to speech. Bell argued that speakers modify their speech according to the audience they are addressing. However, finding ways to quantify style in terms of other factors has proved difficult. One attempt is that of Schilling-Estes (2004), who examined how phonological/phonetic and other linguistic variables vacillated in a conversation according to the topic being discussed.

Sociolinguists have always insisted that stylistic variation is internalised. Weinreich, Labov and Herzog (1968) formulated this idea into the notion of 'structured homogeneity'. That is, variation is not random, but is correlated with social class and other social factors and with speaking style. More contentious is the notion that these patterns are phonologised in speakers' minds. This idea formed the basis for 'variable rules' during the 1970s. Variable rules were a response to generative phonology, which predominated at the time. They consisted of formal phonological rules that generated probabilistic outputs according to the social class of the speaker and the speaking style. C.-J. Bailey (1973) even formulated a notion called the 'isolect', in which each social class/speaking style combination represented a separate speech form (sometimes identical with that of other class/style combinations) with its own probabilities. Although sociolinguists have abandoned variable rules (Fasold 1991), they certainly still hold that knowledge of the stylistic appropriateness of phonological and phonetic variants is part of a speaker's linguistic competence.

1.3.2 Transcription of data in sociolinguistics

When sociolinguists appeared on the scene in the 1960s, they used impressionistic transcription exclusively for some years, following the practice of dialectologists. It remains the most common data coding technique for all kinds of variables. Reliance on impressionistic transcription is particularly pervasive for consonantal variables. The most heavily studied consonantal variables are probably rhoticity/non-rhoticity (*r*-fulness/*r*-lessness) – that is,

retention vs. loss of /r/ or mutation to [ə] in words such as *here* and *hard*; consonant cluster reduction, as in *past* being pronounced as *pas'*; and mutation of historical /θ/ (as in *thick* and *both*) and /ð/ (as in *this* and *brother*), that is, [θ~tθ], [t], and [f] for /θ/ and [ð~dð], [d], and [v] for /ð/. All of these variables have consistently been analysed impressionistically.

Sociolinguists typically do not transcribe entire words and phrases in narrow phonetic symbols, as dialectologists did. Instead, they usually transcribe only the variable – a particular segment or intonation contour – in question. They also do not ordinarily transcribe sounds as narrowly as dialect geographers did, often simply coding tokens in a binary or other simple manner. For example, Wolfram (1969) tabulated the incidence of *r*-ful and *r*-less pronunciation in words such as *here* and *hard* as a binary variable and of various reflexes of historical /θ/ and /ð/ (e.g. for the variable θ) as [θ], [tθ], [t], or [f]. Because his concern was with social correlates of linguistic variables, he had no need to transcribe the rest of the words in which those sounds occurred.

Of special concern in sociolinguistics are segmental phonological mergers. They have been examined using an array of methods, including impressionistic transcription, acoustic analysis, speaker judgements, and perception experiments. I will discuss them at greater length in §1.4.2.

Transcription of intonation used to be exclusively impressionistic. In addition, intonation studies tend to focus on intonational meaning or on cross-linguistic differences. As a result, dialectal studies of intonational differences are not extensive. Moreover, dialectal work is less well developed in English than in some continental European languages. A few dialectal variations in intonation have been examined thus far. Most earlier studies used purely impressionistic transcription and tended to focus on whether tones show rising, level, or falling patterns or some combination, such as a fall followed by a rise, or vice versa, as well as whether rises and falls start from a high, mid, or low tone. Examples include Pellowe and Jones's (1978) analysis of Tyneside intonation and Douglas-Cowie, Cowie and Rahilly's (1995) study of Belfast intonation. In fact, intonational patterns in northern England, Scotland, and Northern Ireland differ significantly from those of southern England and from each other and thus have attracted a number of dialectal studies: see Cruttenden (1995) for a review. In North America, the earlier studies of how African American intonation differs from European American intonation (Tarone 1973; Loman 1975) used impressionistic transcription.

In recent years, intonational analyses have generally employed the Tone and Break Index (ToBI) transcription system (Beckman and Hirschberg 1994). One important innovation ToBI has brought to intonational research is combining impressionistic transcription with acoustic analysis. Researchers examine spectrograms with superimposed autocorrelation pitch tracks, which

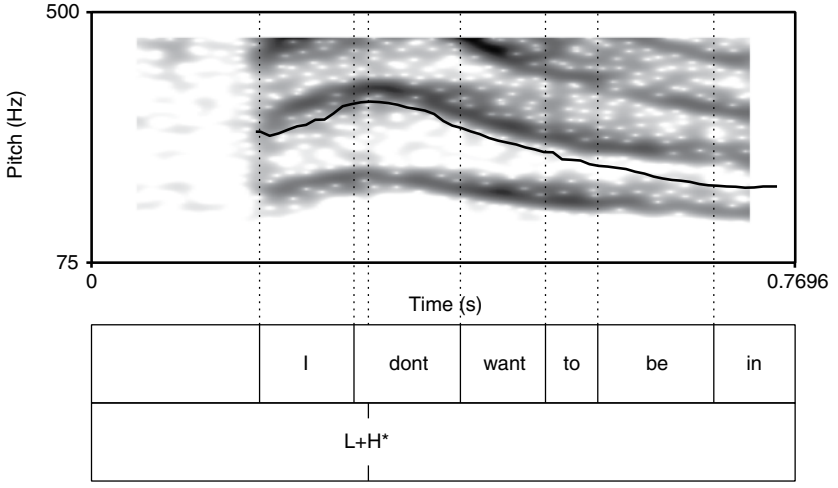


Figure 1.2 Pitch track superimposed on narrowband spectrogram, with ToBI tone tier, for the phrase 'I don't want to be in' uttered by a European American female from Warren County, North Carolina. ToBI break index and miscellaneous tiers are not shown

provide estimates of the fundamental frequency (F_0), and then impressionistically identify kinds of tones based on the pitch tracks. It treats intonation in terms of two phenomena: pitch accents, which are pitch movements that stand out prominently, and edge tones, which are the contours that occur at the end (and occasionally the beginning) of 'intonational phrases' and 'intermediate phrases'. Both pitch accents and edge tones are differentiated into different types depending on whether F_0 falls, rises, or is stable within them and, for certain pitch accents, depending on whether a rise or fall is immediate or delayed. The system also includes a 'break index', which classifies any kind of boundary into a hierarchy ranging from the word level up to the intonational phrase level. ToBI is essentially a phonological system and, hence, assumes that all transcribed forms are phonologically specified. That is, a pitch accent such as H* or a boundary tone such as L% is assumed to be equivalent to a phoneme. Figure 1.2 shows a spectrogram and superimposed pitch track of an L+H* pitch accent. L+H* denotes a pitch accent that begins with a relatively low F_0 and rises to a high plateau. The rise in F_0 is supposed to be perceptually distinctive. F_0 is indicated in Figure 1.2 both by the pitch track – the black line – and by the contours of the harmonics, which are the dark bands. The rise in this example extends through the word *I* and peaks during the word *don't*.

Recent work on dialectal variation in English has used systems such as ToBI that combine pitch tracking and impressionistic transcription. Grabe (2004)

and Grabe *et al*. (2000) used a derivative of ToBI called Intonational Variation in English (IViE) to study intonational variation in Belfast and in cities across England. They focused on whether each dialect tends to compress or truncate tones when they are produced rapidly. In the United States, ToBI and related systems have been used for analysis of intonation in African American and Mexican American English (e.g. Jun and Foreman 1996; Goodwin, Goodwin and Yaeger-Dror 2002).

One well-studied example is the 'high rising terminal', which involves rising final boundary tones in intonational phrases where they would not be expected in most varieties of English. It has been studied in both purely impressionistic studies and in studies using ToBI (Guy *et al*. 1986; Britain 1992; Fletcher, Grabe and Warren 2005; Warren 2005), as well as a perception study eliciting attitudes (Guy and Vonwiller 1984). High rising terminals are most associated with Australian and New Zealand English but have also been found elsewhere.

Finally, voice quality is an area of pronunciation that can show dialectal and sociolectal variation but which is rarely examined by variationists. Speech pathologists dominate studies of voice quality. They have developed methods for impressionistic assessment of voice quality by trained evaluators. The evaluators rate factors such as breathiness and creakiness, jaw protrusion, pharyngeal constriction, and nasality. Such techniques were adopted successfully by Esling (1978) and Stuart-Smith (1999) to test sociolectal variations in Edinburgh and Glasgow, respectively. Most sociolinguistic studies of voice quality have examined it instrumentally, however (see below).

1.3.3 Acoustic methods

The appearance of Labov, Yaeger and Steiner (1972) marked the advent of acoustic phonetic methods in the study of dialectal variation. This study showed that spectrographic analysis, which had been new to phonetics only twenty-five years earlier, could be used to differentiate dialectal vowel variants. The method they used to estimate vowel formant values was to measure the peaks of harmonics in narrowband spectrograms where the harmonics coincided with vowel formants. At that time, the process was tedious and required printing out large numbers of spectrograms. Even so, the authors demonstrated that conversational speech was suitable for large-scale acoustic analysis. Moreover, they showed that acoustic analysis readily illuminates dialectal variation in vowel quality. They illustrated their results on plots with the first formant on the *y*-axis and the second formant on the *x*-axis; both axes were shown in reversed orientation in order to simulate conventional vowel diagrams (see, e.g. Labov *et al*. 1972; Labov 1994). The plots showed measurements of the nuclei of individual vowel tokens, usually enclosed inside ellipses. From these plots, the

authors described vowel shifting patterns and attempted to extract some general principles governing the shifting of vowels. In fact, vowel shifting and the principles behind it have become a major thrust in sociolinguistics. Because of their importance, §1.4.1 below covers them in more detail.

Since 1972, acoustic analysis techniques have advanced and Labov and others have taken advantage of the technological improvements. The most important innovation is linear predictive coding (LPC; Atal and Hanauer 1971). LPC is a computational method of estimating formant frequencies that superseded the laborious method of estimation from harmonics used by Labov *et al*. (1972). LPC became widely accessible within a few years, helped along by the advent of personal computers.

Many of the newer studies have departed from the presentation methods of Labov *et al*. (1972). Even Labov has abandoned enclosure of tokens inside ellipses in formant plots. While Labov still plots each token in his works, he and some other authors have found it useful to show mean values of the tokens for a particular phoneme or sound class. Mean values often make the relative positions of different phonemes clearer, which can be useful for demonstrations of vowel shifting. They also make diphthongal movements easier to plot coherently. Two examples, one illustrating a speaker from the same locality as the speaker in Figure 1.1 and the other a starkly differing dialect from Northern Ireland, are shown in Figure 1.3.[2]

In the plots, keywords connected with an equal sign indicate merged classes. The relative height and advancement of each vowel are indicated by its relative position in the plot. Another kind of display that is used infrequently is a trajectory display, in which a vowel is sampled at intervals and the formant values at each interval are plotted, with lines connecting successive measurements. Trajectory analyses are most useful for diphthongs. For specialised uses, factors besides the first two formants, such as the frequency of the third formant or the duration of the vowel, can be shown.

Acoustic analysis permits examination of some variables that cannot be analysed impressionistically. For example, the length of a segment can be gauged only in general terms impressionistically. However, spectrographic analysis allows researchers to measure it to the millisecond. Statistical comparisons can then tease out phonological length differences. Undershoot of segments – that is, when they do not reach their 'target' values – especially of vowels, can be examined by comparing formant values against durations (Lindblom 1963). Diphthong glides are quite difficult to transcribe impressionistically with any exactitude, but are relatively easy to measure with spectrograms and LPC. Such work shows that their phonological representations can be more specific than the traditional designation as /h/, /w/, or /y/=/j/. Other factors, such as phonation (degree of breathiness/creakiness), can be measured far more precisely with spectrograms. Sociolinguists

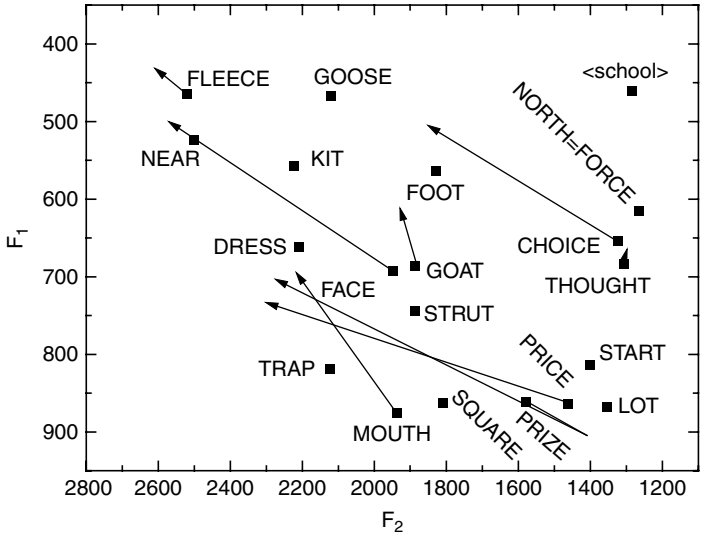


Figure 1.3a Formant plot of the mean values of the vowels of a European American female, born 1902, from Hyde County, North Carolina. Arrows indicate the gliding of diphthongs

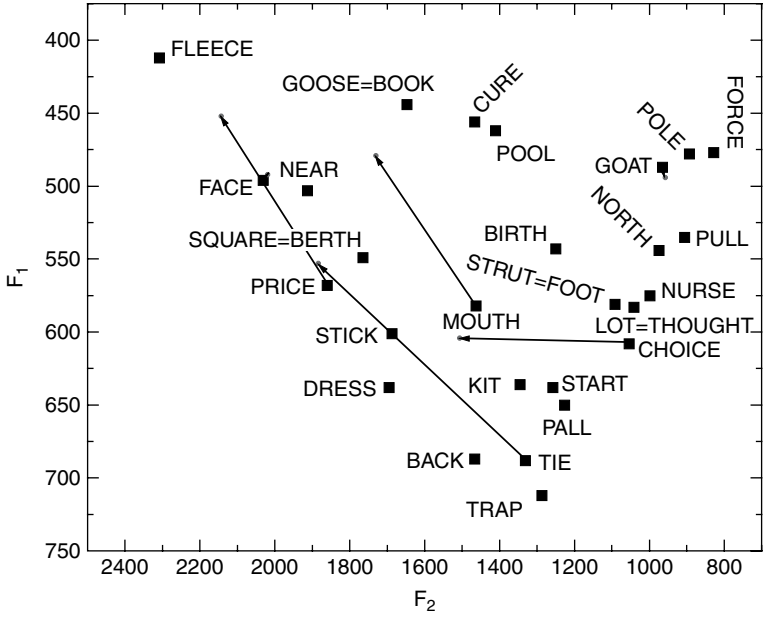


Figure 1.3b Formant plot of the mean values of the vowels of a male speaker, born 1943, from County Tyrone, Northern Ireland

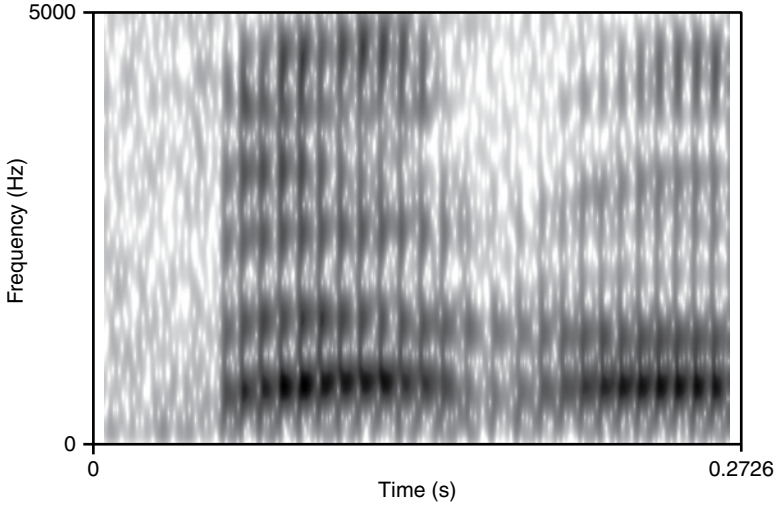


Figure 1.4 Spectrogram of the word 'better' spoken by a white male from Newcastle upon Tyne, England. The /t/ is realised as creaky voicing with no 'stop gap' (silent period for the stop occlusion)

typically regard any form of slow vocal fold vibration as 'creakiness', in large part because it is hard to hear further differences without special training. However, acoustic measures of jitter readily distinguish irregular vibration – more properly called 'roughness' or 'harshness' (Laver 1980) – from slow but regular vibration.

Acoustic analysis of non-vocalic variables, both consonantal and prosodic, has been scarce. Sociolinguists have generally assumed that impressionistic analysis is adequate for consonantal variables. In addition, methods of analysing consonants acoustically are poorly known in sociolinguistics. Nevertheless, a series of studies on variation in /t/ in two cities in northern England, Derby and Newcastle (Docherty *et al*. 1997; Docherty and Foulkes 1999; Foulkes and Docherty 2006) has demonstrated the usefulness of spectrographic analysis of consonantal variation. Glottalisation of /t/ occurred in several types and in complex interactions with social factors. For example, Figure 1.4 shows a spectrogram of a glottalised medial /t/ with the /t/ realised only as creakiness, without the blank period indicative of a stop occlusion. In Derby, final /t/ showed a number of distinct realisations that could pass for releases impressionistically. Existing phonological treatments are inadequate to account for the variety of forms or the variation that individual speakers show.

Another consonantal variable that has barely been examined in English is voice-onset time, widely known as VOT in phonetic circles. VOT is relevant for

syllable-onset stops and is calculated as the distance in milliseconds between the stop release and the onset of voicing. For a truly voiced stop, it is a negative number; for an aspirated stop, it is strongly positive; and for a voiceless unaspirated stop, it is close to zero. Syrdal (1996) found small differences in VOT for /p/ and /b/ across American dialects. Substrate influence sometimes influences VOT, as Heselwood and McChrystal (1999) noted for Panjabi-British speakers.

Substrate influence has been shown to influence syllable-coda consonants as well. Purnell, Salmons and Tepeli (2005) and Purnell, Salmons,Tepeli and Mercer (2005) examined phonetic cues for coda stops in Wisconsin communities with German-American backgrounds. They measured contours of the first formant and F_0, degree of vocal fold vibration, and length of the preceding vowel, all of which serve as cues to the phonological voicing of stops. All of these cues showed realisations unusual in American English, at least among older speakers, apparently because of German influence. The realisation of the preceding vowel can serve as a cue, too. In Thomas (2000), I found that the height of PRICE glides was one cue that Anglo speakers used to determine whether a following consonant was voiced or voiceless. However, Mexican American speakers produced a much smaller difference in the glides than Anglo speakers and used it as a cue to a much lesser extent, apparently due to their Spanish substrate. These studies found that the phonetic cues used for a phonological distinction can differ from those in mainstream varieties, demonstrating that phonology needs to account not just for contrastive features but also for the details of how distinctions are made.

The fact that phonetic cues can be internalised just as contrastiveness is upsets traditional notions of phonology, yet it is not the only non-contrastive factor that is internalised. Fourakis and Port's (1986) comparison of epenthetic stops in American English with their absence in South African English (e.g. in *dense*, American [dɛnts] vs. South African [dɛns]) showed that the relative timing of articulatory gestures could be internalised. That is, in *dense*, the occlusion gesture lasts longer than the voicing and velum-lowering gestures in American English, giving the impression of an epenthetic [t], but it does not last longer in South African English.

Acoustic analysis is also useful for studying prosody. Intonation was covered above, but another aspect of prosody is timing. Students of language variation have barely begun to examine the timing of speech. What study there has been has revolved around the issue of stress-timing vs. syllable-timing, or prosodic rhythm. In stress-timing, intervals between stressed syllables are supposedly relatively constant, while with syllable-timing, intervals between syllables are relatively constant. The determination that prosodic rhythm exists on a continuum and not as a binary feature paved the way for the recent development of methods to quantify it. Two methods were developed to distinguish

relatively syllable-timed Singapore English, with its mostly Chinese substrate, from stress-timed British English. Low, Grabe and Nolan (2000) compared durations of adjacent vowels, while Deterding (2001) compared durations of adjacent syllables as a whole. Both methods successfully differentiated syllable-timing from stress-timing. A few published studies (Gut 2002; Udofot 2003; Carter 2005; Thomas and Carter 2006) have used these methods for other forms of English, mostly varieties with substrate influence.

Phonological theory seldom addresses voice quality except as a secondary feature of segments, but variationist studies show that some aspects of it are internalised. Three such aspects that have been examined acoustically are breathiness, nasality, and overall F_0. Breathiness is measured either by comparing the lower harmonics, because breathy voicing shows greater spectral decay than modal or creaky voicing, or by measuring high-frequency noise (Hillenbrand, Cleveland and Erickson 1994), which is greater in breathy voicing. Di Paolo and Faber (1990) and Henton and Bladon (1985), for example, have used breathiness in variation studies. Nasality can characterise individual speakers, particular vowels, or particular vowels in certain dialects. It is rather difficult to measure but can be quantified by comparing amplitudes of oral and nasal formants or by fitting speakers with a device that measures nasal airflow. Plichta (2006) used both methods to show that Michigan English used vowel nasality as a secondary feature of the TRAP vowel, which most other varieties do not. As for overall F_0, it is gauged fairly easily by taking mean or median values from pitch tracks. Hudson and Holbrook (1981) and Walton and Orlikoff (1994) used such methods in reporting that low overall F_0 characterised African American English.

1.3.4 Data analysis

The practice of using broad phonetic transcription that facilitates coding of data into two or a few discrete categories has influenced the way sociolinguists analyse data. Statistical tests that are appropriate for discrete data, particularly logistic regression, are necessary. For example, the oft-analysed consonant cluster reduction is normally treated as a binary variable, the affected consonant being either present or absent (see, e.g. Wolfram 1969; Guy 1980; Bayley 1994; Santa Ana 1996). The presence or absence of the deletable consonant becomes the dependent variable. Independent variables consist of linguistic factors, social factors, and sometimes stylistic factors. For consonant cluster reduction, linguistic factors might be whether the next word begins with a consonant or vowel and whether the deleted consonant is part of the word root – as with *past* – or a verbal suffix, as in *passed*. Social factors might include the social class, ethnicity, sex, and age group of the speaker. A widely used statistical package, called *Varbrul* (or, in its newer version, *Goldvarb*) was developed

by Henrietta Cedergren and David Sankoff (Cedergren and Sankoff 1974) for such analysis. This package analyses data by subjecting them to logistic regression followed by a post hoc test. The results give factor weights for independent variables. Factor weights range from 0 to 1, with a weight above 0.5 favouring the variant being analysed and a weight below 0.5 disfavouring it.

Analysis of linguistic variables as discrete entities, while well-suited to morphosyntactic variables, can obscure details of phonetic and phonological variables, however, and sociolinguists have perhaps been too reliant on discrete analysis. It has constituted a powerful tool for examining social correlations with language. However, the strict adherence to binarity has hindered examination of many variables that are internalised, such as the 'low-level' processes discussed in Thomas (2001). For example, phonological voicing of a consonant can be signalled by different suites of cues, including any combination of vocal fold vibration, F_0 and F_1 contours, aspiration or glottalisation, length of the preceding vowel, and length of the consonant itself. Focusing on binarity thereby limits ways that variationists can address phonology, though this situation is now changing. In certain cases, such as when the dependent variable is a percentage or proportion of the total responses, the variable becomes continuous. In that case, statistical analyses appropriate for continuous variables, such as t-tests, ANOVA, linear regression, and various more complex procedures have been used. Mixed models that are suitable when independent variables are discrete and dependent linguistic variables are continuous are being explored. Now that acoustic analysis has become commonplace, tests appropriate for continuous data have become essential because acoustic data are inherently continuous.

Multivariate analyses have been used occasionally. Various types are suitable for either discrete data, continuous data, or both. Multivariate analyses allow researchers to determine how different linguistic variables are correlated with each other and with demographic/social features of speakers. One type of multivariate analysis, principal component analysis, became well known among sociolinguists through Horvath's (1985) analysis of linguistic variation in Sydney, Australia. Horvath showed that speakers could be divided into two distinct groups based on their linguistic behaviour. Stuart-Smith, Timmins and Tweedie (2007) employed both principal component analysis and another multivariate analysis, cluster analysis, in a study of consonantal variation in Glasgow English. Cluster analysis produces a dendrogram that links each individual to other individuals according to how similar their speech is.

1.4 Issues of special concern

Sociolinguists and, in some cases, historical and experimental linguists have been especially concerned with certain theoretical problems. Three merit

special attention here. First, historical linguists and quantitative sociolinguists, led by William Labov, have focused a great deal of attention on developing theories of how sounds shift. Second, the mechanisms of mergers have also attracted considerable attention. Third, a smaller amount of attention has been directed at how speakers differentiate vowels in terms of acoustic boundaries. Each of these problems requires its own methods of data collection.

1.4.1 Vowel shifting theories

The copious research on vowel quality has led to theories about the motivations for vowel shifting, which was in fact what Labov *et al*. (1972) intended. Before Labov *et al*., there was discussion of the 'principle of least effort' and of push and pull chains and vowel dispersion (e.g. Saussure 1986: 147–8). The principle of least effort stated that speakers expend as little effort as possible to express themselves. This principle explains processes such as assimilation and deletion, especially assimilation-related changes such as conditioned sound changes, in which a sound shifts only in a particular phonetic context. Assimilatory shifts make one sound more like another sound when they occur adjacently. For example, in many varieties of English, vowels are backed when they occur before /l/, which itself is realised either as a velar [ɬ] or as a velar vowel such as [o]. However, the principle of least effort has little explanatory power for other kinds of shift. Discussions of vowel dispersion and push or pull chains (e.g. Martinet (1952) and Moulton (1962)), offered explanations for those other shift types. In a push chain, one sound encroaches on the space of another, inducing the other vowel to shift out of the way. In a pull chain, one sound shifts and leaves behind an unfilled space, and in turn another vowel moves into that space, probably because no perceptual confusion results. In practice, it is often difficult to distinguish push chains from pull chains.

Labov *et al*. (1972) and Labov (1991, 1994) examined vowel shifting patterns in a variety of languages and English dialects. Patterns of shifts that appeared repeatedly across languages were identified, and then principles to explain those patterns were formulated. Considerable emphasis was placed on *peripherality*. Peripherality refers to whether a vowel lies along the outer edge of the vowel envelope or more toward the interior, which is determined from acoustic analysis. The more important principles are that peripheral vowels rise or move to the front along the periphery, non-peripheral vowels fall, and vowels can change their peripherality under certain circumstances. These principles apply most readily to languages such as Germanic languages that have a tense/lax or long/short vowel contrast, which is usually – though not always – equivalent to peripheral/non-peripheral.

These principles provide a new way of viewing phonology. Labov and his colleagues treat peripherality as a phonological feature. Even though there

are reasons to question whether peripherality itself is the driving force behind chain shifts (Thomas 2003), Labov's team has demonstrated how acoustic analysis can provide insights into phonological structure and the nature of sound change.

Unfortunately, work on principles governing non-vocalic change has not progressed far. Consonantal changes can be seen in terms of push and pull chains and various weakening or strengthening processes, but no widely accepted set of underlying principles has emerged. Prosody is even more poorly understood. Uncertainty about phonological primitives and the lack of historical data hinder work on intonational change. Other aspects of prosody, including rhythm and word stress, have similar problems.

1.4.2 Mergers

A phonological merger occurs when two sounds that had formerly been contrastive become pronounced alike so that they no longer contrast. Mergers come in two types, conditioned and unconditioned. A conditioned merger is one that occurs only in some phonetic contexts. A well-known example in English is the merger of the DRESS and KIT vowels before nasals, as in *pin* and *pen*, which is common in more southerly parts of the United States. An unconditioned merger is one that occurs in all phonetic contexts. Perhaps the best-known unconditioned merger that is now spreading in English is that of the LOT and THOUGHT vowels, which is common in Canada, the United States, and Scotland. There has also been some discussion about how mergers spread: whether by transfer of words from one class to the other, by the two classes steadily approaching each other until their differences disappear, by the two classes becoming something that includes the acoustic space of both classes, or as substrate influence from a language that lacks a similar distinction (Labov 1994; Herold 1997).

As noted earlier, minimal pairs are one method used to test for phonological mergers. An example of a minimal pair for which speakers vary is *cot/caught*, which differs only in that the first word has the LOT vowel and the second the THOUGHT vowel. These pairs are used to test for phonological mergers: if a subject pronounces the words alike, he or she has the merger, and if the subject pronounces them differently, he or she maintains the distinction. However, the results should be used in conjunction with spontaneous and/or reading-passage speech because, as Labov (1994) discusses in detail, some speakers produce mergers or distinctions in minimal pairs that they do not produce in other styles. Minimal pairs induce speakers to focus their conscious knowledge of the language on the words, and sometimes their conscious knowledge does not reflect their ordinary speech. Nevertheless, minimal-pair tests work in most instances.

In addition to determining whether a subject pronounces words in a minimal pair the same, researchers can also ask the subject whether he or she thinks the words sound alike. This method was used extensively by Labov *et al*. (2006). They usually referred to subjects' responses as 'perception', though *speaker judgement* might have been a better name.

Nonetheless, experiments in what is more properly called speech perception have been used to test for mergers. An important early study was that of Janson and Schulman (1983). They tested whether the Swedish vowels short /ɛ/ and short /e/ could be distinguished by subjects from two dialects of Swedish, one of which maintained the distinction and the other of which merged the two vowels. Using a speech synthesiser, they created a continuum of vowel qualities in a /sVt/ frame and asked subjects to identify what word they heard. As it turned out, subjects from both dialects failed to distinguish /ɛ/ and /e/. Janson and Schulman concluded that some distinctions could be maintained in speech production but not be utilised in perception to distinguish words.

There was a flaw in Janson and Schulman's experimental design, however, and Labov, Karen and Miller (1991) exposed it. Janson and Schulman forced their subjects to tap into their conscious knowledge of Swedish. In ordinary conversational interactions, however, such conscious knowledge seldom operates. People rely on some deeper kind of phonological 'knowledge' to recognise segmental distinctions. A different kind of experiment was needed to test whether subjects could recognise variably distinguished sounds. Labov, Karen and Miller devised an experiment, dubbed the 'coach test', to examine whether words such as *merry* and *Murray* could be distinguished by Philadelphians. Subjects listened to a story about a coach in which the outcome of the story rested on whether a particular phrase included *Merion* or *Murray in*. They were then asked questions for which the answers depended on which phrase they had heard. Then they were asked to listen to the story again, but this time, unknown to them, they heard the opposite phrase. They were asked questions to determine whether they recognised what they heard the second time. The results showed that native Philadelphians were impaired in their ability to recognise the distinction, while non-Philadelphians were not.

Another type of perception experiment used to test for phonological mergers is called a *commutation test*. In a commutation test, subjects listen to words that they or another speaker of their dialect have uttered and are asked to identify the word. The assumption is that if they consistently identify such words correctly, they can recognise the distinction being examined, but if their answers are close to random, they cannot recognise it. For instance, if a subject can reliably distinguish local pronunciations of *pool* and *pull*, one can assume that the subject has internalised the phonological distinction between pre-/l/ allophones of the GOOSE and FOOT vowels as it pertains to his or her native dialect. Labov has used commutation tests in a number of studies (e.g. Labov

et al. (1991)), and other authors have employed them as well (e.g. Di Paolo and Faber 1990). One advantage of commutation tests is that, when subjects recognise a distinction that the researchers did not find in their speech, the test can demonstrate that subjects are sensitive to some cue besides those that the researcher expected the distinction to involve.

Rae and Warren (2002) adapted a different type of perception experiment to test the merger of the NEAR and SQUARE vowels in New Zealand English. They used minimal pairs, such as *fear* and *fair*, and played recordings of one of the members of a pair with another word, which was either semantically related to the NEAR/SQUARE word or unrelated to it, to form a couplet. They also included distractor couplets and couplets with non-words. Subjects pushed buttons to indicate whether the second member of the couplet was a real word or not and their response times were measured. Certain couplets without a semantic connection, such as *chair/shout*, showed slower response times than couplets such as *cheer/shout* with a semantic connection. However, quick responses for couplets such as *cheer/sit* indicated that the NEAR and SQUARE vowels were merged because *cheer/sit* would sound the same as *chair/sit*.

1.4.3 Boundaries between phonemes and goodness tests

Studies of the boundaries between phonemes are not usually conducted in speech production, partly because tokens – at least for neighbouring vowel phonemes – usually show some overlap and partly because tokens in different phonetic contexts may show different boundaries, complicating the picture considerably. In fact, overlap often figures in discussions of chain shifting, as noted earlier for push chains. In a push chain, the two sounds are assumed to overlap at some point. However, experiments on phoneme boundaries have occasionally been conducted in speech perception. Similar experiments in which subjects rate how well different stimuli match their conception of a particular phoneme have also been conducted, albeit rarely.

The best-known experiments on boundaries between phonemes have been conducted on Swedish (Janson 1983, 1986) and demonstrated that speakers of different regional dialects of Swedish and different birth cohorts could hear sounds differently. In English, a similar study was conducted earlier by Willis (1972). Willis played a series of synthetic vowels representing a continuum to speakers from Buffalo, New York, and a neighbouring community in Canada. Speakers from the two locales differed strongly in what vowels they identified each stimulus as and where the boundaries occurred. It turns out that the results Willis found match later findings about how vowels of the dialects of Buffalo and Ontario differ in production. Results are not always so closely matched with production, however. Niedzielski (1999) found that subjects in Michigan identified vowels from other regions with their own speech, not vowel variants

that they themselves actually produced. Unlike Janson and Willis, she played different recorded variants to subjects instead of synthesised stimuli on an acoustic continuum.

Another approach to the boundary issue has been cross-dialectal identification, in which speakers of one dialect are asked to identify stimuli uttered by a speaker of a different dialect. A number of studies have taken this approach, such as Flanigan and Norris (2000) and Labov and Ash (1997), both comparing dialects of American English, and Trail, Ball and Müller (1995), who tested how listeners from England identified South African vowels. Listeners experience difficulty with some vowels from unfamiliar dialects, but, surprisingly, they sometimes have trouble recognising vowels in their own dialect.

A different approach to assessing subjects' identifications of variants found in their own speech is through goodness ratings. In this sort of experiment, subjects are asked to rate on a scale how closely stimuli match their own pronunciation. Peeters (1991) conducted such an experiment on certain diphthongs or long vowels with speakers of British English, Dutch, and German and the results matched production norms in each of the languages. Sociolinguists have not generally adopted this method, however.

1.5 Where next?

Key readings are Docherty *et al*. (1997), Foulkes and Docherty (1999), Gilles and Peters (2004), Kurath and McDavid (1961), Labov (1966, 1994), Orton, Sanderson and Widdowson (1978), Pederson *et al*. (1986–92), Thomas (2002), and Wolfram (1969).

More specifically, for examples of how traditional dialectology handles phonological variables, one cannot do better than Kurath and McDavid (1961) for the east coast states of the USA or Orton *et al*. (1978) and Anderson (1987) for England. Modern ways of examining geographical variation in phonology are exemplified notably by LAGS (Pederson *et al*. 1986–92) and the ANAE (Labov *et al*. 2006). The recent *Varieties of English* series (Burridge and Kortmann 2008; Kortmann and Upton 2008; Mesthrie 2008; Schneider 2008) provides an exhaustive survey of phonological and phonetic variation across the entire English-speaking world.

Early sociolinguistic approaches to variation in pronunciation and determining how it is correlated with social factors and speaking style are well represented by Labov (1966), Wolfram (1969), and Trudgill (1974). Key innovations in the treatment of social variables appear in Milroy and Milroy (1985), Horvath (1985), and Eckert (1988). Labov (1994, 2001, and a third volume in preparation) discusses his views of sound change thoroughly. The various papers in Foulkes and Docherty (1999) comprise a number of forward-looking approaches to phonological variation.

All of the preceding sources focus on segmental variation, though some papers in the *Varieties of English* series and in Foulkes and Docherty (1999) address other variables. Gilles and Peters (2004) is a fine collection of papers illustrating approaches to intonational variation in various languages, including English. Thomas (2002) reviews perceptual approaches to variation.

The scarcity of papers on topics beyond the production of segments indicates where future research is needed. Prosody and voice quality are in particularly sore need of work. Perception is receiving some attention but not as much as it warrants. Within segmental production, consonants have not attracted much acoustic work, as Docherty and Foulkes (1999) point out. Theories on shifting of consonants and intonation remain to be formulated. In the larger context of phonology, variation should be used more extensively to address issues such as how detailed phonological specifications are. As noted earlier, dialectal variation suggests that many 'low-level phonetic' features, such as the particular cues used for making contrasts or the relative timing of articulatory gestures, are actually internalised. Furthermore, how much knowledge of stylistic and register variation can be considered part of a speaker's phonology? To what degree is a speaker's acumen about the way other speakers talk represented phonologically, especially when the speaker undergoes accommodation to other speakers? Docherty *et al*. (1997) contend that it is important to examine the behaviour of many speakers in order to find full answers to such questions and attain a complete view of phonology. Scholars who study variation in language should take the collection of phonological data as a key part of their mission.

2 How to make intuitions succeed: testing methods for analysing syntactic microvariation

Isabelle Buchstaller and Karen Corrigan

2.1 Introduction

Dialects of the same language are known to vary systematically with respect to the proportional frequency with which different syntactic constructions are used productively. However, many syntactic variables[1] are relatively rare in spoken interactions of the kind elicited by sociolinguistic interviews, particularly when the variants in question are stigmatised within the community in which they are present but also because of the open-endedness of the syntactic component. Relic features like 'for-to' complementisers, for example, are not only restricted to older generations of speakers, but their frequency in interviews even within this social group is also delimited by the fact that complementiser constructions are only one amongst numerous structural possibilities for conveying grammatical and pragmatic meaning. Due to the low token frequency of such variants, the investigation of large-scale dialectal variation within the syntactic component has increasingly come to rely on the collection and analysis of introspective judgements. This reliance has initiated an important discussion about the linguistic status and empirical appropriateness of judgement data within the field of dialectology. A number of scholars (Schütze 1996; Cowart 1997; Cornips and Poletto 2005; 2008) have drawn our attention to the fact that, if proper care is taken to control for potentially interfering, though independent, linguistic constraints – for example, lexical frequency/familiarity, pragmatic plausibility and sentence length – as well as for extra-grammatical factors – such as the social profile of the speakers, fatigue, memory limitation and ordering effects – native speakers can indeed be found to produce systematic patterns of acceptability ratings when using these methods.

A version of this paper was presented at *NWAV35* and at *Sociolinguistics Symposium 17*. We would like to thank audience members for their constructive comments, from which this paper has greatly benefited. We are also grateful to our fieldworkers, Tejshree Auckle, Laura Bailey, Jonathan Burrows, Sophie Robinson and Dominic Thompson as well as to Newcastle University's Faculty Research Fund and Vacation Scholarship Schemes, for the grants which made the pilot projects reported on here possible. We would like to acknowledge the assistance of Anders Holmberg, Leonie Cornips and Cecilia Poletto in the design of our questionnaires. Thanks also go to David Adger, Frans Gregersen and Bill Kretzschmar for their comments on earlier drafts.

Yet there are still 'continuing doubts about the empirical reliability and theoretical interpretation of judgment data' (Cowart 1997: 2).

More recently, a relatively new sub-discipline of linguistics that has come to be known as 'socio-syntax', whose orientation is generative as well as variationist, has started to problematise data collection techniques that rely on more traditional methods of introspective elicitation (see the collections by Cornips and Corrigan 2005 and Trousdale and Adger 2007 for exemplification). The hallmark of this new paradigm has been the adoption of an approach described as 'layered' by Benincà and Poletto (2007) because it combines different data-collection methodologies. The overall aim is to mitigate the impact of social intervention and task effects that are not related to the syntactic variation being investigated and to produce appropriate data-sets for cross-linguistic and cross-dialectal comparisons. Given that the socio-syntactic approach is relatively new, there is no absolute consensus as to what constitutes the most consistent and objective methods for collecting intuitions which are stable and comparable.

What is needed, therefore, is the establishment of 'best practices' for measuring grammatical acceptability. In this chapter, we aim to introduce some of the methods that are commonly used in research on morphosyntactic variation. As a second step, we will put these instruments to the test and report on some findings from a recent study which set out to investigate the extent to which these methods produce reliable, consistent and therefore comparable results. By doing so, we hope to demonstrate that tapping into native-speaker intuitions regarding variability in English, while providing a rich and varied source of evidence that can complement low token numbers from interviews, is not always straightforward and, in certain respects, is partially determined by related issues raised elsewhere in the volume (see the contributions by Guy, Montgomery and Beal and Trousdale, in particular).

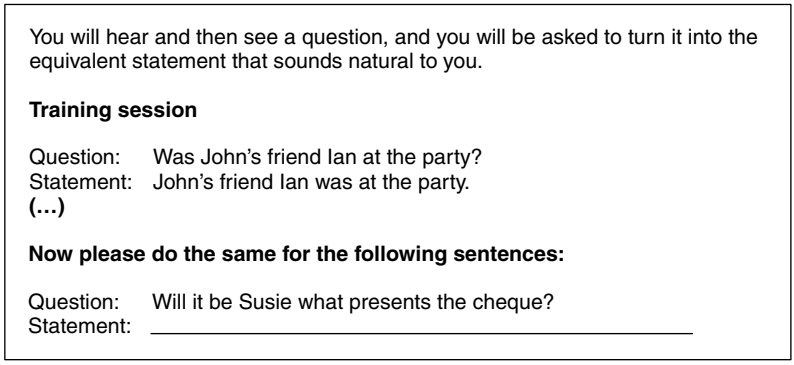
2.2 Morphosyntactic variation: a review of methods

The measurement of linguistic acceptability ratings should aim at maintaining the standards for empirical research set throughout the sciences regarding reliability and replicability (Cowart 1997). Sociolinguists and dialectologists have indeed developed sophisticated methodologies for tracking linguistic diversity in English varieties with respect to the phonological and lexical levels (see Thomas this volume; Kerswill *et al*. 1999; Britain 2002 and Milroy and Gordon 2003 for the British Isles, as well as Kretzschmar *et al*. 1993 and Labov *et al*. 2006 for dialect regions in North America). However, 'best practices' in accessing English vernacular morphosyntactic data that is naturalistic have not yet been fully determined. To date, most research on variation and change in English has focused on highly local (usually urban) communities

at the expense of an integrated, comparative account of dialect morphosyntax across wide tracts of geographical space. As Kortmann (2002) observes, large-scale analyses of syntactic variability within the English-speaking world are, notably, absent. With the exception of the nascent comparative research programme recently undertaken by Beal and Corrigan (2005) and Tagliamonte (2002) and (2008), we are not really in a good position to make global claims about the geospatial or social patterning of morphosyntactic variation more widely. The lack of uniform methodologies means that we do not yet have the kinds of data 'that would allow us to investigate differences in the syntax of Newfoundland and Vancouver Englishes, or of Cornish and Tyneside dialects' (Bauer 2002: 107–8). This chapter introduces the kinds of elicitation test which could finally make this possible.

The first of these is unique as far as tapping into native-speaker intuitions is concerned since it aims to collect data that is linked to both production and introspection. Given the interest within certain linguistic frameworks, such as those of the generative tradition, in accessing grammaticality judgements via questionnaires, many researchers assume that any data that is culled from questionnaires is perceptual per se. However, this is not always so, as we will demonstrate in our discussion below of a judgement task called 'reformulation'.[2] It has been used extensively in dialect atlas projects in the Netherlands (SAND, www.meertens.knaw.nl/projecten/sand/sandeng.html) and in Scandinavia (SCANDIASYN, http://uit.no/scandiasyn?Language=en). In its original form, reformulation encourages informants to 'translate' or 'reformulate' a sentence from the standard variety into their local dialect. However, this method as usually employed might not be the most applicable in a British context where normative ideologies abound, since it presupposes a situation of relatively low prescriptive pressure in which informants are comfortable providing the dialectal equivalent of the 'standard'. In our research in Northern England, therefore, we adapted this task with the aim of circumventing any prescriptive judgements that might interfere with an informant's genuine response. We did this by giving the informants a sentence already in the vernacular and asking them to perform a syntactic transformation. In our case, this entailed 'translating' an interrogative sentence containing a dialect feature into a declarative one to determine whether they persisted with the vernacular variant in the new structure or introduced a standard variant during 'translation' (see Figure 2.1).

One of the great advantages of the reformulation task is that it allows researchers to systematically collect production data concerning syntactic constructions that might be difficult to obtain in more 'natural' speech events, for example, because they are very rare. Hence, results from this test can be used to complement other production data, such as the classic sociolinguistic interview, in establishing whether the informants have productive use of the vernacular feature in question.


You will hear and then see a question, and you will be asked to turn it into the equivalent statement that sounds natural to you.

Training session

Question: Was John's friend Ian at the party?
Statement: John's friend Ian was at the party.
(…)

Now please do the same for the following sentences:

Question: Will it be Susie what presents the cheque?
Statement: _____

Figure 2.1 An example of the reformulation task investigating relative-clause marking preferences

More traditional introspective research tasks seeking to uncover the socio-geographical patterning of syntactic variation have generally been either pseudo-quantitative or qualitative (Beal 2004 and Hughes *et al*. 2005, for instance). Thus, informants are asked to judge between either binary 'grammatical'/'ungrammatical' options (often designated '√'/'*' in the most traditional types of research) or between a wider range of options (often symbolised '?'/'??'/'?*'/'*' as Schütze (1996: 45) notes). Aside from the ambiguity of the use of symbols such as these across different studies (see Schütze 1999, inter alia), there is also the issue of finding a principled method for mathematically measuring degrees of acceptability of this kind. A controlled and systematic process of data collection is obviously important in order to produce a comparative sample of judgements and indeed any truly graded mathematical conception of grammaticality. Hence, in the so-called direct grammaticality judgement task, rather than using '√'/'*', many researchers in the field of socio-syntax ask their informants to decide whether or not they personally would or would not use a particular construction by giving a *yes/no* response (see Labov 1972b: 21, 1996: 78,100). This has the added advantage of making such judgements psychologically real for the informants rather than having them deal with abstract grammatical notions of acceptability or grammaticality. Also, since inter-informant variability depends largely on the ability of individuals to devise a pragmatic context in which a sentence could be acceptable, a good method to reduce variance across results is to embed the stimulus sentence into a short text (Schütze 1996: 151). Providing contextualisation precludes informants having to envisage a suitable discourse context for themselves, which they may find particularly challenging when faced with sentences of only marginal acceptability. In our version of these tests, therefore, we presented each sentence containing the variant for testing at the end of

Please listen to and then read the following sentences. For each question, one sentence will be in bold font. Please indicate whether or not you personally would use that sort of sentence by circling Y or N.

Some friends were having coffee. One complimented her friend on the cake she had made. '**It was Jackie as gave me the recipe**' she admitted.

Would you use this sentence? Y / N

Figure 2.2 An example of the direct grammaticality judgement task investigating relative strategies

a short contextualising paragraph. In order to ensure that the informants were aware of which section of text they should be rating, the 'test' sentence was marked in bold font as in Figure 2.2.

The direct grammaticality judgement test produces 'nominal' output, that is, non-numerical/qualitative responses such as 'sentence A is the same or different from sentence B' which can therefore be counted and reported as frequency percentages.

We will now present a range of other techniques commonly used in dialect syntax projects which (with the exception of the pictorial elicitation task) produce output that is at least 'ordinal' in nature, that is, values for sentences A and B can be ranked as 'more' or 'less', which means they can be ordered (first preference, second preference and so on), hence allowing researchers to perform more powerful mathematical procedures such as the calculation of averages, standard deviations or medians.

In contrast with the direct grammaticality judgement test, in the indirect grammaticality judgement task informants do not have to declare whether or not they personally use a certain variant. Instead, they are asked if they recognise vernacular forms used by other people in their locale. The task asks informants to rate individual sentences by assigning them a number which is associated with a corresponding verbal descriptor (see Labov 1975, 1996). The scale of judgements we have used in our research can be seen in Figure 2.3 below. Although concerns have been raised about the reliability of results produced by scaling of this kind, the task has one main advantage, namely, that it is simple for the informant to understand and that it produces results which are readily quantifiable (Cowart 1997: 72).

There are some key differences between the direct and the indirect grammaticality judgement tasks: by its nature, the indirect grammaticality judgement task exerts considerably less prescriptive pressure on the informant. They are not losing face if they say that people in their area are using these features while simultaneously not claiming to use them themselves. Such judgements can be instructive with respect to tapping into implicit language attitudes as well as interesting when triangulated with informants' own performance data

Please rate the following sentences by circling one option on the following scale:

1 This type of sentence would never be used here – it seems very odd.
2 This type of sentence is not very common here but it doesn't seem too odd.
3 I have heard this type of sentence locally but it's not that common.
4 People around here use this type of sentence a lot.

For example:

If you heard the sentence below but thought that it is not very common in your area, you would circle 2 as we have done below.

Who do you think **that came** to see George yesterday?

1----------(2)----------3----------4

When you judge these sentences, please pay particular attention to the words in bold. Now please do the same for the following sentences:

1. Beth was complaining that her grandchildren were always so busy. '**They divven't visit me any more**' she said.

1----------2----------3----------4

Figure 2.3 An example of the indirect grammaticality judgement task investigating negation

and their responses from direct grammaticality judgement tasks. Thus, culling two types of judgement about identical features from the same informant not only allows the investigation of the degree to which a particular feature is productive in a community but it also permits the researcher to gain insight into the extent to which it may be stigmatised.

A task that sits squarely between these types of judgement test is the pictorial elicitation task used with considerable success in the SAND project, as noted in Cornips and Jongenburger (2001). Pictorial elicitation presents informants with an image alongside a short sentence containing a feature for testing which relates in some respect to the event depicted in the image. The informants are asked to: (i) examine the picture and the corresponding sentence and then (ii) give *yes/no* responses to the question 'Would you (or any local person) use this kind of sentence?'

Note that the manner in which the acceptability question in Figure 2.4 is worded means that the informants did not have to admit to using the vernacular construction themselves. As discussed above, this is useful in that it alleviates prescriptive pressures. Furthermore, by formulating the task in this way, we hoped to get positive responses for features that might no longer be used

Sam worked in the park because he couldn't get peace nowhere else.

Would you (or any local person) use this kind of sentence? Yes/No

Figure 2.4 An example of the pictorial elicitation task investigating judgements on multiple negation

productively by the informant but could still be present in the community dialect.

The final test discussed here is the so-called 'magnitude estimation' task, which is a method commonly employed in psychological and psycholinguistic experiments. Linguists have applied this test to cases of dialectological/syntactic variation, where the stimuli to be rated are sentences thought to differ in their grammatical acceptability (Bard *et al*. 1996; Cowart 1997; Featherston 2005). An important first step in the application of this method to issues of grammaticality is that informants are provided with a reference stimulus to which they assign any positive integer value of their choice. This stimulus tends to be a sentence that is suboptimal but not entirely ungrammatical (see the sentence in bold in Figure 2.5 below, as well as Schütze 1996 and Cowart 1997). Informants are then asked to compare other sentences to this reference stimulus. They are encouraged to give a higher rating to sentences which they deem to be 'better' in terms of grammaticality by comparison to the reference and a lower rating to those which appear 'worse'. This experimental format seeks to ensure that the informants rate the test sentence(s) (1–3 in Figure 2.5) in proportion to how (un-)acceptable they find the reference stimulus.

The informant who completed the questionnaire in Figure 2.5 rated the reference sentence as a '10'. They then allocated a much better score, namely '18', to the first test sentence, thereby rating it as considerably more acceptable than the reference sentence. Sentences 2–3, which are fillers, generated ratings of '8' and '5', and are thus being judged as less acceptable than the reference sentence.[3] The stimulus sentence 3, which contains an instance of multiple negation, was thus judged to be considerably less acceptable than the reference stimulus and much worse than test sentence 1, for instance. An important advantage of magnitude estimation as a method, therefore, is that it allows informants to make distinctions about the acceptability of sentences which are as subtle as they perceive them to be. It also gives them the freedom of choice to create their own individual scale to which they can confidently relate. Furthermore, sentence ratings can be readily compared since they pattern along a linear interval scale. In fact, since magnitude estimation tests yield results that are expressed as

Reference sentence:	**Your rating:**
I'm going home and got an umbrella.	10
Now, please rate all sentences below in relation to the sentence above:	
Sentence	**Your rating**
1 The man put his coat on the hanger.	18
2 That's what I hate, is that she's always late.	8
3 I'm not going to eat nothing hot no more.	5

Figure 2.5 An example of an informant's response to a magnitude estimation task investigating a sentence containing vernacular negation (sentence 3)

'interval' data (i.e. equal intervals on the scale of 'more or less' and thus scalar in nature), they are amenable to powerful parametric statistical tests.

However, in spite of the lengthy practice session which our fieldworkers conducted before the actual test was administered, the magnitude estimation task proved too complex for some informants with low numeracy skills (see Buchstaller and Corrigan 2008 and Buchstaller *et al*. forthcoming for discussions of the problems encountered with informants who failed to master the test for this reason). These generally tended to be the older, less educated, working-class speakers in the investigation, who paradoxically may well have the greatest tolerance of traditional morphosyntactic features in their dialects (i.e. NORMs – non-mobile older rural males). As such, we felt it to be crucial to adapt the magnitude estimation task so as to make it more user-friendly to such speakers. This was achieved by using an instrument that is based on the principle of 'graphic' rating (Guilford 1954: 270; Taylor and Parker 1964), and hence converting the rating from numerical to visual in nature.

Visual versions of the magnitude estimation test basically ask informants to express judgements by either drawing a line on a scale or marking a cross on a line between two opposing poles. In our project, we opted for drawing a cross on a blank line. Our informants were asked to mark preferred ratings further to the right than those that were dispreferred. As in the classic version of the test, they were first given a reference stimulus, which they rated by marking a cross on the blank line provided. Then they were offered a list of test and filler sentences, which they were to rate in relation to the stimulus sentence. Our adaption of the classic magnitude estimation task is shown in Figure 2.6.

Thus, as in the classic version, the ratings for the test/filler sentences are given relative to that for the reference stimulus – albeit graphically rather than numerically – which therefore functions as a conceptual anchor. Importantly, informants are again given free rein as to their rating of the anchor as well as

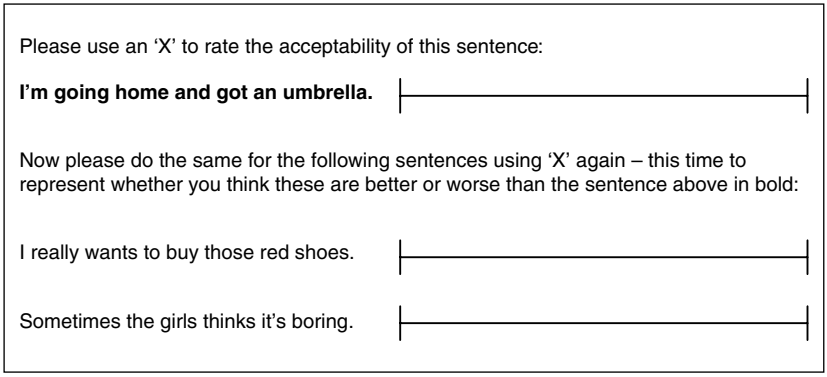


Figure 2.6 An example of a question in the visual version of the magnitude estimation task

the test/filler sentences. In order to convert these graphic markings into numerical form, the length of the line to the point where informants marked a cross was measured (in mm). Although the line-drawing test makes the subsequent quantificational analysis of data rather more time consuming than when using a numerical scoring system, this method circumvents the numeracy problems discussed above. As such, we consider it to be more suited to a wider range of informants (including children as well as older subjects like those targeted in our present investigation (see also Cowart 1997: 73)).

Having described a range of methods commonly used in studies investigating dialect morphosyntax, we will now move on to discuss some general considerations that are important when administering tools for measuring introspective judgements.

2.3 Grammaticality judgement tasks: further considerations

Human performance of any kind is typically patterned as a random scatter of individual observations which cluster around a more or less stable mean (Taylor and Parker 1964; Pashler 2002). This error variance is usually taken account of by offering several tokens of the same type for judgement, thus improving reliability via averaging and checking for outliers. Only if the variance between tokens (within the sentence type) is significantly smaller than between types can we then assume that something of note is occurring regarding speaker judgements. As such, Cowart (1997) suggests four items per feature so as to investigate the variability between instances of the same sentence type. However, depending on the number of features to be tested, coupled with the need to diminish fatigue effects, it is important to strike a balance between the number of exemplifications of a single phenomenon and the amount of time informants can justifiably be asked to spend completing questionnaires. Thus,

in our study, which tested a range of constructions, we restricted ourselves to investigating responses to two test sentences per phenomenon as we were conscious that our informants were already being required to devote over two hours of their time to completing our tasks.[4]

When administering these instruments, researchers also need to bear in mind that 'other factors, such as lexical frequency, (approximate) word length, grammatical complexity (argument structure, subordination patterns, number of adverbials etc.), might make a sentence hard to parse' (Schütze 1996: 164). A good way to identify potential impediments to speaker judgements in advance is to carry out pilot tests. Conducting such trials gives the researcher the chance to identify any orthogonal factors that might impact on informants' judgements and to rectify the questionnaire design appropriately.

The final version of the questionnaire thus piloted will typically consist of pairs of test sentences matched as closely as possible in terms of linguistic structure (as advocated by Cowart 1997: 46). For example, we used the following well-matched sentences in a direct grammaticality judgement task to elicit acceptability ratings for vernacular variants of the second person plural pronoun in subject position:

(1) Yous could share a large pepperoni pizza.
(2) Yous would make a really good team.

Another important consideration in questionnaire design is how to minimise the effect of extralinguistic factors which may influence an informant's ratings. Provided their social characteristics are rigorously controlled for, such factors are known to include the acceptability of preceding sentences or indeed fatigue – particularly as the informant approaches the end of the questionnaire (Schütze 1996: 155; Cowart 1997: 94). There are two strategies which tend to be used to avoid such effects: one is the use of 'filler' sentences, which are arbitrarily interspersed throughout the questionnaires.[5] Cowart (1997: 92) suggests that there should minimally be twice as many fillers as test sentences and ideally three or four times as many. However, depending on the research orientation, this might create an unmanageably long questionnaire leading to boredom, frustration and fatigue. A more manageable strategy might therefore be to follow Schütze's (1996: 193) recommendation of using 'enough' fillers and randomising so that informants are unable to remember the rating they gave to previous similar sentences. In the field of socio-syntax, the questionnaires tend thus to be 'scrambled' (Benincà and Poletto 2007: 51), which effectively means that researchers produce different versions of the same questionnaire with each version containing a different randomisation of test and filler sentences. This means that responses are less likely to be influenced by the order in which test and filler sentences are presented.

The discussion in the previous paragraphs has shown that investigations into morphosyntactic variation can make use of quite diverse methods for accessing introspective judgements. The field has also matured enough to have generated

a general awareness about potential orthogonal factors that might skew such data, as well as mechanisms to circumvent these effects. However, the question remains as to whether we can match findings gleaned from different methods when so many potentially incompatible testing instruments are currently being used. To what extent can we compare divergent data sources collected via rather different elicitation techniques, when we already know that rather diverse results have been produced by even small-scale studies in the same region when different methodologies have been invoked (McDonald and Beal 1987: 45–56)? In other words, can we reliably compare data-sets with respect to multiple negation based on 'indirect grammaticality' judgements from locality A with findings from 'picture elicitation' tasks for the same feature in locality B? It is these questions that the next section will address.

2.4 Testing comparability between instruments: a pilot study

We will now discuss some findings from a 2007 pilot study in the metropolitan county of Tyne and Wear in the north-east of England. The locations we tested are illustrated in Figure 2.7, namely, Newcastle (henceforth, 'NCL'), Gateshead (henceforth, 'GH') and Sunderland (henceforth, 'SL').

Grammaticality judgements were collected using the five different types of testing instrument illustrated above.

2.4.1 The linguistic phenomena

The specific linguistic phenomena we investigated are detailed in (I–IV) below. They were chosen by virtue of their being traditionally associated with Northern Englishes generally or with North-eastern Englishes more specifically, as described in Beal (1993, 2004), inter alia.

(I) Non-standard negation Multiple negation as well as the presence or absence of the vernacular negator, *divven't* or *dinnit* were both tested.

(3) You know I divven't like mayonnaise.
(4) I don't want to go nowhere else.

(II) Pronominalisation Also of interest was the non-standard second person plural pronoun, often spelled *yous*, which has been identified by Beal (1993: 205) as a feature of Tyneside English:

(5) Yous make a really good couple.

(III) Relativisation strategies This investigation focused on variability in the relative clause markers used in subject, animate, restrictive relatives, such as (6–8) below.[6] The vernacular variants examined were *as* (6), *what* (7) and *zero* (8).

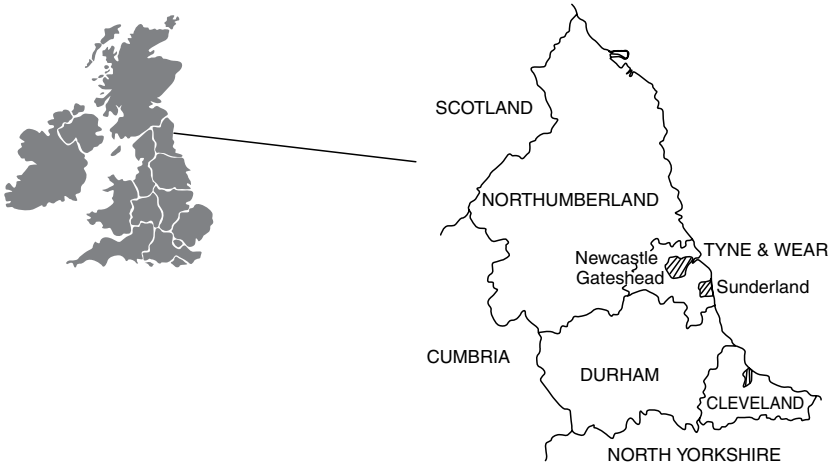


Figure 2.7 Map of the fieldwork locations

(6) It'll be the nurse as sees you next time.
(7) You can be the one what chooses the film.
(8) That's the man Ø lent me some money.

(IV) The Northern Subject Rule The verbal paradigm of many traditional Northern dialects (and others influenced by them) is constrained by the so-called 'Northern Subject Rule' (NSR) (see Murray 1873; Pietsch 2005a,b). According to this 'rule', subject noun phrases in clauses also containing present tense verbs attract an *-s* suffix on the latter even when these are not third person singular in function. The NSR was examined in three different environments, namely, with a noun phrase subject (as in 9), with a pronominal subject separated from the verb by an adverbial or modifier (as in 10), and with conjoined nouns forming the subject (as in 11).

(9) The children in the nativity play talks very clearly.
(10) I really likes to run by myself, but not when it's dark.
(11) Reality TV and sitcoms makes me laugh.

These four domains of typically Northern English dialect grammar acted as stimuli for the battery of acceptability judgement tasks being investigated.

2.4.2 The data collection process

In all, twelve speakers in the over-sixty-five age bracket participated in the study (two males and two females from NCL, GH and SL, respectively). All the participants were targeted via the 'friend-of-a-friend' approach (Milroy

and Gordon 2003) and each of them would be categorised as 'working class' according to the scheme identified by the year 2001 National Statistics Socio-economic Classification of Analytic Classes (see www.ons.gov.uk/about-statistics/classfications/current/ns-sec/index.html). The informants were also selected on the basis that they maintain dense social networks in their local community (Milroy and Gordon 2003), were born and raised there, and have lived in their community at least until the age of eighteen and nowhere else for more than seven years.

Completing the five types of task took, on average, forty-five minutes. For every task the informants were required to complete, they listened to a recording of the sentence they were to rate. These were uttered by a male, working-class native of SL, GH and NCL respectively. They then re-read the sentence themselves and, finally, recorded their judgement on the questionnaire itself as appropriate (i.e. circling *yes/no*, providing a score and so on). The recording ensured that the informant understood the sentence in context, even if they were unfamiliar with the written form of a dialectal feature presented on the questionnaire. Furthermore, as is recommended by Schütze (1996: 193), it gave the sentences consistency of pronunciation and intonation across all informants in each separate location. In cases where the testing method required a more complex response than the *yes/no* type, a demonstration of the task was given by the fieldworker, and the informants were given a short practice session. In the discussion of results to follow, we focus on the extent to which findings from different types of testing method are consistent.

2.4.3 Data analysis

Let us begin by examining the reformulation task, which tested whether informants would use the dialectal features productively. In a previous pilot in Newcastle and Gateshead in 2006, informants were required to convert a declarative sentence containing a vernacular variant into an interrogative one. None of the reformulations provided by the informants contained any of the vernacular variants we were expecting. This was an interesting outcome since this strategy has been reported to have worked extremely well for the compilers of SAND/ASIS (Cornips and Jongenburger 2001; Benincà and Poletto 2007). Hence, in 2007, for our second pilot, we constructed the test so that the reformulation was from a more complex construction to a simpler one, that is, from an interrogative to a declarative. This adaptation of the technique in the second pilot proved much less problematic for informants and generated a numerical outcome, which, as we will see, was not ideal in certain respects either. Table 2.1 plots the occurrence of vernacular features (multiple negation, second person *yous*, Northern Subject rule and vernacular relatives) in the local informants' (NCL, GH and SL) reformulations. Hence, the higher the number per cell, the more willing the

Table 2.1 *Results for the reformulation task*
Interrogative → Declarative

	Reformulation totals		
Construction	NCL	GH	SL
Negation	3	4	3
Yous	5	3	0
NSR	2	6	0
Relatives	2	6	2

informants in this locality are to carry over this particular vernacular variant into the corresponding declarative sentence. Importantly, in this version of the test, reformulations containing all the vernacular variants of interest were constructed with ease, albeit with different frequencies across localities (about which we will have more to say later).

An important factor that needs mentioning here is the fact that, of the five questionnaire tasks, reformulation was the only one that required an original response from the informants in writing. Interestingly, this effect was clearly visible in the responses from a few informants, who rephrased the beginning of each interrogative-to-declarative reformulation. For example, they transformed, the initial *do*-support version to SVO but did not engage with structural aspects of the rest of the sentence.

Hence, informant BA from GH produced the following reformulations:

(12a) He did say he didn't want nothing to do with her.
(12b) You do want to come with us.
(12c) Finished your dinner have yous.

In fact, we strongly suspect that a certain proportion of the informants were simply replicating large chunks of the interrogative sentences into declaratives. This is suggested by our fieldworkers' notes which mention that informants tended to 'copy parrot-like', especially towards the end of this task (see also Schütze 1996: 191). Hence, responses gained via reformulation tasks of this kind need to be scrutinised very carefully.

We will now discuss the results for three further tests, the direct and indirect grammaticality judgement and pictorial elicitation tasks, focusing mainly on their comparability. Table 2.2 depicts the aggregated results for these, represented here as averages divided by both locality and testing method. Higher numbers imply that informants rated the variants as being more acceptable, while lower ones indicate more negative responses. For ease of orientation,

Table 2.2 *Average results for several testing methods by locality* (Higher numbers = more acceptable ratings)

Type of task	Vernacular features investigated			
Pictorial elicitation task	NSR	*Yous*	Relatives double conj.	Negation
NCL	1	**1.75**	0.75	**2**
GH	1	1	**0.92**	1.75
SL	**1.25**	1	0.83	0.75
Indirect judgement	NSR	*Yous*	Relatives double conj.	Negation
NCL	2.17	**3.31**	2.06	3.31
GH	2.81	3.19	**2.95**	**3.38**
SL	**3.08**	1.77	2.21	1.9
Direct judgement	NSR	Yous	Relatives double conj.	Negation
NCL	0.17	0.38	**0.5**	0.25
GH	0.42	0.38	0.33	**0.63**
SL	**0.75**	**0.62**	0.25	0.3

informant ratings in bold indicate that speakers are more accepting of the vernacular variants.

Generally speaking, the results in Table 2.2 yield important differences as well as consistencies across the three different testing methods. One might deduce from the figures given in the first column that all of the survey methods produce similar outcomes for the NSR, namely that the SL informants seem to be the most accepting of sentences containing instances of this vernacular feature. Unfortunately, though, this is the only occasion where all three testing methods lead to a similar result. Thus, the findings from the pictorial elicitation and indirect grammaticality judgement tasks for vernacular *yous* in column 2 suggest that the NCL informants are most accepting of this construction. However, the direct grammaticality judgement task for the exact same variant yields a different result, namely that it is the SL speakers who are more tolerant.

Furthermore, as the next column shows, the results of the pictorial elicitation and indirect grammaticality judgement tasks suggest that the GH informants are most positively disposed to the use of vernacular relative clause markers of various types. However, responses to the direct grammaticality judgement task for these variants generate a further conflicting result, indicating instead that speakers from NCL are, in fact, more accepting of vernacular relative markers than their peers.

These results might indeed support the view outlined above that direct grammaticality judgement tasks place so much normative pressure on

informants that they offer more prescriptive judgements for features that are shown to be at least marginally acceptable via other testing methods. That this is not, however, always the case is testified to by the figures in the last column of Table 2.2, representing informant responses to sentences with vernacular negative markers. In this case, the direct and indirect grammaticality judgements would lead one to assume that it is the GH informants who are the most accepting of this feature overall. However, responses to the pictorial elicitation task suggest that variants of this traditional dialect variable are, in fact, more robust in NCL than they are elsewhere in Tyne and Wear.[7]

The inconsistency across results yielded by the three methods employed here raises doubts as to the comparability of findings we have from various other research projects and reported in Barbiers *et al*. (2002). We still do not know to what extent we can compare, say, results from locality A collected via an indirect grammaticality judgement task with results from locality B culled from a pictorial elicitation task. For example, if we had administered just a single testing instrument – let's say a pictorial elicitation task – we would have confidently reported that informants in NCL are more accepting of vernacular negation strategies. Had we chosen a direct grammaticality judgement task instead, our results, by contrast, would have led us to conclude that it is, instead, the GH speakers who are more accepting of vernacular negation. The heart of the problem, therefore, lies in the fact that much of our previous knowledge of grammatical variation across the region is currently based on research which has adopted very different methodologies. And, as we hope to have demonstrated here, these methods are not necessarily comparable.

Testing the reliability of findings based on different methods commonly used in dialectology has thus not only revealed the limits of cross-test consistency, it also serves as a cautionary tale about the potential pitfalls of comparing results yielded from different testing methods. In our concluding sections, we discuss the results of the last test used in our research, namely, the magnitude estimation task.

2.5 How to make intuitions succeed

Table 2.3 displays the collective responses to the magnitude estimation task used here from all informants across the three localities of north-east England, divided by their location and gender.

As with the results for the other tests described above, Table 2.3 should be read such that higher numbers indicate that informants were more accepting of the sentences containing the vernacular variants. Hence, generally, and indeed rather uniformly, our GH subjects deemed all the constructions we tested as most grammatical (with the exception of the *yous* variant, which was

Table 2.3 *Average results for the magnitude estimation task* (Higher numbers = more acceptable ratings)

	Negation	*Yous*	NSR	Relatives
NCL	5.42	**6.57**	4.28	2.8
GH	**6.28**	5.55	**5.7**	**4.8***
SL	4.33	4.48	4.95	4.21
Male	3.42	4.54	4.21	3.44
Female	**6.93****	**6.49***	**5.8****	**4.49**

* p < .05, ** p < .01

rated as most acceptable by the NCL group). This finding, which is perhaps more uniform than one might expect compared to the quite erratic results from the three tests highlighted as problematic above, is especially interesting since it matches perfectly with the production data culled from the reformulation task. As has already been demonstrated in Table 2.1, it was the GH informants who were the most eager to carry over vernacular features into the reformulation task, except for the *yous* pronouns, which were more frequent amongst the NCL respondents. Hence, the results of both the reformulation and the magnitude estimation task correlate with respect to geographical space.

Moving on to the gender patterns in Table 2.3, we notice that our female informants are consistently less prescriptive, rating all the features tested as more acceptable. This is interesting, firstly, since it appears to contravene expected male versus female trends with respect to vernacularity (Romaine 2005) and, secondly, because the same female informants were, in fact, considerably more standard than their male peers regarding their production data (see Buchstaller and Corrigan 2008 and Buchstaller *et al*. forthcoming).

Generally, the results generated by the magnitude estimation task seem to be the most robust of all the testing instruments described thus far, indicating that females and speakers from GH are consistently more accepting of all the vernacular variants being tested (bar the use of the *yous* variant). Upon further scrutiny, however, the magnitude estimation task also revealed a certain amount of inter- as well as intra-speaker variability. As such, we would not wish at this stage to endorse magnitude estimation without reservation, as the tests would need to be undertaken on much larger population samples to warrant such conviction. However, it would seem, from the results presented here at least, that magnitude estimation is a method which can systematically capture native-speaker intuitions and thereby be used to uncover reportable and consistent patterns with respect to syntactic variation across either geographical or social space. It furthermore provides an important benefit not offered by

the other testing instruments described here, namely, that it extends to informants a suitably wide choice of grammaticality levels and it also allows them to personalise their choice so that they can be more confident about exactly what their ratings mean to them.

2.6 Where next?

This chapter began with the suggestion that we lack 'best practices' in accessing naturalistic informant responses to vernacular morphosyntactic data across English-speaking regions. Of particular note was the scant regard paid to this level of the grammar in traditional atlases like the *Survey of English Dialects* reported in Orton *et al*. (1962–71). While more recent and geographically expansive surveys such as those described in Volume II of the *Handbook* edited by Kortmann and Schneider (2004) document morphosyntactic variation globally, the methodology that underpins them remains less sophisticated and coherent than those of the large-scale Dutch, Italian and Scandinavian dialect atlas projects. These have provided new insights into the most appropriate and objective methods for collecting, measuring, describing and comparing information about syntactic patterning across space that could also be applied to the English context. These novel techniques are summarised in Benincà and Poletto (2007) and in the ground-breaking precursor Cornips and Poletto (2005). The question of adapting consistent and appropriate methodologies is a critical one, for two reasons: (1) the rise of comparative sociolinguistics, which has been made possible by the wider availability of large electronic corpora of vernacular English data (Beal *et al*. 2007a, b; D'Arcy, this volume) and (2) the expansion of syntactic atlas projects (particularly in Europe but elsewhere too), which rely on large-scale data collection across wide tracts of geographical space and are also comparative in purpose, as noted in Barbiers *et al*. (2002).

The investigation reported here of different methods used for accessing grammaticality judgements in three neighbouring varieties of English has shown that informants tend to produce divergent patterns of judgements depending on the test applied (as Schütze 1996 first predicted). The incongruence between the outcomes of different test types illustrates the fact that the results of testing methods commonly employed in dialect syntax in an English-speaking context need to be treated with much care. We need to solve this methodological paradox by developing theories of grammatical variability as well as consistent methodologies with which to test the variation across divergent communities of speakers. This is especially important since the internal factors which lead speakers to judge a variant differently across methods are multifactorial and still relatively little understood (Adger, p.c. 2008). Further research is needed to collate more empirical evidence to capture precisely what these factors are and exactly how they impact upon speaker judgements.

What is the way forward in the face of such heterogeneous methods and findings? We would like to suggest that, in the long run, as well as sampling suitably large populations, the most consistent results can only be arrived at by employing a multi-method or layered approach, such as that advocated on the basis of the SAND/ASIS fieldwork techniques described in Benincà and Poletto (2007) and Cornips and Poletto (2008). We also felt it to be important in this chapter to raise awareness of the advantages and disadvantages of different judgement task types and the extent to which the results they yield can be directly compared in any meaningful way. In addition, data collection methods which endorse a greater sensitivity on the part of fieldworkers to the needs of respondents seem more likely to produce consistent, reliable and replicable results. Furthermore, we have shown that classic methods, such as magnitude estimation or reformulation tasks, can be sensibly adapted to suit the skills of a wide range of informants. In fact, while we have voiced some reservations regarding the potential for variability within the magnitude estimation task itself, the consistency of results across social attributes would lead us to suggest that well-conceived magnitude estimation tasks, especially in combination with other tasks, can produce relatively stable results that concur with the overall production rates of vernacular variants. This is an important avenue for future research which might test the extent to which such tasks are equally suitable for uncovering the trajectory of morphosyntactic change in a range of localities in the English-speaking world (by comparing responses to the task by different generations of speakers, for instance). The method can also be usefully employed to explore differences that one might attribute to social class or ethnic group membership.

Websites

www.meertens.knaw.nl/projecten/sand/sandeng.html
www.uit.no/scandiasyn?/Language-en
www.ons.gov.uk/about-statistics/classifications/current/ns-sec/index.html

3 Corpora: capturing language in use

Alexandra D'Arcy

3.1 Introduction

Language cannot be invented; it can only be captured. (Sinclair 1997: 31)

The enterprise of investigating language variation is based on access to empirical data – language as actually used by speakers and writers. This is not trivial. We only know what we do about variation in English (or for that matter, in any variety, dialect, register, etc.) through analysis of language in some collection of materials. This collection, 'the corpus', is the foundation of everything we do. The data might consist of a collection of letters and diaries, spoken narratives of personal experience, or a compilation of text logs from instant messaging conversations. The materials that provide data for variation studies are diverse, but what unites them is their empirical validity as representations of language in use and, as a consequence, our dependence on them. The simple truth is that we cannot engage in the study of language variation without access to a corpus of data on which to test our hypotheses, base our analyses, and inform our theories, yet this simple truth masks a number of not-so-simple issues. How are corpora constructed? If a corpus contains spoken language, what is the best way to represent the speech in written format? How are corpora accessed and mined? What methods achieve what results? How should the results be interpreted (i.e. what do they mean, what do they tell us?)? This chapter explores these kinds of questions but it intentionally presents few solutions. As you read it will become clear that answers to these questions are rarely binary choices between 'right' and 'wrong'. There may be a number of possible solutions and the determination of which best meets the needs of a particular project will depend on other factors. For example, how much annotation should be included? How detailed should the transcription system be? The answers to questions like these vary from project to project: there is no standard protocol (see Poplack 2007: x). Before diving headlong into these

I am grateful to the following people, who contributed valuable information and feedback during the writing of this chapter: Karen Corrigan, Elizabeth Gordon, Janet Holmes, Margaret Maclagan, Warren Maguire, April McMahon, Charles Meyer, and Shana Poplack.

types of discussion, however, it may be helpful to define just what is meant by 'corpus' in the field of variation studies, since it is there that language, with its inherent variation, is captured.

3.2 What is a corpus?

Although there are a number of publications dedicated to corpora, some effort is required to find those that include a definition of just what a corpus is. This suggests that most linguists working with empirical data-sets take the referent for granted. But, in fact, there is a wide range of corpus types and, as new technologies are developed and new methodological innovations are made, the types of data available for analysis are increasing. This makes defining what is and is not a corpus challenging because the target is constantly shifting and, depending on one's point of view, the basic composition can differ quite radically. For example, corpus linguists generally view a corpus in the electronic era as a collection of computerised language texts (e.g. Sinclair 1991: 171; Kennedy 1998: 1; Biber *et al*. 1999: 24; Meyer 2002: xi), but sociolinguists would find this too restrictive because the emphasis on computerised texts excludes sound recordings and data collected through methods such as surveys, questionnaires, and wordlists (see Bauer 2002: 98).

At its most basic level, a corpus is evidence, evidence of what was and evidence of what is. It is thus free of prescriptivism (what one should say) and intuitions (what one thinks is said). But it is clear that as a definition 'evidence' is inadequate because it is too vague, while 'a collection of texts' is inadequate because, in this context, it is too restrictive. A useful compromise between these two extremes is the definition provided by Bauer (2002: 98), where a corpus is seen as 'a body of language data which can serve as a basis for linguistic analysis and description'.

Within this broad purview we can make a number of distinctions. The first concerns the thematic categorisation of corpora as either *conventional* or *unconventional* (i.e. dialectal), which aligns largely with the methodological frameworks of corpus linguistics on the one hand and variationist sociolinguistics on the other.[1] The key differences derive from the way in which corpora are constructed (i.e. their composition) and their projected use as either '*end-product* or *tool*', to use Poplack's terminology (2007: xi). These points will be elaborated more fully below, but for now it is sufficient to note that the *raison d'être* of each type of corpus is distinct. This has ramifications for all aspects of corpora, from the type of language targeted to the methodological assumptions guiding data extraction. It is important to bear in mind, however, that the ultimate concern of both corpus linguistics and variationist sociolinguistics is the description of the structure and use of language and that they are complementary modes of enquiry (see Kretzschmar *et al*. 2006). It is also important to

bear in mind that while the discussion of corpus types presents them as discrete entities, they in fact represent what is best described as 'a multidimensional matrix' (Bauer 2002: 100) with fuzzy boundaries and intersecting features. For example, many corpora that would fall into the 'conventional' category include regional and social components (e.g. BNC, Switchboard), but methodologically they were designed to address more universal goals rather than to capture the 'special qualities of speech' of a particular region or social circumstance (Kretzschmar *et al*. 2006: 174).

3.2.1 Conventional corpora

Conventional corpora cross-cut a range of fields of scholarship such as lexicology and lexicography, literary studies, grammar studies, computational linguistics, language acquisition and language pedagogy, as well as descriptive linguistics. In general, the focus of these corpora is written language (Kennedy 1998: 20; also Leech 1993b), typically representing standard and more formal registers. The Brown University Standard Corpus of Present-Day American English (Brown), which became the model for a number of subsequent corpora-construction projects (see below), was specifically designed to capture standard printed American English (Francis and Kučera 1964: xvii). It consists of 500 2,000-word samples that were selected from fifteen categories reflecting two prose types (see Table 3.1): informative (374 samples; nine categories) and imaginative (126 samples; six categories).

Some conventional corpora also include spoken language. As with the written components, the data are drawn from a range of registers (e.g. lectures, interviews, telephone conversations). Half of the texts in the Survey of English Usage (SEU) (Quirk 1968) represent spoken English, monologic and dialogic (see Table 3.2), and similarly varied speech-based texts can be found in the British National Corpus (BNC), the Lancaster-Oslo-Bergen Corpus (LOB), the Freiburg-Lancaster-Oslo-Bergen Corpus (FLOB), and the Freiburg-Brown Corpus of American English (Frown). The individual corpora that comprise the International Corpus of English (ICE) contain more speech-based texts than written ones,[2] while the Santa Barbara Corpus of Spoken American English (SBCSAE) is based entirely on spoken registers.

Corpora such as Brown are constructed with a similar overarching goal: a representation of language in use. Their purpose is open ended in that hypothesis formation follows data collection. ICE, for example, was conceived so as to allow for comparative studies of English worldwide, which is quite distinct from compiling a corpus to answer a specific question about variation across global Englishes. It is precisely their open-ended nature that has made conventional corpora a prolific and valuable resource in the descriptive tradition. Consider just one aspect of English, the modal auxiliary system. Krug

Table 3.1 *The Brown corpus (Francis and Kučera 1964; after Kennedy 1998: 24–6)*

Category	Subcategory	Total	%	Category	Subcategory	Total	%	Category	Subcategory	Total	%
Informative prose			75.0	Informative prose cont'd				Imaginative prose			25.0
Press: reportage				Popular lore				General fiction			
	political	14			books	23			novels	20	
	sports	7			periodicals	25			short stories	9	
	society	3		**Total**		**48**	**9.6**	**Total**		**29**	**5.8**
	spot news	9		Belles lettres, etc.				Mystery, detective			
	financial	4			books	38			novels	20	
	cultural	7			periodicals	37			short stories	4	
Total		**44**	**8.8**	**Total**		**75**	**15.0**	**Total**		**24**	**4.8**
Press: editorial				Miscellaneous				Science fiction			
	institutional	10			government docs	24			novels	3	
	personal	10			foundation reports	2			short stories	3	
	letters to editor	7			industry reports	2		**Total**		**6**	**1.2**
Total		**27**	**5.4**		college catalogue	1		Adventure, western			
Press: reviews		17			industry house organ	1			novels	15	
	religion			**Total**		**30**	**6.0**		short stories	14	
	books	7		Learned				**Total**		**29**	**5.8**
	periodicals	6			natural sciences	12		Romance, love story			
	tracts	4			medicine	5			novels	14	
Total		**17**	**3.4**		mathematics	4			short stories	15	
Skills and hobbies					social, behavioural science	14		**Total**		**29**	**5.8**
	books	2			political science, law	15		Humour			
	periodicals	34			humanities	18			novels	3	
Total		**36**	**7.2**		tech. and engineering	12			essays etc.	6	
				Total		**80**	**16.0**	**Total**		**9**	**1.8**

Table 3.2 *The SEU corpus, spoken texts (Quirk 1968; based on Kennedy 1998: 18)*

Category	Subcategory	Total	%	Category	Subcategory	Total	%
Monologue Spontaneous		24	12.0	Dialogue Face-to-face conversation		76	38.0
	oration	10	5.0		surreptitiously recorded	34	17.0
	commentary				non-surreptitiously recorded	26	13.0
	sport	4	2.0				
	non-sport	4	2.0				
Prepared, unscripted oration		6	3.0	Telephone conversation		16	8.0

(2000), a detailed monograph of historical and ongoing grammaticalisation in this system, drew on ARCHER, Brown, Frown, LOB, FLOB, the BNC, and the Helsinki corpora. In more recent, smaller-scale works, Leech (2003) used Brown, Frown, LOB, FLOB, SEU, and ICE-GB, Smith (2003) used SEU and ICE-GB, and Collins (2005) drew on Frown, the SBCSAE, and three of the ICE corpora (GB, AUS, NZ).[3]

What makes conventional corpora particularly well suited to large-scale analyses such as that presented in Krug (2000) is their size. Biber *et al*. (1999: 27) consider a corpus which consists of 50,000 to 2 million words as 'relatively small', while a 'very large' corpus would include over 100 million words. At 40 million words, the Longman Spoken and Written English Corpus (LSWE) represents the 'middle ground' (1999: 28).

In certain respects the size of a corpus depends on the broader goals of the compilers. *Balanced* or *core* corpora, designed to represent a wide range of registers with balanced amounts of text, are typically (but not exclusively) toward the smaller end of the scale (Biber *et al*. 1999: 27). Examples include Brown and LOB. *Opportunistic* corpora, whose primary design feature is size, tend to be very large. According to Biber *et al*. (1999: 27), 'such corpora do not represent registers in a systematic way and give little or no attention to the random selection of texts; they are based on the assumption that all important patterns will be represented if the corpus is large enough'. These types of corpora tend to consist of texts that are already available in electronic form and so can be compiled with relative speed and efficiency. The American National Corpus (ANC) contains an opportunistic collection of texts (see Macleod *et al*. 2000).

The corpora discussed so far are *static*. It is not necessarily the case, however, that once compiled the contents (and by extension, the size) of a corpus are fixed. More recently corpus linguistics has seen the advent of *dynamic* or *monitor* corpora. In a dynamic corpus, new texts are added regularly, sometimes replacing earlier texts (Renouf 1993; Sinclair 1992). The Bank of English, part of the Collins Birmingham University International Language Database (COBUILD), is one such project.

3.2.2 Unconventional (dialect) corpora

Sociolinguists have traditionally depended on unconventional corpora. These are corpora that focus on distinct dialects, be they ethnic, regional, or social. The ultimate goal of dialect corpora is to 'tap the vernacular' (Sankoff 1988: 157), unmonitored, informal, everyday speech. This generally involves a sociolinguistic interview in which speakers are encouraged to converse as 'naturally' as possible (Labov 1984; Tagliamonte 2006a), but it can also involve more indirect means of accessing vernacular norms such as wordlists, questionnaires, and grammaticality judgements (see Buchstaller and Corrigan, this volume). The fundamental role of speech harks back to dialectological traditions in which the object of interest was *traditional dialect*, the 'speechways of the folk' (Kurath 1972: 13), but in modern sociolinguistics it was Labov's seminal work on phonological variation in Martha's Vineyard and New York City (Labov 1963, 1966, 1969, 1972a, b) that laid the foundations for the empirical investigation of *parole*, language as actually spoken.[4]

Dialect corpora are sometimes referred to as *specialised*. This is because they are designed with a particular research question in mind. Hypothesis formation precedes data collection. For example, Poplack (1989) constructed a corpus of spoken Canadian French to examine the linguistic effects of long-term, stable bilingualism on the language in both its minority (Ottawa) and official (Hull) language guises. For this project, ethnographically inspired sociolinguistic interviews were collected from 120 speakers stratified by age, sex, and the status of French in their neighbourhood. The result: 270 hours of spontaneous dialogue and 3.5 million words. The more recent Quebec English Corpus (Poplack *et al*. 2006) was designed to 'assess the impact of a majority language on the structure of the minority language in a situation of long-term contact' (p. 186). Data collection was carefully planned with this goal in mind: target communities were selected based on the proportion of English mother-tongue claimants, ranging from just 1.5 per cent to 86 per cent, while the choice of informants crucially stratified speakers according to the socio-political period during which they acquired English. This corpus, while somewhat smaller than the Ottawa-Hull corpus, nonetheless includes 340 hours of informal conversation and 2.8 million words, collected from 183 speakers. At the other end of the

scale, D'Arcy (2001, 2005b) collected a corpus of spoken English to examine the effects of parentage (local vs. 'from away') on the acquisition of local dialect features in St John's, Newfoundland. This corpus contains data from just 16 speakers (all female, aged 8–11 and 16–17) and consists of approximately 10 hours of conversation representing just 55,000 words. The Ottawa-Hull, Quebec English, and St John's parentage corpora clearly differ in scale but they were similarly created with a focused question in mind (the effects of language contact; the effects of dialect contact).

The Ottawa-Hull and Quebec English corpora are noteworthy for a number of reasons (rigorous sampling methods, accountability to the data, etc.), but among specialised corpora their size is of particular significance. At 3.5 million words, the Ottawa-Hull corpus remains – two decades after its construction – one of the largest dialect corpora. The Michigan Corpus of Academic Spoken English (MICASE) contains 1.8 million words (Simpson *et al.* 2002), the York corpus (Tagliamonte 1998) 1.5 million words, the Wellington Corpus of Spoken New Zealand English 1 million words (Vine *et al.* 1998), the Bergen Corpus of London Teenage Language (COLT) 500,000 words (Haslerud and Stenström 1995), the Northern Ireland Transcribed Corpus of Speech 400,000 words (Kirk 1992), the Ayr corpus 120,000 words (Macaulay 1991b). The size of these corpora relates directly to the extreme time demands involved in constructing them. Depending on the complexity of the dialogue (number of participants, background noise, voice quality, fluidity, etc.), it takes approximately ten hours to transcribe orthographically one hour of speech. This is because, in conversation, speakers can produce 120 words per minute, amounting to over 7,000 words per hour (see Biber *et al.* 1999: 27). For the St John's parentage corpus, each interview, which consisted of two friends talking together for 45 minutes to 1.5 hours, took a full day to transcribe. For the Quebec English corpus (340 hours of speech), 2,471 hours were invested in transcribing the materials while a further 1,536 hours were spent checking and correcting the transcription files (Poplack *et al.* 2006: 194 f.).

In contrast to conventional corpora, sociolinguistic dialect corpora tend to be *private* (i.e. not available to linguists in general; see Bauer 2002). The creation of 'private resources' remains the default for unconventional corpus projects (Kretzschmar *et al.* 2006: 180). This derives from their very nature as specialised data-sets. Because each corpus is designed by a particular researcher to answer a particular question, it remains the property of the primary investigator(s). Moreover, the informed consent documents often do not stipulate making the materials public. Figure 3.1 contains the section from the Origins of New Zealand English (ONZE) project consent form that details access to, and use of, the data. While this agreement does allow for samples to be used in public domains (e.g. short excerpts may be heard

through online journals with sound access, akin to the transcribed examples that appear in articles), it does not allow for the data to be made generally available.

In some ways, though, ONZE presents a special case, demonstrating how the line between *public* and *private* is not clear-cut (cf. Bauer 2002). ONZE consists of three separate collections (see Gordon *et al*. 2007): the Mobile Unit Archive, a collection of interviews gathered by the New Zealand Broadcasting Service between 1946 and 1948, the Intermediate Archive, an ad hoc collection of oral histories gathered by a range of individuals during the 1990s, and the Canterbury Corpus, a socially stratified judgement sample for which sociolinguistic interviews have been ongoing annually since 1994. Whereas the copyright for the Canterbury Corpus is held by the University of Canterbury, that for the Mobile Unit recordings is held by the Sound Archives of Radio Zealand. Thus, while all the ONZE materials are available to bona fide researchers at the University of Canterbury (see Fig. 3.1), only the Mobile Unit data can be accessed more generally, either through the Sound Archives or the Alexander Turnbull Library in Wellington. The requirement to travel to New Zealand, however, renders the Mobile Unit less public in terms of ease of accessibility than, for example, the Newcastle Electronic Corpus of Tyneside English (NECTE), which can be accessed on the Web (www.ncl.ac.uk/necte/).

3.2.3 The time dimension

All corpora are bound in time as either *synchronic* or *diachronic*. The former represent language at a particular point. Brown and LOB capture American and British English respectively in 1961, Frown and FLOB do the same for 1991. COLT was collected in 1993, the Quebec English corpus in 2002. In contrast, a diachronic corpus represents language over a period of time. The diachronic part of the Helsinki Corpus (Kytö 1996) contains English texts from 700 to 1700, covering almost the whole of the Old English period through to the end of the Early Modern English period, but smaller windows are also possible. The Linguistic Atlas of Early Middle English (LAEME) covers the years 1150 to 1325 (Laing and Lass 2007), the Linguistic Atlas of Older Scots (LAOS) 1380 to 1500 (Williamson 2008), the Corpus of Nineteenth-century English (CONCE) 1800 to 1900 (Kytö *et al*. 2000).

Many dialect corpora are synchronic. At the foundation of the variationist enterprise is the apparent time hypothesis, the assumption that the vernacular stabilises after adolescence. Apparent time is thus a theoretical construct that allows diachrony to be viewed from a synchronic perspective: generational differences among speakers sampled at the same time are assumed to be temporal analogues, reflecting historical stages of the language (for extensive discussion, see Tagliamonte and D'Arcy 2009). Thus, the three archives of ONZE, which

AGREEMENT

I agree that the recording of my interview and accompanying material be:

1. Held in the University of Canterbury linguistics archives.
2. Made available to bona fide researchers.
3. Quoted in published work or broadcast or used in public performance in full or in part.
4. Used for teaching purposes.
5. Used as an illustration on a web site (short and anonymous, non-personal excerpts only).

Signature of Interviewer: _____

Signature of Interviewee: _____

Date: _____

Figure 3.1 Excerpt from the ONZE informed consent form (Canterbury Corpus)

contain data from speakers born in the period from 1851 to 1987, together capture the full history of New Zealand English despite the fact that the earliest recordings were made following World War II (see Gordon *et al*. 2004).

3.2.4 Advantages (and things to think about)

The strengths and weaknesses of any particular corpus-construction philosophy depend in part on the goals of individual researchers. In the end, no corpus can provide data for all linguistic phenomena, variable or otherwise (see Meyer 2004), but each corpus type has its advantages.

A crucial design feature of text-based corpora such as Brown and LOB, and more recently the individual corpora of ICE, is comparability. Indeed, ICE was designed specifically for this purpose (Greenbaum 1992), just as LOB was intended to be the British counterpart to Brown (Johansson *et al*. 1978). Other corpora using the Brown model are the Kolhapur corpus (Shastri 1988), the Wellington corpus (Bauer 1993), ACE, and Frown and FLOB, the Freiburg versions of Brown and LOB, which were intentionally constructed as direct replicas.

The shared sampling methods render these corpora compatible at a fairly high level of confidence, but a certain amount of caution is nonetheless

required. For example, does variation simply reflect text-selection or publishing practices in the respective varieties? Brown includes six fiction categories (cf. Table 3.1), but when the Wellington Written Corpus of New Zealand English was constructed, the difficulty of matching these led to the decision to put all fiction into a single category. At the same time, a major New Zealand category – children's fiction – was omitted because it was not part of the model (see Bauer 1993; cf. Kennedy 1998). As discussed by Biber (1988), there is also the possibility for variation within a genre. For example, academic prose has a number of sub-genres (e.g. natural vs. social science), as does press reportage (political vs. cultural).

Speech-based corpora are necessarily more idiosyncratic in nature. This is unremarkable: 'the underlying theoretical goals and assumptions of the researchers are quite distinctive' (Beal *et al*. 2007b: 2), a fact that has consequences for every aspect of the corpus from the nature of the data itself (group discussion with lots of overlap to the 'interactive written discourse' of instant messaging (Ferrara *et al*. 1991: 8)) to the sample represented (speakers of all ages to just one age group, e.g. 65 years and older). This does not mean that the comparative method is moot. As Tagliamonte (2002: 729) points out, '[c]omparison has always been at the root of sociolinguistics'. It has played a central role in variationist theory, from discourse pragmatics to morphosyntax (e.g. Poplack and Tagliamonte 2001; Tagliamonte and Smith 2006; Buchstaller and D'Arcy 2009). In addition to taking the vernacular as a keystone, many specialised corpora share a number of socially stratified categories (e.g. age, sex, ethnicity, education level, etc.). The key is to construct (or situate) corpora with similar purposes and/or analogous design features.

The bottom line is that regardless of corpus type, comparisons must always be approached with caution. We need to be aware of why a corpus was constructed and how the contents might vary, from differences in composition such as those highlighted by the fiction categories of the Brown and Wellington corpora to differences in the stylistic guidelines adhered to, either by different newspapers (e.g. the *Guardian* vs. the *Daily Mirror*) or by different sections within the same newspaper (e.g. sports vs. other sections) (cf. Meyer 2004).

In general, conventional corpora are much larger than unconventional corpora. Their sheer size, coupled with the range of genres and registers represented, makes them the lifeblood for dictionaries and grammars. They are also fundamental to analyses of frequency effects, lexical variation, and grammatical variation, for which vast amounts of data are required. These types of corpora are less effective for the study of discourse features (which can vary throughout a text) because the data are typically compiled from fragments rather than texts in their entirety (e.g. ICE samples 2,000-word excerpts). Different parts of a text may also be characterised by 'marked lexical and syntactic differences' (Stubbs 1996: 32; cf. Meyer 2004: 347). While this may be less critical

for newspaper texts, which tend to be shorter, texts from other genres may be systematically cropped to fit with the overall design of the corpus. Given their textual basis, these types of corpora are also not usually amenable to phonological or phonetic analysis, though in some cases it is possible to access the original sound files.

Despite the range of registers represented in conventional corpora, the spoken components 'have typically been collected in restricted or artificial settings' (Biber *et al*. 1999: 28). To date, the most representative samples are those of the LSWE and SBCSAE, which contain naturally occurring spoken discourse. The vernacular emphasis of dialect corpora circumvents this issue, since the ultimate goal is to obtain 'real language in use' (Milroy 1992: 66). As summarised by Shana Poplack (p.c. 26 June 2008; emphasis in original):

> What distinguishes our corpora is not simply size (corpora constructed from newspapers or other written text will always be exponentially larger), but the fact that they consist of the *real* speech of *real* people, sampled in such a way as to answer specific research questions.

Specialised corpora are purpose built and they present an authentic model of the variety of speech from which all others are calibrated. The data are 'infinitely more rich than the precategorized material in other disciplines' (Sankoff 2005: 999).

3.3 What's out there?

The purpose of this section is to give a sense of existent English corpora. It is impossible, however, to provide a complete overview. There are simply too many and corpus construction projects (public and private) are likely to continue ad infinitum. But it is also the case that many public corpora come at a cost, literally. The International Computer Archive of Modern and Medieval English (ICAME) collection, for example, costs 3,500 NOK for an individual user licence (at the time of writing, roughly equivalent to 685 USD, 345 GBP, or 440 EUR).[5] For those without the necessary funds (i.e. most students), these fees present the ultimate barrier.

In what follows, some public and free corpora are briefly outlined. The list is by no means exhaustive; it is simply intended as a starting point for students interested in variation in English.

3.3.1 Dialect atlases

Dialect atlases are an excellent source of data for studying variation and most university collections include at least one (some have more than a hundred). Both the geographic representation and the historical time depth of English dialect atlases allow for innumerable investigations of lexical, phonological,

and phonetic variation across time and space. The most recently published is the *Atlas of North American English* (Labov *et al*. 2006), but for a point of historical comparison one can also find the *Linguistic Atlas of New England* (Kurath *et al*. 1939–43). Online, there is the *Linguistic Atlas Projects*, a portal to a number of atlas projects in the United States (e.g. African American and Gullah Project (AFAM), LAGS, LAMSAS, etc.). You can also access the *Dialect Topography Project* (Chambers 1994), which investigates words (and their pronunciation) used both in Canada and in regions of the United States that border Canada.

3.3.2 The Oxford Text Archive

The Oxford Text Archive is a repository for literary and linguistic resources. Most of the holdings are in text format, but some audio and video files are archived as well. All the texts can be accessed for free simply by submitting your email address (used to send the link to the text of interest), but for those marked 'restricted' users are required to register before the resource can be downloaded.

3.3.3 Text- and speech-based corpora

Among traditional text-based and speech-based corpora, there are a few that can be accessed via the Internet for non-profit academic research.[6] In most cases, a password is required, obtainable by downloading the appropriate access request form and/or licensing agreement.

- *Brown University Standard Corpus of Present-Day American English*: Via a guest account (as opposed to purchasing a membership), the full text of the Brown corpus can be accessed through the Linguistic Data Consortium, LDC Online. Guests can also access an indexed collection of Arabic, Chinese, and English newswire text, the Switchboard and Fisher collections of telephone speech, and the American English Spoken Lexicon.
- *Buckeye Natural Speech Corpus*: The Buckeye corpus is a sociolinguistically stratified corpus of unmonitored casual conversations from Columbus Ohio. It includes data from forty speakers (male and female, over 40 years old and under 30 years old) in text and audio format. The materials can be accessed for research and teaching purposes after submitting a completed licence agreement.
- *Corpus of Early Ontario English, pre-Confederation section*: CONTE-pC is a diachronic, text-based corpus of early Canadian English with three genres (newspaper texts, diary entries, letters). It is similar in design to ARCHER (A Representative Corpus of Historical English Registers), enabling comparisons

with other historical varieties of English (see Dollinger 2008: 99–119). At the time of writing, CONTE-pC is in the final proof-reading stage but once complete it will be available through the Oxford Text Archive.[7] Period covered: 1776 to 1849.

- *International Corpus of English*: The ICE corpora include both written and spoken texts. Of the eight completed regional corpora (thirteen others are currently under construction), five can be accessed free of charge through the ICE site: East Africa, Hong Kong, India, Philippines, and Singapore.
- *Newcastle Electronic Corpus of Tyneside English*: NECTE is a public dialect corpus from Tyneside (Allen *et al*. 2007). It consists of two synchronic corpora, one from the late 1960s and one from the early 1990s. The materials are available in a variety of formats (digitised audio, standard orthographic transcription, phonetic transcription, POS-tagged) and may be accessed by students (undergraduate and postgraduate), academics, and members of the public for bona fide research purposes (e.g. class projects, research) upon submitting the access request form.
- *Santa Barbara Corpus of Spoken American English*: The SBCSAE contains naturally occurring discourse from across the USA (e.g. Alabama, California, Montana, New Mexico, Washington, etc.). Most of the conversations are face-to-face interactions, but some record other modes of discourse such as telephone conversations, lectures, medical interactions, and narratives of personal experience.[8] The SBCSAE can be purchased in CD or DVD format from the LDC or the transcripts and their corresponding audio files can be downloaded from TalkBank.
- *Scottish Corpus of Texts and Speech*: The SCOTS corpus contains written and spoken texts of Scots and Scottish English, and includes a handful of Scottish Gaelic texts as well. The corpus covers the period 1945 to 2007, though most of the spoken texts (which are synchronised with the audio recordings) were recorded after 2000. After agreeing to the terms and conditions outlined on the site at www.scottishcorpus.ac.uk/termsandconditions.html, SCOTS can be searched online at no charge to the user.

3.3.4 The World Wide Web

Finally, the Web itself can be a corpus and there are search engines available for this purpose. Two in particular have been designed to retrieve linguistic data from the Web: WebCorp (Renouf 2003; Morley 2006) and GlossaNet (Fairon 2000). GlossaNet is an automated service that monitors the websites of more than 100 newspapers. Once a search item, dates, and intervals are specified, GlossaNet applies the queries and the results are emailed to the user in the form of a concordance (a display of the search item with its surrounding context). Because GlossaNet builds new corpora every day, downloading

current editions of newspapers, it is a dynamic corpus. WebCorp is more versatile. It can 'piggy-back' on existent search engines (Google, Altavista, Metacrawler) and is not limited to newspapers. A basic search will result in concordance lines of the query item, but the program also has a built-in suite of tools that enable a number of more advanced searches like pattern matching (e.g. *is * nice* will match *is so nice*, *is really nice*, *is very nice*, etc.) or specifying the target domain (e.g. the New Zealand academic domain, *ac.nz*, the BBC website, *bbc.co.uk*, the Canadian government, *gc.ca*).

3.4 Constructing a corpus

It is often the case that existent (or accessible) corpora are not suited to a particular project. Building your own corpus is challenging yet extremely rewarding; challenging because there are numerous issues to consider, rewarding because in the end you have a personalised window on the very issue in which you are interested. This section highlights four central concerns of corpus construction: representativeness, transcription, annotation, and accessibility. Overarching these four concerns is the ultimate function of the corpus. Is it intended to stand alone (e.g. as a specialised data-set, such as the Quebec English corpus) or to complement an already existing corpus (e.g. with periods that align in some way to those in another corpus, as with CONTE-pC and ARCHER)? When undertaking any corpus-construction project – regardless of scope or size – there are a number of important considerations to bear in mind. The ways in which the issues raised in this section are addressed by any individual project depend on the answers to questions such as the following:

- Is the corpus strictly for personal use (e.g. for a class project, honours paper, Ph.D. dissertation, etc.)?
- Is it important that the corpus be free-standing because the research imposes certain requirements, or is it crucial that it be compatible with some other corpus?
- Could the corpus be used for research other than that which spurred its construction?
- What will happen to the corpus once it is complete?
- Might someone else be interested in the data? If so, how might they access it (e.g. through personal request, a website, or not at all; as text files only or with audio)?
- If the corpus is to be public, what is the protocol for transcription, annotation, or digitisation?

3.4.1 Representativeness

A primary aim of all corpus-construction projects is *representativeness*, a model of the population or universe to be sampled. The question is how to

best achieve this. Some corpora aim for representativeness through size, but the more common method is through sampling. Sankoff (2005: 1000) notes, 'A more useful notion of representativeness requires not that the sample be a miniature version of the population, but only that we have the possibility of making inferences about the population based on the sample.'

Balanced corpora seek representativeness through the range of genres sampled. There is, however, no comprehensive taxonomy of genres from which to select (Kennedy 1998: 62). Further, given the population of newspaper texts alone, is one million words representative? How many papers to sample? What sections to sample? Sinclair (1991: 20) has suggested that when compiling a written corpus, the texts must minimally differentiate between fiction and non-fiction, or formal and informal, etc.

For specialised dialect corpora the question of representativeness is particularly complex. At the heart of the matter is the target population (i.e. the sampling universe). For the Quebec English corpus (Poplack *et al.* 2006), the sampling universe was the population of anglophones in Quebec and Ontario, Canada. Represented were individuals of different ages, sexes, different times and conditions of acquisition of English, different socio-political statuses (minority vs. majority), etc.

A corollary to sampling is defining a speaker/writer of a given variety. In compiling CONTE-pC, Dollinger (2008: 103–6) was faced with delimiting what, in the historical context, makes a text Ontarioan. For contemporaneous corpora, the increased mobility of many populations is a confounding factor. For ICE-NZ, a native speaker was one who had lived in New Zealand since before the age of ten, had not spent more than ten years (or more than half his/her lifetime) overseas, and had not returned from a trip overseas within the last year (Holmes 1996). In less mobile communities, the definition of a native speaker might be more rigid. In the St John's parentage corpus (D'Arcy 2001, 2005a), speakers were born and raised in the city or its immediate surrounds and had spent little or no time outside the province.

Things to think about:

- *Speakers/Texts*: Who/Which? How many genres?
- *Demographics*: What social factors might be relevant? Which should be incorporated and which should be controlled for? (e.g. age, gender, ethnicity, education, occupation, region, housing type, neighbourhood, linguistic background, time spent in community)
- *Registers*: What style(s) will best serve your needs? (e.g. monologic or dialogic, formal or informal, free conversation, narratives of personal experience, task-related)
- *Size*: How much data is needed? If written, full texts or partial texts and how many of each? If spoken, how many speakers and how long should each recording be?

3.4.2 Transcription

If the corpus includes spoken data, the speech will need to be recorded in written format.[9] Before the first word is typed it is necessary to decide on a format. For the Toronto English corpus (Tagliamonte 2006b) transcriptions were created as Word documents, with no link between the audio and text files, but for the ONZE project (Gordon *et al*. 2007) they are made using *Transcriber*, a freeware utility that synchronises the text with the corresponding part of the recording. NECTE, on the other hand, intended as a public corpus, conformed to emerging global standards for the encoding of text; the files are in the form of TEI-conformant XML syntax (alternative models are provided by TalkBank and the LDC; see, e.g. MacWhinney 2000 for TalkBank or www.ldc.upenn.edu/Creating/ for LDC conventions).

Language is inherently variable but speech is inherently messy, full of false starts, hesitations, repetitions, and the like. In making a corpus machine readable, the major challenge is to ensure that the recorded speech is represented 'faithfully and consistently' (Tagliamonte 2006a: 55). Inevitably, a transcription is only an interpretation; 'it can never be so detailed and precise as to provide for the recreation of the full sound' (Macaulay 1991a: 282). But more detail does not mean increased quality: as detail increases (e.g. ellipses, reductions, spelling pronunciations), the more cumbersome the transcription becomes.

Most researchers stress the need to follow standard orthographic conventions unless there are strong motivations for proceeding otherwise. A case in point is the use of dialect forms that do not appear in standard dictionaries (e.g. *nae* for *no*; *nowt* for *nothing*; *tiv* for *to*; *whae* for *who*). Less agreed upon is the use of standard punctuation. Some feel that using full stops, commas, and question marks is critical (e.g. Preston 1985, 2000; Tagliamonte 2006a, 2007), while others reserve their use for special cases. The ONZE protocol stipulates that question marks can be used, especially in cases where the intonation indicates a question but the syntax does not, but that commas and full stops should not be used. Instead, a full stop with a space on either side is used to mark a short hesitation (though the recent decision to introduce syntactic parsing has necessitated a revision of the punctuation protocol). Decisions also have to be made regarding colloquialisms (ONZE allows *gonna*, *gotta*, and *wanna*, but not *hafta*, *woulda*, or *mighta*), hyphenation (which affects word counts and concordances), and other sundries associated with unscripted dialogue (overlapping or incomprehensible speech, backchannelling cues, etc.). In the end, consistency is crucial. The following are some questions to consider when designing a transcription protocol (after Macaulay 1991a: 287):

Things to think about

- What is the purpose of the transcriptions? (e.g. analytical, illustrative)
- Will others have access to them? If so, will the representation of dialect features be clear to researchers not familiar with the community or will they be opaque?
- Is there a purpose to a certain representation? What does it buy you and is it consistent with other decisions?
- Are the features predictable from general phonetic rules, and if so, are they better left out of the transcriptions? (e.g. consonant cluster simplification, assimilation, vowel reduction)

3.4.3 Annotation

A key feature of corpus construction is annotation. This refers to marking-up the text with explicit information about its linguistic form and content. The most common type is *part of speech* (POS) or *grammatical* tagging, which affixes a label to each word indicating its grammatical function. Table 3.3 lists some examples from the tagset used in the BNC, generated by the CLAWS program (Garside 1987). Some corpora are also *parsed*, which means they have been tagged with structural, syntactic information (e.g. clause structure, *wh–* traces).

Annotations are the traditional domain of conventional corpora, but they are not necessarily restricted to written texts. In the BNC, for example, the spoken section is tagged, while in ICE-GB, the spoken texts are parsed as well as tagged. Specialised dialect corpora have historically consisted of 'raw' orthographic transcription, with minimal mark-ups, if any. For researchers building specialised corpora, the transcriptions represent a tool for uncovering variation and the patterned constraints on heterogeneity; there is less emphasis on automated data extraction since the methods of variationist sociolinguistics often necessitate careful consideration and delimitation of both the variable context and the individual variants of a particular variable. However, some digital corpora do incorporate annotation. NECTE, for example, has POS-tagging, while the whole of the ONZE archive is automatically tagged with the CELEX information (orthography, phonology, morphology, syntactic word class, and frequency; see Baayen *et al*. 1995). Depending on the type of variable in question, this kind of mark-up can vastly facilitate data extraction. Consider the paradigmatic sociolinguistic variable, word-final unstressed *-ing* in words like *running* and *singing*. By performing a query across morphology (+ ing) and syntax (verb) in the Mobile Unit, the software designed for mining ONZE (ONZEminer, Fromont and Hay 2008[10]) is able to automatically search the archive and return the results, 3,499 tokens, in just under 7.5 minutes. These can then be exported directly to Excel. Figure 3.2 displays a sample of the results.

Table 3.3 *Sample POS tagset from the BNC (CLAWS, v.5)*

Tag	Denotation	
AJ0	Adjective, unmarked	tall, nice
AJC	Adjective, comparative	taller, nicer
AJS	Adjective, superlative	tallest, nicest
NN0	Noun, neutral for number	sheep, fish
NN1	Noun, singular	cat, tooth
NN2	Noun, plural	cats, teeth
NP0	Noun, proper	Canterbury, Elizabeth
PNP	Pronoun, personal	he, she
PNX	Pronoun, reflexive	himself, herself

3.4.4 Accessibility

Two primary motivations for making corpora public are scientific enquiry (the results are confirmable and replicable) and descriptive adequacy (the results can be tested against other data-sets). The decision to go public must be made at the outset. Explicit permissions must be sought from all data sources (written materials face copyright issues; spoken materials face ethics issues). Moreover, you need to think about how the corpus will be accessed by others. Will you establish a website (like that for NECTE or SCOTS) or will you distribute it through an established catalogue like the Oxford Text Archive or the LDC? If the latter, what are their protocols? If the former, then the question of how to sustain the project must be planned for from the beginning (Denbo *et al*. 2008: 1). For how long do you intend the digital resource to be available and maintained? Sustainability has a number of facets: the need to provide a host (e.g. a university research centre like the BlueFern computing services facility at the University of Canterbury or a national institute like the British Universities Film and Video Council), the need to update the technical format and the content, and the need for financial and technical support to maintain the digital resource. These issues are fundamental to the availability of any public corpus and comprise some of the most pressing concerns facing digital corpora today.

3.5 How you use a corpus informs what you find

The most important skill is to be able to ask insightful questions which address real issues and problems in theoretical, descriptive and applied language studies (Kennedy 1998: 3).

The questions we bring to a particular data-set and the way(s) in which we seek the answers ultimately inform the types of answers we find. There is the obvious disclaimer here: whatever we find in a corpus is only what that

No.	Transcript	Speaker	gender	YOB	region	Sync	URL	Text	Match transcript
2	AdaAitcheson-01.trs	Ada Aitcheson	F	1878	Coastal Otago	78.003	http://uccorpus/lay	- that ah was bought . when he was going th	going
3	AdaAitcheson-01.trs	Ada Aitcheson	F	1878	Coastal Otago	88.137	http://uccorpus/lay	- told the jeweller he was going to Australia I	going
4	AdaAitcheson-01.trs	Ada Aitcheson	F	1878	Coastal Otago	114.766	http://uccorpus/lay	we still have the clock . the clock is still goin	going
5	AdaAitcheson-01.trs	Ada Aitcheson	F	1878	Coastal Otago	126.473	http://uccorpus/lay	Still going very well . yes	going
6	AdaAitcheson-01.trs	Ada Aitcheson	F	1878	Coastal Otago	148.956	http://uccorpus/lay	er with land here grazing sheep I think	grazing
7	AdaAitcheson-01.trs	Ada Aitcheson	F	1878	Coastal Otago	156.506	http://uccorpus/lay	how pleased he was to think that there was	going
8	AdaAitcheson-01.trs	Ada Aitcheson	F	1878	Coastal Otago	198.697	http://uccorpus/lay	that house is still being used today	being
9	AdaAitcheson-01.trs	Ada Aitcheson	F	1878	Coastal Otago	203.474	http://uccorpus/lay	-- he er stayed here farming . on different s	farming .
10	AdaAitcheson-01.trs	Ada Aitcheson	F	1878	Coastal Otago	247.498	http://uccorpus/lay	each carrying . a child and luggage . I	carrying .
12	AdaAitcheson-02.trs	Ada Aitcheson	F	1878	Coastal Otago	26.209	http://uccorpus/lay	which was then working about half a mile a	working
13	AdaAitcheson-02.trs	Ada Aitcheson	F	1878	Coastal Otago	70.393	http://uccorpus/lay	and he had no difficulty whatsoever . in ah	borrowing
14	AdaAitcheson-02.trs	Ada Aitcheson	F	1878	Coastal Otago	70.393	http://uccorpus/lay	and he had no difficulty whatsoever . in ah	renting
15	AdaAitcheson-02.trs	Ada Aitcheson	F	1878	Coastal Otago	114.484	http://uccorpus/lay	but the one that I have heard my own fath	being .
16	AdaAitcheson-02.trs	Ada Aitcheson	F	1878	Coastal Otago	210.553	http://uccorpus/lay	ideal for . growing anything really	growing
17	AdaAitcheson-02.trs	Ada Aitcheson	F	1878	Coastal Otago	233.456	http://uccorpus/lay	. and one other family . one other branch of	living
18	AliceMackie-03.trs	Alice Mackie	F	1877	Coastal Otago	110.826	http://uccorpus/lay	farming	farming
19	AliceMackie-05.trs	Alice Mackie	F	1877	Coastal Otago	92.465	http://uccorpus/lay	was telling you about the two Chinamen in I	telling
20	AliceMackie-05.trs	Alice Mackie	F	1877	Coastal Otago	129.469	http://uccorpus/lay	and burn them f~ on the fire instead of buy	buying

Figure 3.2 Sample of Excel spreadsheet generated by ONZEminer, layered search for verbs with the suffix *-ing* in the Mobile Unit Archive (3,499 results, 446,446 ms)

corpus has managed to capture. As Meyer and Nelson (2006: 94) point out, no corpus is representative of the entire potential of a given language. Moreover, no single corpus can satisfy every potential user (Meyer 2004: 348). That is, no matter how carefully constructed a corpus is, certain design features will inhibit certain types of analyses. But there are also less obvious ways in which this generalisation can operate. For example, more and more researchers are turning to instant messaging and other forms of mediated communication for their data (e.g. Baron 2004; Ling and Baron 2007; Tagliamonte and Denis 2008). The advent of predictive text in these mediums restricts intra-speaker variability, a result that has inevitable consequences for inter-speaker variability. At the same time, the growing availability of public corpora has important descriptive ramifications. If we approach a corpus assuming that a particular feature does the same work there as it does in our native dialect, then the possibility of misinterpreting variability arises. For example, a colleague discussed a researcher who examined the use of the discourse marker *eh*. The analysis was framed using a North American model, a model that was ill suited to the southern hemisphere data being used. The end result was that pragmatic functions were wrongly ascribed to certain uses, invalidating the overall findings. This serves as a caution: when using corpora it is critical to divorce intuitions from interpretation and not allow assumptions regarding particular forms to obfuscate the local, context-dependent meaning of variation. In other words, the data themselves should inform our analyses. Interpretation matters.

Ultimately though 'we are limited in what we discover by what we set out to look for' (Cheshire 1999: 65). Cheshire raised this point in discussing the tendency in dialect research to investigate known variables (i.e. those already analysed). As Bauer (2002: 102) points out, 'replicability […] is a sign of good science'. At the same time, a fresh perspective on a 'known' entity can offer new insights and, borne of hypothesis testing, this too is a sign of good science. A case in point concerns discourse *like*, a ubiquitous feature of casual speech. Investigations of *like* have tended to focus on adolescents and young adults and they have concentrated on the contexts where *like* is used to the exclusion of those where it is not (e.g. Underhill 1988; Miller and Weinert 1995; Andersen 1997, 1998, 2001). The patterns uncovered in these studies were consistent, whether the data were British (Andersen 1997; COLT) or American (Underhill 1988; specialised, private), but they also appeared unsystematic. Consider the following examples from the Toronto English corpus (Tagliamonte 2006b). How would you summarise the distribution of *like*?

1 a. *Like* you have to *like* walk into their room just to see *like* the different *like* half.
  b. I don't really *like* judge people on what music they listen to.

c. A trade that I *like* really like was the one they had got from Jersey.
d. They were *like* so mad they decided to ground me for a week.
e. My mouth was getting incredibly *like* dry.

Data such as these have resulted in lists of possible combinations (e.g. *like* can appear before or within a noun phrase, at the beginning of a sentence) but no coherent theory of what made these combinations possible. Where did they come from? How did they emerge? A central tenet of the variationist paradigm is that variation is not only an inherent aspect of language but that it is structured (Weinreich, Labov and Herzog 1968). It is also unlikely that younger speakers simply 'made up' *like*; they had to have learned it from somebody.[11]

Rather than isolating younger speakers and actual occurrences of *like*, D'Arcy (2005a, 2007, 2008) considered all age groups and examined individual syntactic structures, whether they contained *like* or not. Among the insights provided by this perspective, it became apparent that *like*:

- is used by speakers of all ages (e.g. 45-year-olds and 15-year-olds are differentiated primarily by frequency, not contexts of use);
- is constrained by the syntax (e.g. it follows speaker and subject-oriented adverbs like *really* 'truly' (1b) but precedes degree and manner adverbs like *really* 'intensification' (1c));
- is constrained by semantic factors (e.g. it is probabilistically favoured with verbs that select an agentive subject, like *walk* (1a));
- has developed systematically (e.g. [*like* [DegP AP]] preceded [[DegP] *like* [AP]] (1d > 1e)).

In other words, what looks fairly random when considered one way looks strikingly structured when considered another way. The data remain the same, but the perspective from which they are examined can alter the way we interpret their meaning.

3.6 Summary

Whether conventional or dialectal, balanced or specialised, static or dynamic, big or small, spoken or written, corpora are the foundation upon which variation studies are moored. A corpus is the basis for linguistic analysis and description, capturing language as used by speakers and writers. It is thus revealing. But as researchers we must always remember that a corpus is also an imperfect construct: no corpus can capture all phenomena. And, for those it does capture, the questions we bring to bear will influence our interpretations. But in the end, corpora provide a window on the inherent variability of language and there is nothing more exhilarating than your first view of the results achieved in answer to your question, the issue that led you to the corpus in the first place.

3.7 Where next?

The classic primer in corpus linguistics is Sinclair (1991), while the contributions in Beal *et al.* (2007a) represent the state of the art on specialised corpora. Key readings in sociolinguistic data collection are Labov (1972c), Sankoff and Sankoff (1973), and Milroy (1987), and more currently, Milroy and Gordon (2003) and Tagliamonte (2006a). Poplack (1989) is foundational for issues surrounding sociolinguistic corpus construction and data handling; for careful discussion of text-based corpus construction, see Meyer (2002). On representativeness in data sampling and corpus construction see Sankoff (2005) and Biber (1993). A good starting point is Francis and Kučera (1964), the companion to the *Brown Corpus*, which established the model for subsequent corpora projects. On annotation in text corpora, see Leech (1993a). For discussion of the issues involved in representing speech in writing, see Ochs (1979), Macaulay (1991a), and Tagliamonte (2007). Kennedy (1998) provides a history of English corpus linguistics and a summary of key research in the field. A valuable resource for those interested in the burgeoning field of web-based corpus studies is Hundt *et al.* (2007). Biber *et al.* (1999) is an excellent reference grammar based on corpora representing British and American English; it is a good place to start when looking for possible project ideas. Online, David Lee's *Bookmarks for Corpus-Based Linguists* (Lee 2001) is an invaluable resource for all corpus-related issues.

Websites

Site	http(s):// (note: all these sites were last accessed 4 August 2010)
ANC	americannationalcorpus.org/
Bookmarks for Corpus-Based Linguists	personal.cityu.edu.hk/~davidlee/devotedtocorpora/CBLLinks.htm
BNC	www.natcorp.ox.ac.uk/
Buckey Corpus	buckeyecorpus.osu.edu/
Dialect Topography	dialect.topography.chass.utoronto.ca/
GlossaNet	glossa.fltr.ucl.ac.be/
ICAME	icame.uib.no/
ICE	ice-corpora.net/ice/
LDC	www.ldc.upenn.edu/ For a guest account: online.ldc.upenn.edu/login.html To access Brown: secure.ldc.upenn.edu/intranet/
LAEME	www.lel.ed.ac.uk/ihd/laeme1/laeme1.html

LAOS	www.lel.ed.ac.uk/ihd/laos1/laos1.html
Linguistic Atlas Projects	us.english.uga.edu/
MICASE	micase.elicorpora.info/
NECTE	www.ncl.ac.uk/necte/
Oxford Text Archive	ota.ahds.ac.uk/
SCOTS	www.scottishcorpus.ac.uk/
TalkBank	talkbank.org/
	For SBCSAE talkbank.org/data/local.html
TEI	www.tei-c.org/index.xml
Transcriber	trans.sourceforge.net/en/presentation.php
WebCorp	www.webcorp.org.uk

4 Hypothesis generation

Hermann Moisl

4.1 Introduction

The aim of science is to understand reality. An academic discipline, philosophy of science, is devoted to explicating the nature of science and its relationship to reality, and, perhaps predictably, both are controversial; for an excellent introduction to the issues see Chalmers (1999). In practice, however, most scientists explicitly or implicitly assume a view of scientific methodology based on the philosophy of Karl Popper (Popper 1959, 1963), in which one or more non-contradictory hypotheses about some domain of interest are stated, the validity of the hypotheses is tested by observation of the domain, and the hypotheses are either confirmed (but not proven) if they are compatible with observation, or rejected if they are not.

Where do such hypotheses come from? In principle, it doesn't matter, because the validity of the claims they make can always be assessed with reference to the observable state of the world. Any one of us, whatever our background, could wake up in the middle of the night with an utterly novel and brilliant hypothesis that, say, unifies quantum mechanics and Einsteinian relativity, but this kind of inspiration is highly unlikely and must be exceedingly rare. In practice, scientists develop hypotheses in something like the following sequence of steps: the researcher (i) selects some aspect of reality that s/he wants to understand, (ii) becomes familiar with the selected research domain by observation of it, reads the associated research literature, and formulates a research question which, if convincingly answered, will enhance scientific understanding of the domain, (iii) abstracts data from the domain and draws inferences from it in the light of the research literature, and (iv) on the basis of these inferences states a hypothesis to answer the research question. The hypothesis is subsequently tested for validity with reference to the domain and amended as required.

Linguistics is a science, and as such uses or should use scientific methodology. The research domain is human language, and, in the process of hypothesis generation, the data comes from observation of language use. Such observation can be based on introspection, since every native speaker is an expert on the

usage of his or her language. It can also be based on observation of the linguistic usage of others in either spoken or written form. In some sub-disciplines like historical linguistics, sociolinguistics, and dialectology, the latter is in fact the only possible alternative, and this is why D'Arcy (this volume) stresses the importance of linguistic corpora in language variation research: corpora are 'the foundation of everything we do'.

Traditionally, hypothesis generation based on linguistic corpora has involved the researcher listening to or reading through a corpus, often repeatedly, noting features of interest, and then formulating a hypothesis. The advent of information technology in general and of digital representation of text in particular in the past few decades has made this often-onerous process much easier via a range of computational tools, but, as the amount of digitally represented language available to linguists has grown, a new problem has emerged: data overload. Actual and potential language corpora are growing ever larger, and even now they can be on the limit of what the individual researcher can work through efficiently in the traditional way. Moreover, as we shall see, data abstracted from such large corpora can be impenetrable to understanding. One approach to the problem is to deal only with corpora of tractable size, or, equivalently, with tractable subsets of large corpora, but ignoring potential data in such an unprincipled way is not scientifically respectable. The alternative is to use mathematically based computational tools for data exploration, as developed in the physical and social sciences, where data overload has long been a problem. This latter alternative is the one explored here. Specifically, the discussion shows how a particular type of computational tool, cluster analysis, can be used in the formulation of hypotheses in corpus-based linguistic research.

The discussion is in three main parts. The first describes data abstraction from corpora, the second outlines the principles of cluster analysis, and the third shows how the results of cluster analysis can be used in the formulation of hypotheses. Examples are based on the *Newcastle Electronic Corpus of Tyneside English* (NECTE), a corpus of dialect speech (Allen *et al*. 2007). The overall approach is introductory, and as such the aim has been to make the material accessible to as broad a readership as possible.

4.2 Data creation

'Data' comes from the Latin verb 'to give' and means 'things that are given'. Data are therefore things to be accepted at face value, true statements about the world. What is a true statement about the world? That question has been debated in philosophical metaphysics since antiquity and probably before (Flew and Priest 2002; Bunnin and Yu 2009; Zalta 2009), and, in our own time, has been intensively studied by the disciplines that comprise cognitive science (for example, Thagard 2005). The issues are complex, controversy abounds,

and the associated academic literatures are vast – saying what a true statement about the world might be is anything but straightforward. We can't go into all this, and so will adopt the attitude prevalent in most areas of science: data are abstractions of what we observe using our senses, often with the aid of instruments (Chalmers 1999).

Data are ontologically different from the world. The world is as it is; data are an interpretation of it for the purpose of scientific study. The weather is not the meteorologist's data – measurements of such things as air temperature are. A text corpus is not the linguist's data – measurements of such things as average sentence length are. Data are constructed from observation of things in the world, and the process of construction raises a range of issues that determine the amenability of the data to analysis and the interpretability of the analytical results. The importance of understanding such data issues in cluster analysis can hardly be overstated. On the one hand, nothing can be discovered that is beyond the limits of the data itself. On the other, failure to understand relevant characteristics of data can lead to results and interpretations that are distorted or even worthless. For these reasons, a detailed account of data issues is given before moving on to discussion of analytical methods.

4.2.1 Formulation of a research question

In general, any aspect of the world can be described in an arbitrary number of ways and to arbitrary degrees of precision. The implications of this go straight to the heart of the debate on the nature of science and scientific theories, but to avoid being drawn into that debate, this discussion adopts the position that is pretty much standard in scientific practice: the view, based on Karl Popper's philosophy of science (Popper 1959, 1963; Chalmers 1999), that there is no theory-free observation of the world. In essence, this means that there is no such thing as objective observation in science. Entities in a domain of enquiry only become relevant to observation in terms of a hypothesis framed using the ontology and axioms of a theory about the domain. For example, in linguistic analysis, variables are selected in terms of the discipline of linguistics broadly defined, which includes the division into sub-disciplines such as sociolinguistics and dialectology, the subcategorisation within sub-disciplines such as phonetics through syntax to semantics and pragmatics in formal grammar, and theoretical entities within each subcategory such as phonemes in phonology and constituency structures in syntax. Claims, occasionally seen, that the variables used to describe a corpus are 'theoretically neutral' are naïve: even word categories like 'noun' and 'verb' are interpretative constructs that imply a certain view of how language works, and they only appear to be theory-neutral because of familiarity with long-established tradition.

Figure 4.1 The NECTE dialect area

Data can, therefore, only be created in relation to a research question that is defined on the domain of interest, and that thereby provides an interpretative orientation. Without such an orientation, how does one know what to observe, what is important, and what is not?

The domain of interest in the present case is the *Newcastle Electronic Corpus of Tyneside English* (NECTE), a corpus of dialect speech interviews from Tyneside in north-east England (see Figure 4.1) (Allen *et al*. 2007).[1]

Moisl *et al*. (2006) and Moisl and Maguire (2008) began the study of the NECTE corpus with the aim of generating hypotheses about phonetic variation among speakers in the Tyneside dialect area using cluster analysis. The

research question asked in that work, and which serves as the basis for what follows here, is:

> Is there systematic phonetic variation in the Tyneside speech community, and, if so, what are the main phonetic determinants of that variation?

These studies went on to correlate the findings with social data about the speakers, but the present discussion does not engage with that.

4.2.2 Variable selection

Given that data are an interpretation of some domain of interest, what does such an interpretation look like? It is a description of entities in the domain in terms of variables. A variable is a symbol, and as such is a physical entity with a conventional semantics, where a conventional semantics is understood as one in which the designation of a physical thing as a symbol together with the connection between the symbol and what it represents are determined by agreement within a community. The symbol 'A', for example, represents the phoneme /a/ by common assent, not because there is any necessary connection between it and what it represents. Since each variable has a conventional semantics, the set of variables chosen to describe entities constitutes the template in terms of which the domain is interpreted. Selection of appropriate variables is, therefore, crucial to the success of any data analysis.

Which variables are appropriate in any given case? That depends on the nature of the research question. The fundamental principle in variable selection is that the variables must describe all and only those aspects of the domain that are relevant to the research question. In general, this is an unattainable ideal. Any domain can be described by an essentially arbitrary number of finite sets of variables; selection of one particular set can only be done on the basis of personal knowledge of the domain and of the body of scientific theory associated with it, tempered by personal discretion. In other words, there is no algorithm for choosing an optimally relevant set of variables for a research question.

Which variables are suitable to describe the NECTE speakers? In principle, when setting out to perform a classification of a speech corpus, the first step is to partition each speaker's analogue speech signal into a sequence of discrete phonetic segments and to represent those segments symbolically, or, in other words, to transcribe the audio interviews. To do this, one has to decide which features of the audio signal are of interest, and then to define a set of variables to represent those features. These decisions were made long ago with respect to the NECTE interviews.

NECTE is based on two pre-existing corpora, one of them collected in the late 1960s by the *Tyneside Linguistic Survey* (TLS) project (Strang 1968;

OU		PDV	(code)	states	lexical examples
	NL				
I	i:	i:	0002	i i̲ i̲ i̲ ɨ i̲	week, treat, see
		I	0004	i̲ i̲ i̲ i̲ ɨ	week, relief
		ɛ	0006	e̲ ɛ̲ e̲ ɛ̲	beat
		eI	0008	e̲ i ∂ɨ e̲ ɛ̲ i̲	see
		Iə	0010	i ɛ i ɛ i ∂	feed
		Ii	0012	ii(back) ii(low) i̲	we see

Figure 4.2 Extract from the TLS transcription scheme

Pellowe *et al*. 1972), and the other in 1994 by the *Phonological Variation and Change in Contemporary Spoken English* (PVC) project (Milroy *et al*. 1997). For present purposes we are interested in the sixty-three interviews that comprise the TLS component of NECTE, and it happens that the TLS researchers had already created phonetic transcriptions of at least part of each interview. This saved the NECTE project the arduous labour of transcription, but at the same time bound us to their decisions about which phonetic features are of interest, and how they should be symbolically represented as variables. Details of the TLS transcription scheme are available in (Allen *et al*. 2007) as well as at the NECTE website;[2] a short excerpt from the TLS transcription scheme is given in Figure 4.2.

Two levels of transcription were produced, a highly detailed narrow one designated 'states' in Figure 4.2, and a superordinate 'Putative Diasystemic Variables' (PDV) level which collapsed some of the finer distinctions transcribed at the 'states' level. We shall be dealing with the less detailed PDV level.

4.2.3 Variable value assignment

The semantics of each variable determines a particular interpretation of the domain of interest, and the domain is 'measured' in terms of the semantics. That measurement constitutes the values of the variables: height in metres = 1.71, weight in kilograms = 70, and so on. Measurement is fundamental in the creation of data because it makes the link between data and the world, and thus allows the results of data analysis to be applied to the understanding of the world.

Measurement is only possible in terms of some scale. There are various types of measurement scale, and these are discussed at length in, for example, any statistics textbook, but for present purposes the main dichotomy is

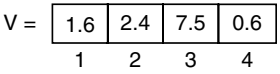


Figure 4.3 A numerical vector

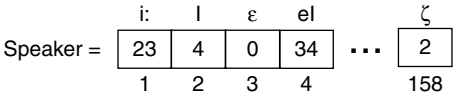


Figure 4.4 A NECTE data vector

between numeric and non-numeric. Cluster analysis methods assume numeric measurement as the default case, and for that reason the same is assumed in what follows. Specifically, we shall be interested in the number of times each speaker uses each of the NECTE phonetic variables. The speakers are therefore 'measured' in terms of the frequency with which they use these segments.

4.2.4 Data representation

If they are to be analysed using mathematically based computational methods, the descriptions of the entities in the domain of interest in terms of the selected variables must be mathematically represented. A widely used way of doing this, and the one adopted here, is to use structures from a branch of mathematics known as linear algebra. There are numerous textbooks and websites devoted to linear algebra; a small selection of introductory textbooks is Anton (2005), Poole (2005), and Blyth and Robertson (2002).

Vectors are fundamental in data representation. A vector is just a sequence of numbered slots containing numerical values. Figure 4.3 shows a four-element vector, each element of which contains a real-valued number: 1.6 is the value of the first element v_1, 2.4 the value of the second element v_2, and so on.

A single NECTE speaker's frequency of usage of the 158 phonetic segments in the transcription scheme can be represented by a 158-element vector in which each element is associated with a different segment, as in Figure 4.4. This speaker uses the segment at $Speaker_1$ twenty-three times, the segment at $Speaker_2$ four times, and so on.

The sixty-three speaker vectors can be assembled into a matrix M, shown in Figure 4.5, in which the 63 rows represent the speakers, the 158 columns represent the phonetic segments, and the value at M_{ij} is the number of times speaker i uses segment j (for $i = 1\ldots63$ and $j = 1\ldots158$).

This matrix M is the basis of subsequent analysis.

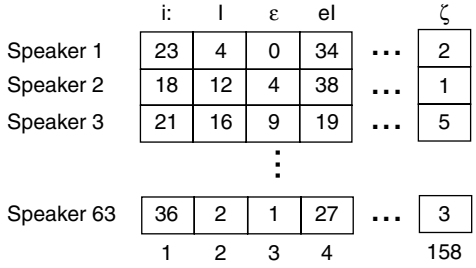


Figure 4.5 A fragment of the NECTE data matrix M

4.3 Data analysis

Once the data matrix has been created, a variety of computational methods can be used to classify its row vectors, and thereby the objects in the domain that the row vectors represent. In the present case, those objects are the NECTE speakers. The discussion is in three main parts:

- Part 1 motivates the use of computational methods for clustering.
- Part 2 introduces a fundamental concept: vector space.
- Part 3 describes how clusters can be found in vector space.

All three parts of the discussion are based on the NECTE data matrix M developed in the preceding section.

4.3.1 Motivation

We have seen that creation of data for study of a domain requires description of the objects in the domain in terms of variables. One might choose to observe only one aspect – the height of individuals in a population, say – in which case the data consists of more or less numerous values assigned to one variable; such data is univariate. If two values are observed – say height and weight – then the data is bivariate, if three trivariate, and so on up to some arbitrary number n; any data where n is greater than 1 is multivariate.

As the number of variables grows, so does the difficulty of classifying the objects that the data matrix rows represent by direct inspection. Consider, for example, Figure 4.6, which shows a matrix describing nine people in terms of a single variable *Age*. It's easy enough to classify these people into three groups: young (1–3), middle-aged (4–6), and old (7–9) just by looking at the matrix.

If one adds a second variable *Weight*, as in Figure 4.7, classification based on direct examination of the matrix is a little more difficult.

	Age
Person 1	14
Person 2	12
Person 3	15
Person 4	41
Person 5	47
Person 6	43
Person 7	83
Person 8	76
Person 9	81

Figure 4.6 Univariate data

The groups are the same as before, and there is a correlation between age and weight: the young group weighs least, the middle-aged group weighs most, and the old group weighs a little less than the middle-aged one.

Now increase the number of variables to, say, six, as in Figure 4.8. One can spend a long time looking at these numbers without coming up with a coherent grouping. And what if the number of variables is increased even more to, say, the 158 variables of the NECTE data matrix M? That matrix is too large to be shown here in its entirety, so only a dozen variables are given for nine of the speakers in Figure 4.9, but even this is sufficient to make the required point. Group these speakers on the basis of this phonetic segment frequency data. Difficult? Impossible? Try all 158 variables, and classify not just 9 but 63 speakers.

In general, as the number of variables grows, so does the difficulty of understanding the data, that is, of conceptualising the interrelationships of variables within a single data item on the one hand, and the interrelationships of complete data items on the other. The moral is straightforward: human cognitive make-up is unsuited to seeing regularities in anything but the smallest

	Age	Weight (kg)
Person 1	14	25
Person 2	12	21
Person 3	15	26
Person 4	41	83
Person 5	47	82
Person 6	43	80
Person 7	83	71
Person 8	76	73
Person 9	81	72

Figure 4.7 Bivariate data

	Age	Weight (kg)	Height (m)	Size of family	Years worked	Trips abroad
Person 1	14	25	1.4	5	2	2
Person 2	12	21	1.36	5	0	0
Person 3	15	26	1.5	4	1	1
Person 4	41	83	1.74	7	15	46
Person 5	47	82	1.72	3	17	23
Person 6	43	80	1.66	6	21	0
Person 7	83	71	1.65	2	36	12
Person 8	76	73	1.68	5	34	29
Person 9	81	72	1.81	4	42	0

Figure 4.8 Multivariate data

	$d_{initial}$	ɛɻ	n	$b_{initial}$	aI	$K_{initial}$	ɪ	a	k_{final}	tʃ	æ	p_{medial}
Speaker 1	22	19	177	39	6	44	13	11	47	10	37	8
Speaker 2	27	6	210	32	9	45	18	8	40	17	46	6
Speaker 3	32	16	188	57	8	27	23	6	29	6	42	6
Speaker 4	33	20	191	45	6	47	21	16	40	3	42	7
Speaker 5	43	27	304	58	13	53	28	12	74	14	76	10
Speaker 6	34	9	202	54	14	26	14	14	45	5	53	6
Speaker 7	33	0	222	27	54	47	27	11	40	16	51	18
Speaker 8	22	16	186	41	3	56	19	10	29	8	53	8
Speaker 9	30	27	214	54	12	29	20	6	45	7	54	8

Figure 4.9 Multivariate NECTE data

collections of numerical data. To see the regularities we need graphical aids, and that is what clustering methods provide.

4.3.2 Vector space

Although it is just a sequence of numbers, a vector can be geometrically interpreted (Blyth and Robertson 2002; Anton 2005; Poole 2005). To see how, take a vector consisting of two elements, say $v = (30,70)$. Under a geometrical interpretation, the two elements of v define a two-dimensional space, the numbers at $v_1 = 30$ and $v_2 = 70$ are co-ordinates in that space, and the vector v itself is a point at the co-ordinates (30,70), as shown in Figure 4.10.

A vector consisting of three elements, say $v = (40, 20, 60)$ defines a three-dimensional space in which the co-ordinates of the point v are 40 along the horizontal axis, 20 along the vertical axis, and 60 along the third axis shown in perspective, as in Figure 4.11.

A vector $v = (22, 38, 52, 12)$ defines a four-dimensional space with a point at the stated co-ordinates, and so on to any dimensionality n. Vector spaces of dimensionality greater than three are impossible to visualise directly and are therefore counterintuitive, but mathematically there is no problem with them; two- and three-dimensional spaces are useful as a metaphor for conceptualising higher-dimensional ones.

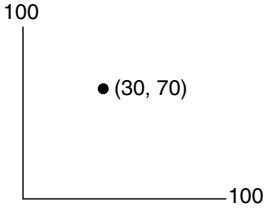


Figure 4.10 A vector in two-dimensional space

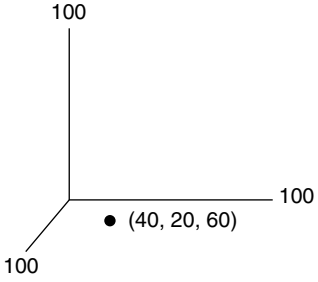


Figure 4.11 A vector in three-dimensional space

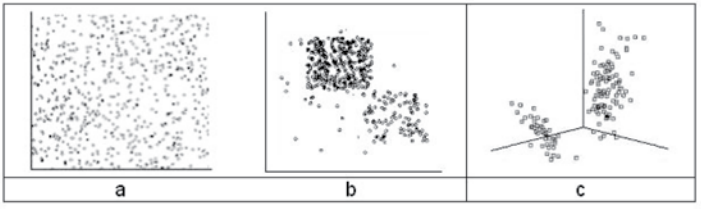


Figure 4.12 Multiple vectors in two- and three-dimensional spaces

When numerous vectors exist in a space, it may or may not be possible to see interesting structure in the way they are arranged in it. Figure 4.12 shows vectors in two- and three-dimensional spaces. In (a) they were randomly generated and there is no structure to be observed, in (b) there are two clearly defined concentrations in two-dimensional space, and in (c) there are two clearly defined concentrations in three-dimensional space.

The existence of concentrations like those in (b) and (c) indicate relationships among the entities that the vectors represent. In (b), for example, if the horizontal axis measures weight and the vertical one height for a sample human population, then members of the sample fall into two groups: tall, light people on the one hand, and short, heavy ones on the other.

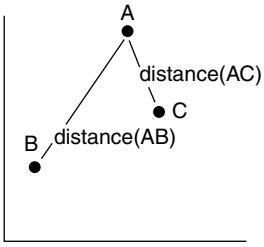


Figure 4.13 Distance between vectors in two-dimensional space

This idea of identifying clusters of vectors in vector space and interpreting them in terms of what the vectors represent is the basis of cluster analysis. In what follows, we shall be attempting to group the NECTE speakers on the basis of their phonetic usage by looking for clusters in the arrangement of the row vectors of M in 158-dimensional space.

4.3.3 Cluster analysis

Where the vectors are two or three dimensional they can simply be plotted and any clusters will be visually identifiable, as we have just seen. But what about when the vector dimensionality is greater than 3 – say 4, or 10, or 100? In such a case direct plotting is not an option. How exactly would one draw a six-dimensional space, for example? Many data matrix row vectors have dimensionalities greater than 3 – the NECTE matrix M has dimensionality 158 – and, to identify clusters in such high-dimensional spaces, some procedure more general than direct plotting is required. A variety of such procedures is available, and they are generically known as cluster analysis methods. This section looks at these methods.

The literature on cluster analysis is extensive. A few recent books are Everitt *et al*. (2001) and Kaufman and Rousseeuw (2005), but many textbooks in fields like multivariate statistical analysis, information retrieval, and data mining also contain useful and accessible discussions, and there are numerous relevant and often excellent websites.

The discussion of cluster analysis is in four parts. The first introduces distance in vector space, the second describes one particular class of clustering methods, the third applies that type of method to the NECTE data matrix M, and the fourth interprets the result of the NECTE analysis.

4.3.3.1 *Distance in vector space* Where there are two or more vectors in a space, it is possible to measure the distance between any two of them and to rank them in terms of their proximity to one another. Figure 4.13 shows

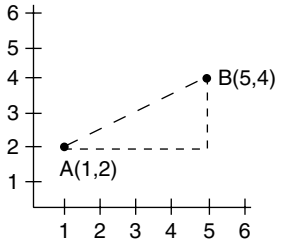


Figure 4.14 Euclidean distance calculation: 'In a right-angled triangle, the square of the length of the hypotenuse is equal to the sum of the squares of the lengths of the other two sides'

a simple case of a two-dimensional space in which the distance from vector A to vector B is greater than the distance from A to C.

There are various ways of measuring such distances, but the most often used is the familiar Euclidean one (Figure 4.14):

$$dist(AB) = \sqrt{(5-1)^2} + (4-2)^2$$

4.3.3.2 *Cluster analysis methods* Cluster analysis methods use relative distance among vectors in a space to group the vectors into clusters. Specifically, for a given set of vectors in a space, they first calculate the distances between all pairs of vectors, and then group into clusters all the vectors that are relatively close to one another in the space and relatively far from those in other clusters. 'Relatively close' and 'relatively far' are, of course, vague expressions, but they are precisely defined by the various clustering methods, and for present purposes we can avoid the technicalities and rely on intuitions about relative distance.

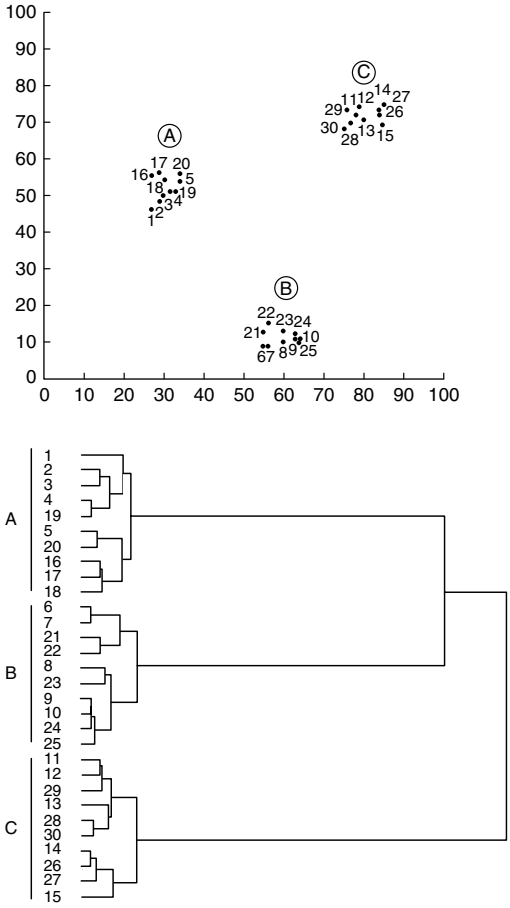
For concreteness, we will concentrate on one particular class of methods: hierarchical cluster analysis, which represents the relativities of distance among vectors as a tree. Figure 4.15 exemplifies this.

Column (a) shows a 30×2 data matrix that is to be cluster analysed. Because the data space is two-dimensional the vectors can be directly plotted to show the cluster structure, as in the upper part of column (b). The corresponding hierarchical cluster tree is shown in the lower part of column (b). Linguists use such trees as representations of sentence phrase structure, but cluster trees differ from linguistic ones in the following respects:

- The leaves are not lexical tokens but labels for the data items – the numbers at the leaves correspond to the numerical labels of the row vectors in the data matrix.

	v1	v2
1	27	46
2	29	48
3	30	50
4	32	51
5	34	54
6	55	9
7	56	9
8	60	10
9	63	11
10	64	11
11	78	72
12	79	74
13	80	70
14	84	73
15	85	69
16	27	55
17	29	56
18	30	54
19	33	51
20	34	56
21	55	13
22	56	15
23	60	13
24	63	12
25	64	10
26	84	72
27	85	74
28	77	70
29	76	73
30	76	69

a



b

Figure 4.15 Hierarchical cluster analysis of two-dimensional data

- They represent not grammatical constituency but relativities of distance between clusters. The lengths of the branches linking the clusters represent degrees of closeness: the shorter the branch, the more similar the clusters. In cluster A vectors 4 and 19 are very close and thus linked with very short lines; 2 and 3 are almost but not quite as close as 4 and 19, and are therefore linked with slightly longer lines, and so on.

Knowing this, the tree can be interpreted as follows. There are three clusters labelled A, B, and C, in each of which the distances among vectors are quite small. These three clusters are relatively far from one another, though A and B are closer to one another than either of them is to C. Comparison with the vector plot shows that the hierarchical analysis accurately represents the distance relations among the thirty vectors in two-dimensional space.

Given that the tree tells us nothing more than what the plot tells us, what is gained? In the present case, nothing. The real power of hierarchical analysis lies in its independence of vector space dimensionality. We have seen that direct plotting is limited to three or fewer dimensions, but there is no dimensionality limit on hierarchical analysis – it can determine relative distances in vector spaces of any dimensionality and represent those distance relativities as a tree like the one above. To exemplify this, the 158-dimensional NECTE data matrix M was hierarchically cluster analysed, and the results of the analysis are shown in the next section.

4.3.3.3 *Hierarchical cluster analysis of the NECTE data* Recall that the NECTE data is a 63×158 matrix M in which each of the 63 rows represents a speaker, each of the columns represents a phonetic segment, and the value at M_{ij} is the number of times speaker i uses phonetic segment j. Each row vector is therefore a phonetic profile of a different NECTE speaker; the aim is to classify the speakers in terms of the similarity of their phonetic profiles or, put another way, in terms of the relative distances among the row vectors in the 158-dimensional space. The resulting tree is shown in Figure 4.16.

Plotting M in 158-dimensional space would have been impossible, and, without cluster analysis, one would have been left pondering a very large and incomprehensible matrix of numbers. With the aid of cluster analysis, however, structure in the data is clearly visible: there are two main clusters, NG1 and NG2; NG1 consists of large subclusters NG1a and NG1b; NG1a itself has two main subclusters NG1a(i) and NG1a(ii).

4.4 Hypothesis generation

Given that there is structure in the relative distances of the row vectors of M, what does that structure mean in terms of the research question?

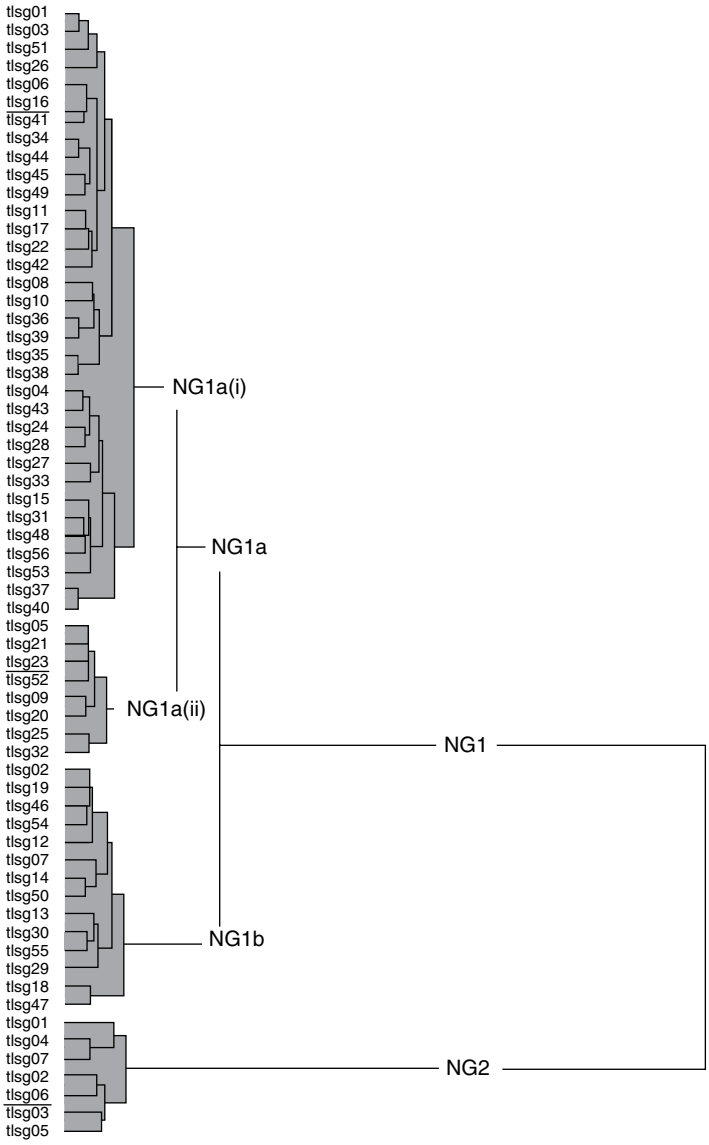


Figure 4.16 Hierarchical cluster analysis of the NECTE data matrix M

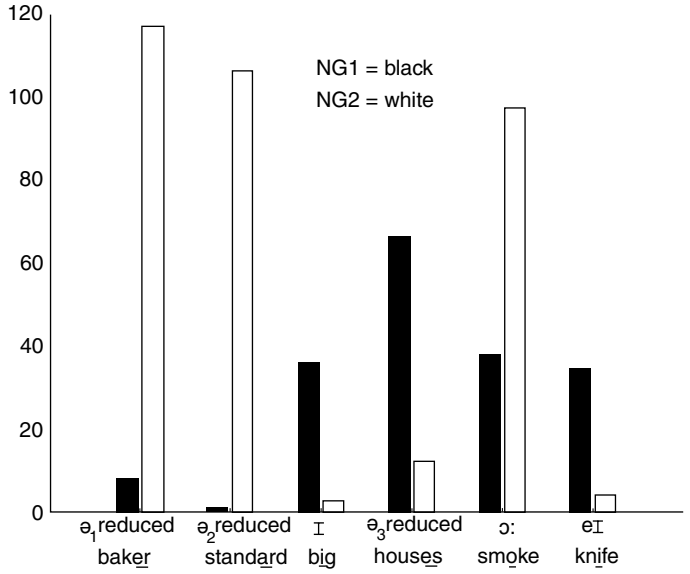


Figure 4.17 Co-plot of centroids for NG1 and NG2

Is there systematic phonetic variation in the Tyneside speech community, and, if so, what are the main phonetic determinants of that variation?

Because the row vectors of M are phonetic profiles of the NECTE speakers, the cluster structure means that the speakers fall into clearly defined groups with specific interrelationships rather than, say, being randomly distributed around the phonetic space. A reasonable hypothesis to answer the first part of the research question, therefore, is that there is systematic variation in the Tyneside speech community. This hypothesis can be refined by examining the social data relating to the NECTE speakers, which shows, for example, that all those in the NG1 cluster come from the Gateshead area on the south side of the river Tyne and all those in NG2 come from Newcastle on the north side, and that the sub-clusters in NG1 group the Gateshead speakers by gender and occupation.

The cluster tree can also be used to generate a hypothesis in answer to the second part of the research question. So far we know *that* the NECTE speakers fall into clearly demarcated groups on the basis of variation in their phonetic usage. We do not, however, know *why*, that is, which segments out of the 158 in the TLS transcription scheme are the main determinants of this regularity. To identify these segments, we begin by looking at the two main clusters NG1 and NG2 to see which segments are most important in distinguishing them (Moisl and Maguire 2008).

The first step is to create for the NG1 cluster a vector that captures the general phonetic characteristics of the speakers it contains, and to do the same for the NG2. Such vectors can be created by averaging all the row vectors in a cluster using the formula

$$v_j = \frac{\sum_{i=1..m} M_{ij}}{m}$$

where v_j is the jth element of the average or 'centroid' vector v (for $j = 1 \ldots$ the number of columns in M), M is the data matrix, Σ designates summation, and m is the number of row vectors in the cluster in question (56 for NG1, 7 for NG2). This yields two centroid vectors.

Next, compare the two centroid vectors by co-plotting them to show graphically how, on average, the two speaker groups differ on each of the 158 phonetic segments; a plot of all 158 segments is too dense to be readily deciphered, so the 6 on which the NG1 and NG2 centroids differ most are shown in Figure 4.17.

The six phonetic segments most important in distinguishing cluster NG1 from NG2 are three varieties of [ə], [ɔː], [ɪ], and [eɪ]: the Newcastle speakers characteristically use $ə_1$ and $ə_2$ whereas the Gateshead speakers use them hardly at all, the Gateshead speakers use $ə_3$ much more than the Newcastle speakers, and so on. A hypothesis that answers the second part of the research question is therefore that the main determinants of phonetic variation in the Tyneside speech community are three kinds of [ə], [ɔː], [ɪ], and [eɪ]. The subclusters of NG1 can be examined in the same way and the hypothesis thereby further refined.

Having formulated two hypotheses about Tyneside speech, they need to be tested against additional evidence from a source or sources other than NECTE and amended or even discarded if that is what the evidence requires.

4.5 Summary

This discussion set out to show how one type of computational analytical tool, cluster analysis, can be used to generate hypotheses about large digital linguistic corpora when the data abstracted from them is too complex to be interpreted by direct inspection. This approach to hypothesis generation is useful primarily when dealing with corpora in languages that have been relatively little studied, such as endangered languages, but even for intensively studied ones like English, where hypotheses can usually be generated from the existing research literature, cluster analysis can produce surprises, as Moisl and Maguire (2008) showed for Tyneside English.

4.6 Where next?

The foregoing discussion was introductory, and anyone wishing to use cluster analysis in actual research applications has some additional reading to do. There is no shortage of such reading: the literature on cluster analysis, both in traditional printed form and on the Web, is extensive. Much of it is, however, quite technical, and this can be an obstacle to those new to the subject. It's important to have a secure intuitive grasp of the underlying concepts before trying to assimilate the technicalities, so a good way into the literature is to start with the Web, using 'cluster analysis' as the search string. There are numerous good and even excellent introductory-level cluster analysis websites, and working through these lays the groundwork for more advanced reading. Romesburg (1984) is an accessible first textbook, followed by Everitt *et al.* (2001); the latter contains an extensive bibliography for further reading.

Knowing the theory of cluster analysis is a necessary but not sufficient condition for using it in research. Software is required to do the actual work. The standard statistics packages available in university and other research environments include a few types of clustering method, but more specialised ones provide a greater range of methods and, generally, better output graphics; a Web search using the string 'cluster analysis software' gives a good overview of what is available. Also very useful are Web directories of cluster analysis and related resources such as Fionn Murtagh's *Multivariate Data Analysis Software and Resources Page* (http://astro.u-strasbg.fr/~fmurtagh/mda-sw/).

The data to be cluster analysed may contain characteristics that can distort the result or even render it invalid as a basis for hypothesis generation. These characteristics, which include variation in the lengths of documents in multi-document corpora, data sparsity, and nonlinearity, must be recognised and where necessary eliminated or at least mitigated prior to undertaking the analysis. Given its importance, the research literature contains surprisingly little on such matters; see Pyle (1999) and Moisl (2007, 2008, 2010).

Finally, anyone proposing to use cluster analysis has to face the reality that, to do so respectably, knowledge of the basics of linear algebra and of statistics is a prerequisite. Some introductory textbooks on linear algebra are Anton (2005), Blyth and Robertson (2002), and Poole (2005); introductory statistics textbooks are too numerous to require individual mention, and are available in any research library as well as on the Web.

Cluster analysis has long been and continues to be a standard data processing tool across a broad range of physical and social sciences. The advent of digital electronic text in the second half of the twentieth century has driven the emergence of research disciplines devoted to search and interpretation of large digital natural language document collections, among them Information

Retrieval (Manning *et al*. 2008), Data Mining (Hand *et al*. 2001), Computational Linguistics (Mitkov 2005), and Natural Language Processing (Manning and Schütze 1999), and here too cluster analysis is a standard tool. As increasingly large digital collections become available for research into linguistic variation, traditional analytical methods will become intractable, and use of the computational tools developed by these text processing disciplines, including cluster analysis, will become the only realistic option.

5 Quantifying relations between dialects

Warren Maguire and April McMahon

5.1 Why we might want to know how similar or different varieties are to each other

What are the historical origins of Standard English, and how might we find out? Are regional varieties of English getting more similar to each other as a result of increased geographical and social mobility? Is it true that varieties of English spoken in former Celtic-speaking areas such as Cornwall, Wales, and the Highlands of Scotland are more similar to Standard English than other varieties? How similar are Scots and English, and are they more or less different than, for example, German and Dutch?

These, and many others like them, are questions which we, as linguists and dialectologists, would like to be able to answer. They all involve determining the relationships between varieties of a language (or indeed between different languages) in a way which goes beyond looking at individual features. In determining and even measuring the similarities and differences between dialects, we can begin to answer questions about their status and history, and the connections between these and the society, history, and geography of the people who speak them. This chapter is about how we determine the degree of relatedness between linguistic varieties – of course, varieties of the same language are related, and are similar, so we need to get beyond yes–no questions and instead figure out how close the relationships between them are. In §5.2, we examine traditional approaches to the issue, specifically the use of isoglosses for determining dialect boundaries. In §5.3, we discuss some problems with this kind of approach, and in §5.4 we give an overview of alternative methods which are designed to avoid the problems inherent in the isogloss approach. In §5.5, we turn to methods for representing these relationships, specifically in the form of maps, trees, and networks. In §5.6, we briefly examine how we validate the results of these kinds of approaches, and identify some key reading in §5.7.

5.2 Shared features

It is an obvious fact that varieties of the same language share many features but differ in others. For example, speakers of English in Scotland and in England

distinguish between the vowel in *cut* and the vowel in *coat*, but speakers of English in Scotland often do not distinguish between the vowels in *cot* and *caught*, whilst speakers of English in England do. When varieties share a particular feature, this may be for a number of reasons: retention of a feature which was found in the variety ancestral to both; shared innovation in the two varieties, perhaps because of geographical proximity; contact between the two varieties at some stage in the past; or perhaps simply chance – the two varieties may have independently developed the same feature. The more complex or unusual a shared innovation is, the less likely that varieties share it by accident. Thus the complex Scottish Vowel Length Rule is found both in Scotland and in northern parts of Ireland, a situation which speaks of the close historical links between these two regions (Harris 1984a).

It is also fairly obvious that these kinds of differences add up. As well as pronouncing *cot* and *caught* the same, speakers of Scottish English are also likely to have the same vowels in *Sam* and *palm*, the same vowels in *foot* and *goose*, and different vowels in *agreed* and *greed*. Speakers in England do not. The result is that Scottish English and English English are quite different, not just because they are different with respect to one feature, but because they are different with respect to many. Again, it is no accident that many speakers of English in northern parts of Ireland align with Scottish speakers as far as these features are concerned rather than with English speakers. The more features varieties have in common, the more similar they are to each other, and assessing the consequent degree of similarity between varieties is crucial if we want to answer all kinds of interesting linguistic, social, and historical questions.

5.2.1 Isoglosses

The traditional approach to assessing the similarity of varieties of a language, and determining the divisions between varieties, was to determine the geographical location of features, since there is an obvious, easily observed connection between geographical distance and linguistic difference. Chambers and Trudgill (1980: 6–8) describe a typical pattern whereby differences between varieties gradually accumulate as geographical distance increases, such that geographically proximal varieties are separated by few linguistic features but geographically distant varieties are separated by many. This means that in a geographical chain of locations, A to Z, the varieties spoken in A and B will be rather similar to each other, variety A will be quite different from variety M, and variety A and variety Z may be extremely different, to the point that they could be considered to be different languages. For some excellent illustrations of the differences which can be found across geographical continua (in this case in Dutch dialects), see Heeringa and Nerbonne (2001).

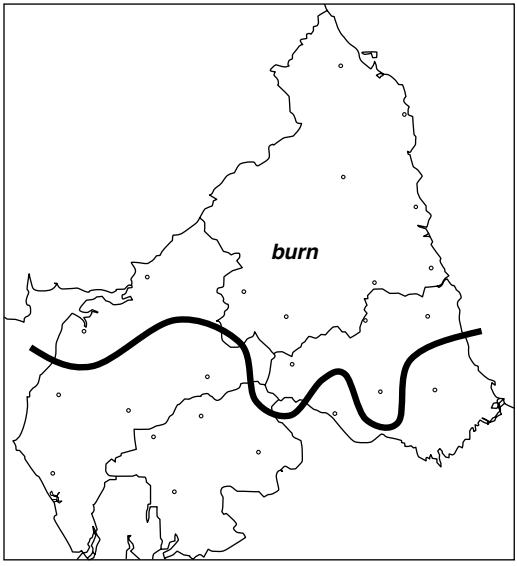


Figure 5.1 Distribution of the response 'burn' to SED Question IV.1.1

Despite the apparently continuous nature of many dialect relations, it is still possible to map the distribution of particular linguistic features. The most common way to do this is to draw *isoglosses*, which are defined by Chambers and Trudgill (1980: 103) as 'the boundaries between two regions which differ with respect to some linguistic feature'. Isoglosses are represented as lines drawn on a map between the locations which differ. An example of an isogloss, derived from the *Survey of English Dialects* (Orton, Sanderson and Widdowson 1962–71, henceforth SED), is given in Figure 5.1.

This map, which covers only the northernmost counties of England, represents the distribution of the response 'burn' to the question *What do you call any running water smaller than a river?* (SED Question IV.1.1). Localities (indicated by points) where the response 'burn' was elicited are found north of the isogloss and localities where 'burn' was not elicited are found to the south of it. The isogloss indicates that localities to the north form a group with respect to this feature. However, note that various responses were elicited south of this area, so that those varieties don't necessarily form a group with respect to this feature.

For further examples of isoglosses, see Chambers and Trudgill (1980: 31, 104, 111, 128) and the many maps in the *Linguistic Atlas of England* (Orton, Sanderson and Widdowson 1978).

The isogloss in Figure 5.1 indicates the geographical distribution of only one linguistic feature, a tiny fragment of the variation which underlies this

particular geographical dialect continuum. What happens when we map more linguistic features and what can we learn about the relationships between dialects in doing so? Chambers and Trudgill (1980: 105–11, 125–7) identify three patterns, each of which has different consequences for the definition of dialects and the similarity between them:

1. *Random patterning*: Isoglosses criss-cross the map without seeming to be related to each other in their distribution. Chambers and Trudgill (1980: 107) note that this is 'recognised as a typical pattern for any region that has a long settlement history' and it is entirely consistent with dialects being related to each other in a continuous fashion. It does not mean that the isoglosses do not add up of course – it is still likely that varieties which are geographically distant from each other will be separated by many isoglosses and will be quite different as a result. But there may not be any obvious dialect boundaries between them.
2. *Bundling*: Isoglosses for different features may follow the same course, or bundle (even if there is some variation in their precise location). Bundling of isoglosses indicates that varieties separated by the bundle are rather different. Chambers and Trudgill (1980: 109) note that 'the significance of a dialect area increases as more and more isoglosses are found which separate it from adjoining areas'.
3. *Transition zones*: Between these two extremes, isoglosses may roughly follow the same course but may not bundle so tightly together. Rather they fan out over a wide area, such that varieties on one side of the isoglosses are rather different from those on the other side, but those in between are transitional.

Although isogloss bundling seems somewhat at odds with the concept of the dialect continuum, Chambers and Trudgill (1980: 127) point out that the two notions are not necessarily contradictory:

> Isogloss bundles … are made up of lines which are in the same vicinity but are hardly ever contiguous. In moving from the region on one side of a bundle to the other, then, one would have the impression of a continuum, since first one feature and then another and eventually another would vary from site to site. Thus the notion of a bundle, which is based on the notion of the isogloss, can be reconciled with geographic gradualness.

The key point is that, although tight bundles of isoglosses may mark sharp transitions between varieties, it is not essential to have tight bundles or, indeed, obvious bundles of isoglosses at all, for varieties to be considerably different from each other. We return to this point in §5.3.

There may well be structure in the distribution of isoglosses, but it need not necessarily be entirely of one or other of these types of distribution. An excellent example of the interaction of (lexical) isoglosses is detailed in Glauser

(1974), which deals with the transition between the dialects of northern England and those of southern Scotland. Glauser draws thirty-nine isoglosses on a single map (p. 250), revealing a mixture of isogloss bundling (particularly along the Scottish–English border), transitional areas (particularly north Northumberland and south Dumfriesshire), and a fair amount of random distribution everywhere, especially in northern England. From Glauser's map, we can tell quite a bit about relations between varieties in this region. For example, locations in Scotland which are separated by few isoglosses are relatively similar, whilst locations in England which are separated from those in Scotland by many isoglosses are relatively dissimilar from them. In some cases, the transition between locations is very sharp and in others it is less abrupt, but it is the number of isoglosses separating varieties which is important for assessing how similar/different they are. In other words, isogloss approaches, such as that adopted by Glauser, are tantamount to a kind of informal quantification of the similarities and differences between varieties and allow us to make general statements about the relationships between varieties – Scottish varieties are relatively homogeneous, for example, but are quite sharply distinguished from English varieties, which in turn are somewhat heterogeneous (at least in comparison with the Scottish ones).

5.3 Problems with isogloss and feature bundles

As was discussed above, Chambers and Trudgill (1980: 109) argue that 'the significance of a dialect area increases as more and more isoglosses are found which separate it from adjoining areas'. This is reflected in the importance isogloss bundles have been given for defining dialects – see, for example, the division between Low German and High German (Chambers and Trudgill 1980: 106) in Germany and adjacent areas, the division between *langue d'oc* and *langue d'oil* in France (Chambers and Trudgill 1980: 111), and the division between northern and non-northern dialects of English in England (Wakelin 1984a: 73). In many cases, these linguistic isoglosses pattern along with distinctions of cultural practice, whether these involve food, customs, architecture or farming, for example. Labov, Ash and Boberg (2006: 41) explain the rationale for selecting language features for defining dialect areas in the following way:

> Dialect geography has traditionally been concerned with the search for a principled basis for dividing dialects and drawing boundaries (or isoglosses) between them … Perhaps the most important consideration in selecting a parameter for dialect classification is the degree of spatial differentiation it displays. Any examination of candidates for dialect markers must reject those that appear to be randomly distributed in space in favor of those with the greatest regional differentiation, no matter how particular or general they are.

This reasoning, which is central to isogloss approaches, seems sensible (throw out the junk and concentrate on the significant patterns), but there are major problems with it which suggest that alternative approaches are necessary for defining the relations between varieties of a language. These problems are discussed in this section, whilst alternative approaches are outlined in §5.4.

If dialect divisions are defined by isogloss bundles then it is precisely the features that bundle which define dialect areas. But how many such features do we need to define a dialect area? Since it is clear that dialects can be very different from each other without being separated by any obvious isogloss bundles (as the dialect continuum model predicts), should other features which don't bundle, or which seem to pattern in random ways, not be considered to be equally important 'dialect markers' if they nevertheless distinguish some varieties from others? And are there inherent problems with determining dialect boundaries on the basis of the geographical distribution of features?

With regards to the number of dialect markers required to define dialects, no definite answer is possible. Although even the close bundling of just two different linguistic features across a wide geographic range is significant (given that it is extremely unlikely for two independent features to follow, at random, the same course for any distance on a map) this does not mean that varieties separated by two (or more) isoglosses need be very different at all. Indeed, they may be much less different than varieties which are separated by dozens of 'random' isoglosses which do not follow any particular pattern. Why should isogloss bundles be privileged in such a case? Instead, randomly distributed features can be important for determining how similar or different varieties are to each other, even if no obvious boundary falls between them. And it need not be assumed that just because features appear to be distributed randomly there is no structure in their distribution – it is possible, for example, that in 100 'randomly distributed' features, two locations might have the same value for 60 of them whilst another variety might have the same value for only 30 of them. Isogloss analyses can't capture these potentially important, cumulative differences. This problem relates to the geographical limitations of isogloss analysis of linguistic features. Even when isoglosses pattern in obvious ways on a map, varieties which are non-contiguous, but which nevertheless share the same variants of particular features, are still represented as different areas. Since it is not difficult to think of historical scenarios which could give rise to such non-continuous, fragmented relationships (e.g. migration), it is clear that isoglosses impose serious restrictions on how we can represent the relations between varieties. When we add to this the fact that it is difficult to capture social variation on maps, since we may have rather different linguistic features for different social groups at the same geographical location, it is clear that isoglosses are not necessarily the best means for representing the relations between varieties.

Related to this issue is the problem of where we place isoglosses on a map. As was discussed above, Chambers and Trudgill (1980: 103) define an isogloss as a line 'drawn between the locations of any two speakers exhibiting different features', and note (p. 105) that any given isogloss 'cuts arbitrarily through what is in fact unknown territory'. They do not tell us exactly where this line should be drawn, and in their example (p. 104) it is essentially drawn freehand. This is problematic, since the placement of the isogloss is, to a certain extent, subjective, and it leaves unanswered the problem of how much leeway we can have when drawing an isogloss – is it acceptable to draw distended loops to include outlying locations within the isogloss, for instance? That is, there is no way of determining what should appear inside and what should appear outside the isogloss, and it leaves open the question of what constitutes a contiguous location.

In order to avoid these problems, we need to define the position of isoglosses objectively. We can do this using *Delaunay Triangulation* and *Voronoi Tessellation* (Delaunay 1934; Voronoi 1907; see also Krämer 1995 and Heeringa 2004: 161–2), which are defined in the following way:

1. Connect locations by drawing circles such that three locations lie on the circumference of each circle but with *no location falling inside any of the circles* (see Figure 5.2a);
2. Locations which are found on the circumference of the same circle are adjacent – connect them with a line; this is *Delaunay Triangulation* (Figure 5.2b); this kind of map is also known as a *beam map*;
3. To define the area of the map which lies closest to each location, draw lines at the exact mid-point of these lines, but perpendicular to them. These new lines will connect with each other in three-way junctions (which are at the exact centre of the circles), forming an area around each location; this is *Voronoi Tessellation* (Figure 5.2c); this kind of map is also known as a *honeycomb map*;
4. This method defines exactly which locations are adjacent, the exact area which is closest to each location, and the exact borders between these areas.

An example of a honeycomb map, of locations in the far northern English counties in the SED, is given in Figure 5.3, which also illustrates how we can use this kind of analysis to represent isoglosses (in this case instances of the response *burn*, as per §5.2.1 above).

Figure 5.3 illustrates some important features of Voronoi Tessellation: it defines exactly the mid-point between locations so that we can place isoglosses objectively (on the edges between the locations); the edges define a tile or cell which contains all points on the map which are closest to the location at the centre of it; it reveals which locations are contiguous and which are not (so

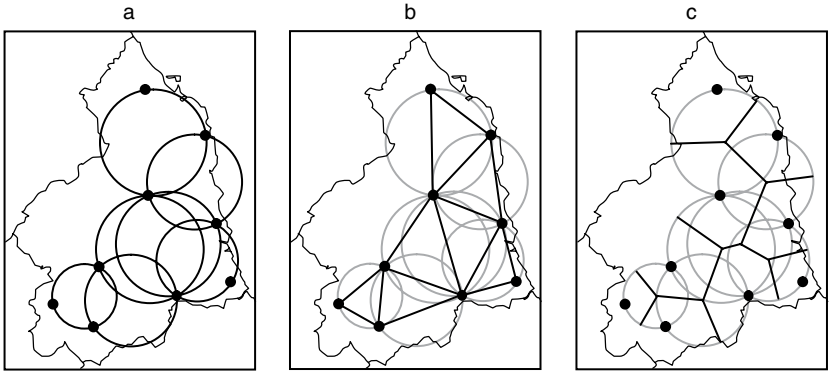


Figure 5.2 Delaunay Triangulation and Voronoi Tessellation of the SED locations in Northumberland

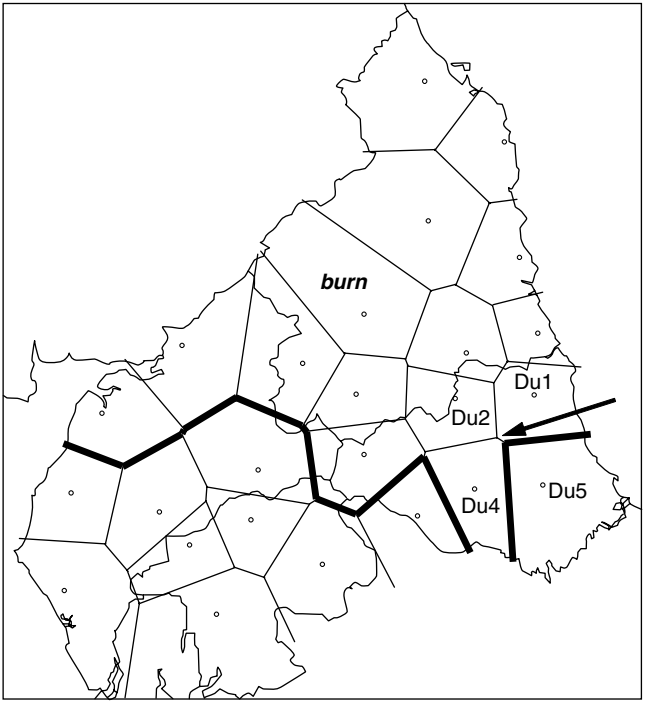


Figure 5.3 Voronoi Tessellation of northern SED locations and the isogloss for *burn*

Du1 and Du4 are contiguous, since they share an edge; Du2 and Du5 are not because they do not); and, consequently, it defines whether or not apparent outliers are geographically connected to other locations with the same variant.

Voronoi Tessellation is extremely useful for representing relations between varieties (see §5.5). There are other methods for making the placement of isoglosses more objective than drawing them freehand, although they don't have quite the precision of Voronoi Tessellation – see, for example, Labov *et al*. (2006: 41–3), Hägerstrand (1952) and Britain (1997).

Returning to issues with defining relationships between varieties, another rather serious problem is that very often those relationships are not of an either/or type but are, instead, of a more/less type. Isoglosses struggle to capture these more complex kinds of relationship. A good example of this is provided by Chambers and Trudgill (1980: 127–37), where they consider the transition between northern English [ʊ] and southern English [ʌ] in the STRUT lexical set (Wells 1982: 131–2). Although they draw an isogloss dividing these two pronunciations (see the map in Chambers and Trudgill 1980: 128), this is a considerable abstraction away from reality. When they examine the border between the two pronunciations more closely, they find not only that there are intermediate forms such as [ɤ], but also that some locations vary between [ʊ] and [ʌ], using the different vowel variants to one degree or another. This means that an isogloss, no matter how objectively defined, cannot represent the reality on the ground. Instead, Chambers and Trudgill resort to other kinds of cartographical representation – using symbols and shadings to indicate varieties with different kinds of pronunciation (p. 133), and using a frequency map, indicating the percentage of [ʊ] vowels in words of the STRUT lexical set at each location (p. 130). They add an isogloss to the frequency map, separating localities with greater than 50 per cent [ʊ] from those with less than 50 per cent [ʊ], but it is the percentages, not the isogloss, which indicate the relationships between varieties. Chambers and Trudgill (pp. 137–42) find a similar situation with the vowel in the BATH lexical set (Wells 1982: 133–5), and it seems likely that this kind of transition is the rule rather than the exception (see Anderson 1987 for ample exemplification). Since transitions of this sort are not really equivalent to isoglosses, defining dialect areas by isogloss bundles is even more problematic.

The final problem with using isoglosses to define relationships between varieties is again a crucial one. As we pointed out above, isoglosses give an indication of how similar or different varieties are (varieties separated by few isoglosses are relatively similar, those separated by many are relatively different). Hence, we could count the number of isoglosses which separate varieties to give us a rough quantification of the similarity between varieties. But all of the problems discussed above mean that this is, in practice, impossible – isoglosses may be 'randomly' distributed so that the number of isoglosses between two

locations isn't equivalent to the difference between them; isoglosses abstract away from a more complex reality; and the features which are mapped may only represent a small fraction of the possible distinctions which may or may not be important in determining relations between varieties. In other words, isoglosses are not particularly well suited for quantification, even if they imply degrees of similarity and difference between varieties. This means that if we wish to measure relationships between varieties we will need other means of analysing and representing the similarities and differences between them. We turn to the issue of how we might do this in the next section.

5.4 Quantifying relations between varieties

In order to avoid the problems inherent in the isogloss approach to analysing relations between varieties of a language, we need to take the following steps:

1. Quantify precisely how similar or different varieties are to each other;
2. Compare all varieties with each other, even non-contiguous ones, so that the relationships between them can be properly appreciated;
3. Compare varieties across a wide range of features, not just those which we judge to be important for one reason or another;
4. Represent the relationships between the many varieties in a way which captures the complexities of the relationships between them.

There are many different ways of quantifying the distance between linguistic varieties, and a number of these are discussed in greater detail below. For further discussion, see Nerbonne and Kretzschmar (2003), Heeringa (2004: 14–24), Kessler (2005), and Nerbonne and Hinrichs (2006). Perhaps the most conceptually simple approach is percentage similarity/difference, which is explained further in §5.4.1. Similar to this (but not expressed as a percentage) is the calculation of Hamming distance, as used, for example, in Embleton and Wheeler (1997) to calculate distances (based on morphological, syntactic, and lexical features) between dialects of English, and in Spruit (2006), to calculate distances (based on syntactic variables) between Dutch dialects. In a rather different kind of approach, a number of researchers have compared the frequency of segments or features in different varieties in phonetically transcribed texts (e.g. Hoppenbrouwers and Hoppenbrouwers 1988, 2001; Moisl, Maguire and Allen 2006, Moisl and Maguire 2008, and Moisl, this volume). Each variety is characterised by a particular frequency of each segment or phonetic feature, and the frequency of all features for each variety constitutes a *vector*. These vectors are compared between varieties and the distances between them calculated using some metric (for example, Euclidean distance). In a similar approach, Szmrecsanyi (2008) calculates the Euclidean distance between English dialects based on the frequency of morphosyntactic features in a corpus of recordings.

A number of techniques have been developed specifically to handle comparison of the phonetics and phonology of varieties, including Levenshtein distance (discussed in detail below), the 'Sound Comparisons' method (also discussed below), and the method described in Maguire (2008) for quantifying similarity by comparing the lexical distribution of phonemes.

An important aspect of quantificational methods which sets them apart from the traditional isogloss approach is that the relationships between varieties are quantified across a wide range of linguistic features. We saw in §5.2 that when even a handful of isoglosses bundle, this might be taken as evidence of a significant dialect distinction, even if hundreds of other linguistic features are acting differently. In the methods described in this section, hundreds of different features are compared, regardless of how they are distributed geographically, so that the patterns which emerge are much more accountable to the data, and the impact of patchily or 'randomly' distributed features is factored in. If there is a sharp distinction between varieties, it is the result of them differing with respect to many linguistic features, not just one or two, and we can be confident, as a result, that the distinctions and groups which emerge are so much more robust.

Once we have a method for measuring similarity or difference between varieties, we apply this to every pairwise grouping of varieties, and we can do this regardless of the geographical relationships between them (which means, for example, that non-contiguous varieties may be compared, and different social varieties from the same location could also be compared). If we have four varieties, A, B, C, and D, we compare the following pairs: A-B, A-C, A-D, B-C, B-D, C-D. The total number of comparisons involved may be calculated using the following formula, where N = the number of varieties entered into the comparison: (N(N-1))/2.[1] So where we have twenty varieties, the total number of comparisons is (20*(20–1))/2 = 190.

The usual way of displaying these similarity scores is in the form of a *matrix*, as in Figure 5.4.

Note that each cell in the matrix contains a similarity or distance score for a unique pairwise comparison. Also note that when the values in the matrix are percentage similarities, it is a *similarity matrix*, and a *distance matrix* can be calculated from the similarity matrix by subtracting the percentage similarities from 100 per cent (Goebl 2006: 413). Furthermore, the similarity of each variety to itself is obviously 100 per cent (as indicated in Figure 5.4).

Although a matrix contains a vast amount of information on the relationships between varieties, it is very difficult to comprehend the patterns in it without further processing. There are many different techniques for analysing the data in a matrix to reveal the structure in it, and only a few of these will be discussed further in this chapter (§5.5). First, however, in the following sections we examine a number of approaches for quantifying the relationships

A	B	C	D	
1.0	0.3	0.3	0.4	**A**
	1.0	0.4	0.3	**B**
		1.0	0.3	**C**
			1.0	**D**

Figure 5.4 An example matrix

between linguistic varieties before turning to the issue of how we interpret and represent the results they give.

5.4.1 Categorical approaches

Quantitative methods for assessing the overall similarity of linguistic varieties were developed for historical linguistics before they were applied to dialectology. Morris Swadesh in particular advocated the use of *lexicostatistics* for quantitative comparison of different languages (e.g. Swadesh 1950, 1952, 1955; see McMahon and McMahon 2005: ch. 2, for a discussion). Lexicostatistics works as follows:

1. Construct a list of meanings which are found in each of the varieties to be compared (Swadesh used basic vocabulary items, arranged in two lists, one of 100 items, another of 200 items);
2. List the lexical item which corresponds to each meaning in the wordlist for each language being compared (so, for example, for the meaning 'dog', the English lexical item is *dog*, the German one is *Hund*; note that we do not compare the English cognate *hound* as this has a *different meaning* in modern English);
3. Compare the list for each language against the list for every other language under analysis (so if we include English, German, French, and Irish in the analysis, we would compare the following pairs: *English–German*, *English–French*, *English–Irish*, *German–French*, *German–Irish*, *French–Irish*);
4. For each meaning being compared between languages, a score of 1 is given if the two lexical items are cognate and a score of 0 is given if they are not (note that this method relies on knowing in advance which words are cognate; matches as a result of borrowing are not considered to be cognates);
5. Calculate the total score for each pairwise comparison of languages and express as a percentage of the number of meanings in the list; this is the *percentage similarity* of the languages (so, for example, if two varieties share 60 cognates in a list of 100 meanings, they will be 60 per cent similar);

6. The distance between languages can be calculated by subtracting the percentage similarity from 100 per cent (so our two languages would have a distance of 40 per cent).

There are a number of crucial points to make about lexicostatistics. Firstly, the meaning-list used is designed to be the one which is most likely to contain cognates in related languages, but it is also intended to be neutral in that its members are not specially selected to emphasise particular features or relationships; however, since the method relies on comparing cognates, it excludes the effects of contact and borrowing from the equation. Secondly, it should only be applied to varieties which are known to be related and for which an assessment of cognate matches is possible. Thirdly, lexicostatistics should not be confused with *glottochronology*, which is an additional (and contentious) method designed to determine the timescale implied in the similarities and distances revealed by lexicostatistics – see McMahon and McMahon (2005: ch. 7).

A number of other issues and problems with lexicostatistics have been identified, some of which are relevant to the application of a similar method to varieties of a single language (see Embleton 2000 and McMahon and McMahon 2005: 40–4), including the universality and representativeness of the meaning-list, multiple synonyms (as in English *little* and *small* for the meaning 'small'), partial overlaps in form (e.g. French *coeur* and Spanish *corazón*), restriction to lexical comparison, and exclusion of borrowings which may reveal just as much about the history of a language as cognates. Various approaches have been developed to deal with these issues (again, see McMahon and McMahon 2005: 40–4). Nevertheless, lexicostatistics provides a useful and objective means of assessing relations between languages in a way which is complementary to the comparative method.

Approaches rather similar to lexicostatistics, at least in terms of the calculation of similarity/distance, have been developed for comparison of varieties of a single language (see, for example, Séguy 1971, 1973; Goebl 1984, 1997, 2006 and 2007; and Goebl and Schiltz 1997). These methods, to which the term *dialectometry* has been applied, work on the principle that it is possible to quantify the distance between varieties using strictly comparable data from dialect surveys. Traditional dialect studies typically surveyed a number of locations at which responses to a linguistic questionnaire were elicited (see, for example, the SED). The data for each location consists of a set of answers to the questionnaire, and we can, in a very similar way to lexicostatistical analysis, compare the answers from each location, scoring 1 when an answer for a pair of locations is the same and 0 when it is not. So two locations answering *burn* to the question *What do you call any running water smaller than a river?* score 1, but two locations answering *burn* and *stream* respectively score 0. The similarity of any two varieties is the percentage of matching answers in the

questionnaire, and the distance between them is derived, as before, by subtracting the similarity from 100 per cent.

Although this method is ostensibly similar to lexicostatistics, there are a number of crucial differences. This kind of method does not rely upon cognacy – answers match if they are the same, whatever the reason for that identity might be – since it is often impossible to tell the source of particular features in a variety. Nor is this method restricted to lexis, although that is a domain where it is easily applied. It can also be used to compare morphology, syntax, and gross phonetic differences (and, indeed, Goebl advocates an approach where varieties are compared across multiple linguistic levels). Furthermore, the 'meanings' in the questionnaire are not necessarily 'basic' but can be considered to give a general overview of the dialects being investigated, as specific features are not chosen to highlight particular relationships (as is done in the isogloss approach). Some of the problems with lexicostatistics also apply here, and decisions need to be made how to deal with them. It might be the case that answers to a single question at different locations have a slightly different meaning, although it could be argued that the questionnaire controls for this to an extent and the differences which exist are not significant enough to create problems. More than one answer might be possible for particular questions, and informants might give one or the other, or both, in which case it might be necessary to include more than one answer for some questions (so that some matches between varieties might be fractions rather than 1 or 0). And answers may be partially similar but not the same, in which case we can count them as entirely different (score 0), or attempt to estimate the extent to which they match (again giving a fraction match rather than 1 or 0).

However, one linguistic domain which categorical approaches struggle to deal with is phonetics, where relationships are often gradient ([i] and [e] are more similar than [i] and [ɑ], for example), and where differences between varieties for a single word, for instance, involve more than just matches or non-matches (because of deletions, epentheses, metatheses, and so on). Other methods have been developed for quantifying the similarity of varieties at the phonetic level, and we discuss two of these in sections 5.4.2 and 5.4.3.

5.4.2 Levenshtein distance

An approach which has proved effective for measuring the difference between varieties at the phonetic level is the calculation of *Levenshtein distance* (Levenshtein 1966; Kessler 1995; Nerbonne and Heeringa 1997; Kruskal 1999; Nerbonne *et al*. 1999; Heeringa and Nerbonne 2001; Nerbonne and Heeringa 2001; Heeringa 2004; and Heeringa *et al*. 2006). Levenshtein distance works on the principle that it is possible to convert any string of characters into any other using a series of deletions, insertions, and substitutions, and the number

of such operations required to do this is the *cost*, or the *distance* between the strings. Using the pronunciation of the word *afternoon* in two varieties of English as an example (see Heeringa 2004: 124), we can calculate the cost of converting one into the other as follows:[2]

Pronunciation 1:	[æəftənʉn]	
Pronunciation 2:	[æftərnun]	
Starting with:	[æəftənʉn]	
Delete [ə]:	[æftənʉn]	1
Insert [r]:	[æftərnʉn]	1
Substitute [u] for [ʉ]:	[æftərnun]	1
Total cost:		3

This conversion can be represented in the following way, which makes the details of the operation clearer:

Variety 1	æ	ə	f	t	ə	—	n	ʉ	n
Variety 2	æ	—	f	t	ə	r	n	u	n
Cost	0	1	0	0	0	1	0	1	0

Thus the cost or distance is 3. Note that this is the 'cheapest' means of converting one of these strings into the other – any other alignment of the two strings and subsequent set of conversions would result in a more costly conversion, as the following scheme indicates (Heeringa 2004: 122):

Variety 1	æ	ə	f	t	ə	n	ʉ	n
Variety 2	æ	f	t	ə	r	n	u	n
Cost	0	1	1	1	1	0	1	0

The cost of converting between the two strings in this alignment is greater (5), so it is rejected.

Although Levenshtein distance provides an efficient way of determining the distance between phonetic strings, there are a number of issues which mean that refinements to the method are desirable. Firstly, comparison of longer words raises the likelihood that the distance between varieties will be increased (as there are more phonetic slots to match); some means of normalising distances is required. Secondly, a method is required to compare words of uneven length. Thirdly, Levenshtein distance, as described above, is rather crude, allowing any segment to be converted into any other at the same cost. Thus converting [p] into [b] is just as costly as converting [p] into [a], although in terms of sound changes or correspondences between varieties, the former is much more likely than the latter. It is only possible to give a brief outline of some of the methods

which have been introduced to address these problems here. For further details, see the references provided above.

In order to normalise for word length, so that long words do not count more than short words, the Levenshtein distance is divided by the number of segmental slots in the match between the two strings. Thus, in the example of *afternoon* above, the score, 3, is divided by the number of slots, 9, giving a normalised distance of 0.33 for this word. Since the least similar these two strings can be is 9 (no exact matches in segment), this is in effect a percentage distance between the two strings (33 per cent) (Heeringa 2004: 132).

The mechanism adopted for *afternoon* can also be used to compare words of uneven segment length. Thus if we compare Irish English [fɪləm] *film* with English English [fɪɬm], we match them as follows:

Variety 1	f	ɪ	l	ə	m
Variety 2	f	ɪ	ɬ	—	m
Cost	0	0	1	1	0

In the discussion of Levenshtein distance thus far, any segment can be converted into any other, and will always incur the same cost. So if we chose to convert [p] into [b], the cost (distance) is 1, and the same goes for the conversion of [p] into [a]. Although aligning the strings using the method described above will prevent this situation from arising in every case, it isn't fool-proof. In the following example (adapted from Heeringa 2004: 131), two pronunciations of the word *shaft* are compared, [ʃæəf] and [ʃaft]. There are two obvious ways of comparing these two strings:

Variety 1	ʃ	æ	ə	f	—	ʃ	æ	ə	f
Variety 2	ʃ	a	—	f	t	ʃ	a	f	t
Cost	0	1	1	0	1	0	1	1	1

Note that in both cases, the cost of conversion is 3, but the percentage distance is different (60 per cent in the first case, 75 per cent in the second). On percentage alone, we might pick the first option, and it also seems preferable another way: it compares [f] with [f] and avoids the comparison of [f] with [t] as a result. But the problem of phonetically similar conversions costing the same as phonetically distant conversions remains ([æ] to [a] costs the same as [ə] to [f]). One way of avoiding the conversion from [ə] to [f] is to stipulate that in the alignment of phonetic strings, vowels may only match with vowels (or nothing) and consonants may only match with consonants (or nothing), but semi-vowels such as [w] and [j] are allowed to match with either (see Heeringa 2004: 125). A more complicated approach, described in Nerbonne *et al*. (1999), is to treat each 'segment' as a bundle of phonetic features, such

that phonetically similar segments share more features in common than they do with phonetically distant ones. The distance between segments can be calculated as in the following example, taken from Nerbonne *et al.* (1999: ix):

	ɪ	e	u	ɪ -e	ɪ -u
Advancement	2 (front)	2 (front)	6 (back)	0	4
High	4 (high)	3 (mid)	4 (high)	1	0
Long	3 (short)	3 (short)	3 (short)	0	0
Lip-rounding	0 (none)	0 (none)	1 (rounded)	0	1
Total				1	5

This approach has the advantage of capturing phonetic detail, but it does have the consequence that the distance between segments must be measured in some other way (e.g. as vectors in Euclidean space).

However Levenshtein distance is calculated, an assessment of the distance between varieties can be made by calculating the difference for many words, selected to represent the sound patterns of the varieties under consideration, and expressing the overall distance as the average distance per word (Heeringa 2004: 133–4). The result is a matrix of distances between varieties which can then be analysed further to represent the relations between varieties (see §5.5 below).

5.4.3 The 'Sound Comparisons' method

An alternative approach to comparison of varieties at the phonetic level which has, at its core, a concern with quantifying fine phonetic differences, is proposed in Heggarty *et al.* (2005), McMahon and McMahon (2005: ch. 8), McMahon *et al.* (2007), and Maguire *et al.* (2010). This method compares phonetic transcriptions of a standard wordlist in varieties of English and other Germanic languages and measures the overall phonetic similarity as an average of the similarity of each of the phonetic segments in the transcriptions being compared.

In order to determine which phonetic features in each variety should be compared with each other, the method uses *node forms* which are, in effect, the ancestor forms of the cognates in each variety. Figure 5.5 demonstrates how the Tyneside English pronunciation of *daughter*, [dɔːt͡ʔɐ(ɹ)], is compared with the Buckie Scots pronunciation [doχtör] through the Proto-Germanic ancestor form, *doxteːr.[3] Note that the Tyneside and Buckie forms are not themselves compared to the Proto-Germanic form – it is instead used to determine which phonetic features of each variety should be compared with which.

Figure 5.5 reveals that [d] in Tyneside should be compared with [d] in Buckie, [ɔː] with [oχ] (i.e. [ɔ] with [o] and [ɔ] with [χ]), [t͡ʔ] with [t], [ɐ] with [ö], and [ɹ], which is only variably present in Tyneside, with [r] in Buckie.

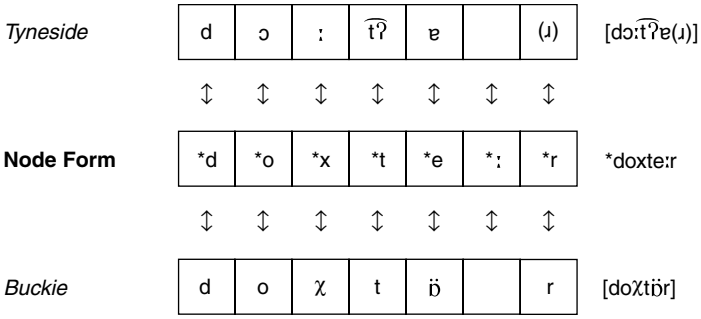


Figure 5.5 Comparison of pronunciations of *daughter* in two varieties of English

The next step is to determine how similar the segments are to each other. This is done by assessing the similarity of the phonetic features which define each segment. Taking [t] and [d] as an example, these two sounds are the same for place and manner of articulation, but are different for voicing. That is, they agree in two out of three features, or are 67 per cent similar. However, there are typically three distinctions of place and manner cross-linguistically, but only typically two distinctions in voicing, so the weighting of each of the features is adjusted to reflect this, with place and manner each being given a weight of 2 and voicing being given a weight of 1. The result is that [t] and [d] are considered to be 80 per cent similar, as they only differ on voicing. [d] and [g] are also considered to be 80 per cent similar, since they agree in voicing (1 point) and in manner (2 points) and although they disagree in place, the distance between them is not as great as the distance between [b] and [g], so they differ by only 1 point on this dimension too. [b] and [g], on the other hand, are only 60 per cent similar, since their places of articulation are further apart (they are the same for voicing, 1 point, manner, 2 points, but not for place, 0 points). Comparing phonetic segments which share even fewer features gives lower similarity scores again, so that [b] and [s] are only 40 per cent similar (0 points for voicing, only 1 point for manner rather than 2, and 1 point for place).

Vowels work in a similar way, although things are complicated by the shape of the vowel space. So a high vowel compared with a high-mid vowel scores 2 out of 3 on the height scale, a high vowel compared with a low-mid vowel scores 1 out of 3, and a high vowel compared with a low vowel scores 0 out of 3. Rounding is scored as 1 (match in rounding) or 0 (no match). The front-to-back dimension is more complicated because there is greater phonetic differentiation in the high part of the vowel space than in the low part, but a similar system applies, with differences in fronting more significant for high vowels than for low vowels.

The system encodes much more detail than it is possible to give here, allowing for comparison of phonetically very similar segments such as [p] and [p^{h}] or [a] and [æ]. Further details of the phonetic comparison system can be found in the references given above, in Heggarty (forthcoming) and at www.languagesandpeoples.com/MethodsPhonetics.htm.

Using this system, a matrix of similarities for every segment compared against every other is constructed, and the overall similarity of varieties is expressed as the average percentage similarity across all segments in the transcriptions being compared. Because the method is capable of measuring very fine phonetic differences between varieties (as opposed to gross differences between different phonetic symbols), it is well suited to the analysis of relationships between accents of the same language and even of social variation within single locations – see Maguire *et al*. (2010) for discussion. And because this method is concerned with capturing the subtleties of the relationships between closely related varieties, it has adopted methods of representation which reveal the complexities of these relationships – see §5.5 for details.

5.5 Representing similarities and differences

As was discussed above, it is desirable to further analyse the relations in a matrix to reveal the structure that it encodes. There are many techniques for doing so, and only a few of these will be discussed further here. Methods used include cartographical representations (discussed below), cluster analysis and trees (discussed below; see also Moisl, this volume), network analysis (also discussed below), multi-dimensional scaling (see Embleton 1987; Embleton and Wheeler 1997; Heeringa 2004: 156–64; Spruit 2006; and Shackleton 2007), and principal components and factor analysis (Shackleton 2005, 2007; Clopper and Paolillo 2006; Labov *et al*. 2006: 146–7; and Nerbonne 2006).

Certain relationships contained in the distance matrix can, with relative ease, be represented cartographically. Taking a single reference variety in the matrix, it is defined by a set of similarity/distance scores to every other variety. The similarity of every variety to this reference variety can be represented on a map by, for example, shading each location such that high similarity to the reference variety is indicated by warm colours or dark shading, and low similarity by cold colours or light shading. This is one of the key methods of representation defined in the works of Hans Goebl (see, for example, Goebl 1984, 2006, 2007). Goebl uses *choropleth* honeycomb maps to indicate the similarity of French varieties to Parisian French, or of English varieties to RP, for example, giving us an instant visual representation of the relationships between standard varieties and regional dialects. Such maps give us an immediate insight into such issues as the spread of standard influence, the origins of the standard language, or recent expansion of the language into new territory. The same kind

of representation can be used to indicate the relationships of all varieties to any single variety.

Although this method is extremely fruitful, it fails to make use of the vast amount of information contained in the matrix since it represents only one dimension of it. Further information can be captured in maps of a similar form by indicating the average similarity of each variety to every other variety (Goebl's skewness maps; see Goebl 2007: 159–60 for illustrations). Thus, for example, locations which are, on average, more similar to other varieties may be indicated by warm colours and varieties which are less similar to other varieties, on average, may be indicated by cold colours.

Further aspects of the data matrix can be displayed cartographically using isoglosses and beams. In honeycomb maps, each location is adjacent to a number of other locations, and shares edges (where the adjacent polygons meet) with them. Each of these edges can be used to represent a similarity or distance value in the matrix. So where the similarity between adjacent varieties is high, we can represent this by making their shared edge thin or lightly shaded. Where the similarity between adjacent varieties is low, we can represent this with a thick or darkly shaded edge. The result is a kind of isogloss map, but one representing gross similarity (or difference) rather than the distribution of particular linguistic features (see Goebl 2007: 161–2 for examples). Closely related to this is representation using beams, where every location is connected to adjacent locations (as defined by Delaunay Triangulation) by a line (or beam). This line can be shaded to indicate relationships between these varieties (so that, for example, a thick line indicates a high degree of similarity). Where geographical groups of varieties are consistently similar to each other, the result is a cluster of obvious beams, but where there are major divisions between geographical areas, there will be little such clustering (again see Goebl 2007: 161–2 for examples).

Although all of these methods are extremely useful, they are still limited in one respect or another since they represent only a subset of the relationships in the matrix. One way of representing all of the values in the matrix is to connect each of the locations on the map to every other location with beams, and to code similarities and differences by shading the beams in particular ways (as with beam maps connecting only adjacent locations). It will almost certainly be necessary to establish some threshold of similarity below which relations are not indicated, or are only faintly indicated, since all varieties are necessarily similar to one degree or another, and the result will likely be a confusing mess of beams. Excellent examples of this approach, used to indicate relations between Dutch varieties, can be seen in Nerbonne and Heeringa (1997) and Nerbonne *et al*. (1999). Note that these maps do not just represent the relationships between adjacent varieties, and are an excellent way of summarising all relations within the matrix and interpreting them in a geographical context.

There are some issues with even these kinds of beam maps, however. Firstly, they are not particularly well suited to representing the relative similarities of *all* varieties, regardless of how similar or different they are. If all relations are indicated using a beam map, the result can be very messy and hard to interpret. Secondly, beam maps force a two-dimensional geographical interpretation on the data, which may be problematic when the relationships between varieties don't correlate well with geographical distance as a result of contact, spread, or other (physical and human) geographical factors. If, for example, traditional west Cornish varieties of English are, on average, more similar to varieties in and around London than their geographic position suggests they should be (perhaps as a result of relatively recent language shift from Cornish to English – see Wakelin 1984b), this similarity would be indicated by more prominent beams connecting west Cornwall with the south-east of England. But these beams might obscure the fact that geographically intermediate varieties don't have quite the same relationships. Furthermore, the similarities of west Cornish varieties may be greater than expected, but they may still be low compared with adjacent varieties and would not, as a result, show up on a beam map with a particular threshold. Thirdly, it is difficult to represent relationships between social varieties of a language on a map, since the relationships between them aren't defined by geographical location but rather by parameters such as age, class, or gender, and beam maps don't resolve this problem.

The best way to get around these issues is to remove geography from the equation (at least temporarily) and to represent the similarities/distances between varieties graphically. Typically this means in *tree* form (trees are also referred to as *dendrograms* and *cladograms*). By 'tree' in this instance, we mean a geometric representation of the similarity/distance values between varieties rather than a genealogical tree (as used, for example, in historical linguistics) which is an interpretation of relative relatedness based on a theory of divergence (McMahon and McMahon 2005: ch. 1). Moisl (this volume) discusses cluster analysis, noting that some measurement of distance is necessary for grouping varieties. Since the distance matrix already encodes distance, hierarchical cluster analysis can equally be applied to the kind of data described in this chapter, the result being a tree which captures the similarity of all varieties to each other. As Moisl discusses, there are many different algorithms for the construction of trees (see also Sneath and Sokal 1973: ch. 5; Heeringa 2004: 140–50, and McMahon and McMahon 2005: ch. 3), but the basic principle is that varieties which are similar are grouped together in the tree and varieties which are far apart appear in different parts of the tree, and the positions of all varieties are mathematically defined. Thus, in the example in Figure 5.6, taken from Moisl, Maguire and Allen (2006), A and B are closest, so they are grouped first; this group is then further grouped with the next closest variety, which, depending upon the precise algorithm implemented, may be D (which

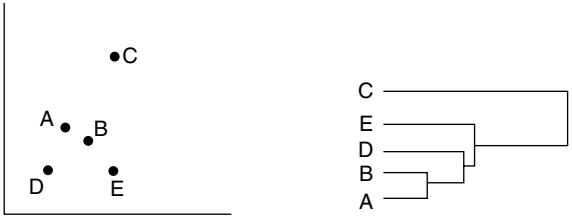


Figure 5.6 Constructing a tree

is closest to the 'average position' of A and B); this is further grouped with E and finally with C; resulting in the tree as indicated (the precise kind of tree again depends upon the clustering algorithm used).

Trees of this sort have a number of properties which make them well suited to representing similarities between varieties. Firstly, they are entirely independent of geography (which is not to say that linguistic relations don't correlate with geographical relations) – so that, for example, geographically non-contiguous varieties which share much in common can cluster together in the tree; social variation can also be captured, since it is just as possible to represent the relationship between younger and older speech at a single location, for example, as it is to represent the relationships between different geographical varieties; and the structure of the relationship between varieties can be revealed, perhaps showing that certain varieties group closely together whilst others group more loosely or are outliers (although it is necessary to understand what the clustering algorithm is doing since it may enforce particular structures on the data). Examples of trees showing the similarities and distances between varieties can be seen in Nerbonne *et al*. (1999), McMahon *et al*. (2007), and in Moisl's chapter, this volume.

Trees are not without their problems, however. Quite obviously, but very importantly, representing relationships between varieties using trees *assumes that the relationships between varieties are tree like*. This might well be the case, particularly when varieties have diverged rather sharply from each other without subsequent close contact. But this is much less likely to be the case when we are dealing with varieties of a single language, since we are likely to be confronted with a dialect continuum without major divisions between varieties. Just taking the artificial, but not unlikely, situation in Figure 5.6 as an example, A and B cluster together, and the combined cluster AB is grouped with D despite the fact that B is closer to E than it is to D. This is a result of the demands of the clustering algorithm, and it is one which fails to capture the complexities of reality. In other words, trees, which by their very nature assume divergence and branching between varieties, can't cope very well with cross-cutting, contradictory, and intermediate relationships.

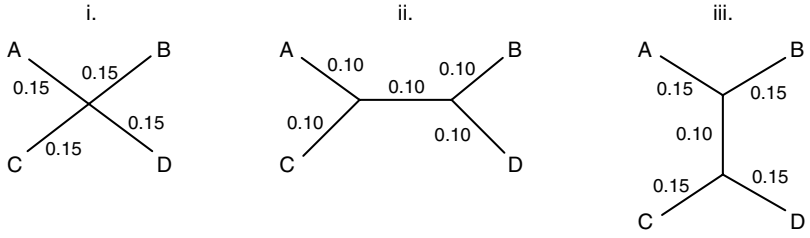


Figure 5.7 Wrong trees for A, B, C, and D

Let's take the example of four varieties, A, B, C, and D which are related as indicated in the matrix in Figure 5.4 above (thus 0.3 means 30 per cent different or 70 per cent similar). If we attempt to draw a tree using these data, we immediately run into a problem. Variety A is 0.3 different from both varieties B and C, but varieties B and C are 0.4 different from each other. Furthermore, variety D is 0.4 different from variety A, but is only 0.3 different from varieties B and C. This means that trees of the type illustrated in Figure 5.7 cannot capture the relationships between the varieties because there are conflicting signals in the data.

Tree (i) gets it wrong because it makes A-D and B-C too short; (ii) makes A-C and B-D too short; and (iii) makes A-C and B-D too long. It doesn't matter how we position the varieties relative to each other, or how long or short we make each of the branches of the tree, we cannot reconcile the contradictory relationships that hold in the data. The only way to do so is to draw cross-cutting lines, or *reticulations*, in the tree (which now doesn't look like a tree), as in Figure 5.8. Note that the distance between two varieties is represented by the shortest route along the lines between them. Introducing the reticulations allows us to capture the contradictions in the relationships between the varieties. Such representations are known as *networks*, and are much better suited to the representation of complex, cross-cutting, seemingly contradictory similarities/distances between varieties.

This discussion necessarily only scratches the surface of the theory and practice of using trees and networks for representing relationships between linguistic varieties – for further discussion of networks, see McMahon and McMahon (2005: ch. 6) and Huson and Bryant (2006). For examples of networks being used to represent relationships between linguistic varieties, see McMahon *et al*. (2007) and Maguire *et al*. (2010).

Although graphical representations of language relations such as trees and networks appear to make cartographical representations of the same relationships redundant, the results of these analyses can be mapped to indicate how the relations revealed by these methods pan out geographically. Thus, for example,

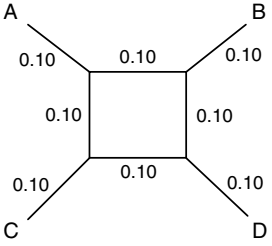


Figure 5.8 A correct 'tree' for A, B, C, and D

Goebl (2007) maps the major clusters revealed in his hierarchical analysis of French and English dialects.

5.6 Validation

The results of quantificational approaches provide us with insights into the complex relationships between varieties which simply cannot be appreciated using traditional methods. That does not mean they answer all the questions we might, as dialectologists, want to answer. We may, for example, be interested in the history and distribution of single features and their social significance, and other methods will be necessary to explore these issues. Quantificational approaches should be seen as complementary to these other approaches. An important final point to be considered in this chapter, albeit briefly, is how we evaluate the results of quantificational methods. How do we determine the effectiveness of different methods for measuring the similarities and differences between varieties, and how do we even know for sure that they are producing coherent and meaningful results? Although the methods compare and quantify features objectively and precisely, which in itself is an advantage over traditional dialectology approaches, other evidence must be brought to bear on the issue, especially if we seek to compare methods with one another.

There is a well-known relationship between linguistic similarity and geographical distance (see Chambers and Trudgill 1980: 6–8), certain exceptions such as colonisation, language shift and contact notwithstanding. It would not speak very highly in favour of quantificational methods if their results did not evidence the same relationship, and so correlation with geographical distance is an important means of evaluation. That the results of quantificational approaches do correlate very well with geographical distance was demonstrated in Séguy (1971), and this correlation has been found in subsequent quantificational approaches (see, for example, Heeringa and Nerbonne 2001; Goebl 2007; and Nerbonne and Kleiweg 2007), although the correlation appears to

be weaker for measurements of morphosyntactic distance (see, for example, Spruit 2006, and Szmrecsanyi 2008). In fact, this correlation is so basic that it can be used as a heuristic to identify significant events in the history of varieties such as migration into new territory (see Goebl 2006, 2007 for discussion).

Gooskens and Heeringa (2004) (see also Heeringa 2004: ch. 7), attempt to validate the results of their quantification of distances between Norwegian dialects (using Levenshtein distance) with the perceptions of differences between these dialects by Norwegian listeners. Having constructed two trees for the fifteen dialects under analysis, one derived from a Levenshtein distance matrix, the other derived from a matrix of perceptual judgements of the fifteen dialects by fifteen speakers of each dialect, they compare the results and find that 'the two dendrograms are rather similar, especially because of the fact that the closer clusters in the one dendrogram are also found in the other one'. (2004: 201). This is indicated by a highly significant correlation between the two distance measurements, even if the correlation is not perfect. Gooskens and Heeringa suggest that this may be the result of listener attitudes and knowledge and the fact that the recordings contained all sorts of detail not encoded in the phonetic transcriptions which are compared using Levenshtein distance (e.g. intonation and syntax). Nevertheless, the match between the two sets of distance measures is remarkably good and suggests that comparing production and perception in this way is a useful means of validating the results of quantificational approaches.

Another way of assessing the validity of quantification results is to compare them against the results from other procedures – for example with those of traditional dialectological and historical linguistic analyses, or with those produced by different quantificational approaches. One example of this approach is outlined in Heeringa *et al*. (2002) (see also Heeringa 2004: 178, and Prokić and Nerbonne 2008), where the results of various methods are compared against a 'gold standard', which is described as 'a classification of language varieties with which (nearly) all experts agree' (Heeringa *et al*. 2002: 446). This 'gold standard' is based on the relationships between Dutch dialects as defined by previous traditional dialect studies, which identified a number of clear, uncontroversial divisions between varieties. The degree to which the results from different quantificational methods agree with the 'gold standard' groups is an indication of how suitable they are for analysing relationships between varieties. There are a number of problems with this approach, however. Firstly, it assumes that the 'gold standard', as defined by traditional dialect approaches, is meaningful, which it might not be, given the concerns raised in §5.3 of this chapter. Secondly, the method proposed in Heeringa *et al*. (2002) compares groups but does not assess how accurate the similarities between varieties within those groups are (since the 'gold standard' does not contain

this information, so there is nothing to compare to). Nevertheless, this study indicates the kinds of approach which can be adopted and, combined with geographical and perceptual comparisons, it is a useful heuristic for assessing the validity of quantificational approaches.

Maguire *et al*. (2010) describe further approaches for assessing the validity of quantificational results. In their analysis, they compare the distances between traditional dialects of English and between modern accents (see Trudgill 1990 for this distinction). Trudgill notes (p. 6) that compared with the traditional dialects of English, 'The Mainstream Modern Nonstandard Dialects differ much less from Standard English and from each other'. This is something which can be tested with quantificational methods and can provide a means of assessing how well the method used analyses the relationships between varieties. Maguire *et al*. find that traditional varieties *are* significantly more distant from each other and from Standard English than modern varieties are. In addition, they find that the relationships between varieties reflect their geographical relations and demonstrate that even though traditional and modern varieties from the same location often vary considerably in their distances from other varieties, they are usually quite closely related to each other. All of these findings combined suggest that the quantificational method used is capturing the relationships between varieties in a realistic way.

5.7 Conclusion

It is probably fair to say that, since the publication of Chambers and Trudgill's *Dialectology* in 1980, there has been something of a revolution in the way dialectologists have come to view relationships between varieties. Beginning with the works of Séguy and Goebl, there has been an increased emphasis on quantifying these relationships in objective ways, and a wide range of methods have been developed for comparing varieties on all linguistic levels, measuring the similarities and differences between them, and further analysing these measurements to reveal the patterns they encode.

Chambers and Trudgill (1980: 112) raised the following issue, which at the time had not been answered and probably was not easily answerable:

> It is undeniable that some isoglosses are of greater significance than others, in the sense that some mark distinctions 'felt' to be culturally important while others do not, some persist while others are transitory, and the like. It is equally obvious that some bundles are more significant than others, in the same sense. Yet in the entire history of dialectology, no one has succeeded in devising a satisfactory procedure or a set of principles to determine which isoglosses or which bundles should outrank some others. The lack of a theory or even a heuristic that would make this possible constitutes a notable weakness in dialect geography.

In this chapter, we suggest that basing an analysis on obvious isoglosses and bundles of isoglosses in a two-dimensional geographical space is less than ideal, and it is here that Chambers and Trudgill's problem partly lies. When many features and localities are compared at the same time, and when the relationships between them are quantified, we are in a better position to judge whether particular relations are more important than others, and to assess which features are contributing most to the similarities and differences between varieties. There are issues with these approaches, not least in determining the validity of the results they produce, but quantificational approaches give us new ways of analysing the relations between dialects in objective ways and, combined with innovative ways of representing these relationships, they promise to answer many questions that traditional approaches leave open.

5.8 Where next?

For a foundational examination of the use of isoglosses in analysing dialect relations, see Chambers and Trudgill (1980: chs. 2, 7, and 8). For a general examination of what is involved in quantifying relations between linguistic varieties and of representation of these relationships in tree and network form, see McMahon and McMahon (2005). For more specific details of methods for quantifying relations between varieties, see the sections above which deal with the different methods: good starting points in English are Goebl (1993, 2006, 2007) and the articles in the appendices to the second volume of *The Computer Developed Linguistic Atlas of England* (Viereck and Ramisch 1997). Heeringa (2004) is an excellent, wide-ranging, though rather technical, discussion of the issues and methods involved in quantifying relationships between dialects. It also provides a good entrance point into the rapidly increasing literature on the subject. In addition, Nerbonne *et al*. (1999), Nerbonne and Kretzschmar (2003), Kessler (2005), and Nerbonne and Kretzschmar (2006) provide good overviews of many of the issues and methods covered in this chapter.

The techniques and methods discussed in this chapter also feed into several other chapters in this volume. In particular, objective and quantifiable measures of relatedness and similarity are key to the interdisciplinary concerns raised by McMahon in Chapter 11, and could well revolutionise some of the forensic issues discussed by Rock (Chapter 9). If measurable comparison of varieties is desirable and helpful, there is a corresponding need to extend these methods to other levels of linguistic analysis; so far, there has been much work on lexis and segmental phonetics, and far less on morphosyntax (though see Spruit 2006; Longobardi and Guardiano 2009; and Szmrecsanyi 2008) and prosody. These innovations also allow linguistic data to feed more reliably and on a more equal basis into interdisciplinary studies involving other kinds of data, whether these are geographical, social, anthropological,

or genetic; see for example www.cecd.ucl.ac.uk/home/. Progress depends crucially on refining our developing methods, on comparing and validating them, and on extending their reach across different aspects of the grammar; only by understanding and seeking to overcome their current limitations can we hope to build more reliable methods of comparison both for work within linguistics and for truly interdisciplinary collaborations.

6 Perceptual dialectology

Chris Montgomery and Joan Beal

6.1 Introduction

Perceptual dialectology is a discipline that investigates what language users themselves think and believe about language. It explores where people believe dialect areas to exist, and the geographical extent of these areas, along with how these people react to spoken language. Consequently, perceptual dialectology is 'speaker-focused', and informs linguistic accounts of how and why language varies. In England, study within a perceptual dialectology framework has been neglected until relatively recently. However, asking non-linguists directly about how and where language varies results in data that can be used alongside other sources in order to 'fill the gaps' in linguists' understanding of how language works.

In this chapter, we discuss the development of the field of perceptual dialectology. We place this area of study within a wider approach to the study of non-linguists' thoughts and beliefs about language, known (perhaps unhelpfully, as discussed below) as folk linguistics. We begin by discussing some of the first academic interest in non-linguists' beliefs about language and the beginning of a perceptual dialectology approach. We go on to discuss further the methods and findings of language attitude research, placing this area of investigation within the field of folk linguistics. We then move on to discuss recent developments in the field, both as a response to the perceived shortcomings of language attitude research, and as a continuation of a tradition of perceptual study. Finally, we introduce and discuss results of recent research undertaken in England.

We are very grateful to members of the Centre for Language and Communication research at Cardiff University whose collegiality and support has made it possible for us to write this chapter. We would also like to thank the editors of this volume who have also provided helpful and constructive feedback on previous drafts of the text. We are delighted to have benefited from their experience and insight. Finally, thanks to the Arts and Humanities Research Council for a period of funded research leave during which this piece was written (AHRC Reference: AH/G007926/1).

6.1.1 Folk linguistics

For many years there has been interest in what non-specialists believe about language, and what these beliefs may signify. Bloomfield termed these beliefs 'secondary responses', in an article in which he also claims that 'the most important [utterances about language] are those which are made in the systematic study of language' (Bloomfield 1944: 45). This underlines the fact that non-linguists' beliefs have been discounted by many linguists as unimportant, as arising from a lack of education or knowledge, and therefore invalid as legitimate areas for investigation. This oppositional view is recognised by Niedzielski and Preston, who state that 'linguists have generally taken an 'us' versus 'them' position' (2003: 1).

Whilst acknowledging this 'us and them' position, Hoenigswald (1966) called for the systematic study of folk linguistics.

> [W]e should be interested not only in (a) what goes on (language), but also in (b) how people react to what goes on (they are persuaded, they are put off, etc.) and in (c) what people say goes on (talk concerning language). It will not do to dismiss these secondary and tertiary modes of conduct merely as sources of error. (Hoenigswald 1966: 20)

Hoenigswald's 'folk' are non-linguists and language users who have no formal linguistic training, and Preston (1993: 334) claims that 'knowledge of the folk categories at every level serves not only folkloric, anthropological, and applied linguistic ends but also general linguistic ones'. We can see, therefore, a source of conflict between linguists on the 'side' of Bloomfield, who regard only the observations made by trained linguists as important in understanding how and why language functions, and those who take the position espoused by Hoenigswald and supported by Preston.

The respective dates of Bloomfield's and Hoenigswald's comments may give the impression that there was no interest in non-linguists' beliefs in the intervening period, but this is not the case. We will detail below not only well-known studies (such as those which developed the matched-guise test), but also others that provided the starting point for present-day perceptual dialectology.

6.2 Perceptual dialectology: situating the field

Perceptual dialectology has a relatively long historical pedigree in various countries including Japan and the Netherlands. Long claims that 'if not "born" in Japan, [it was] at least "raised" there' (Long 1999b: 199), but we can trace the 'birth' of perceptual dialectology to the Netherlands, which saw pioneering research in the 1950s (Rensink [1955] 1999). Since this date, many linguists have further contributed to the body of perceptual dialectology research, notably Preston (1981, 1989, 1999a), Long (1999a, b; Long and Preston 2002;

Long and Yim 2002), Inoue (1999a, b), and Niedzielski (Niedzielski and Preston 2003).

Situating the field of perceptual dialectology has proved difficult for some, not least because of the reservations discussed above. Some have had difficulties in defining where it fits into the wider field of language investigation, although illustration of this is provided in diagrammatical form by Preston (1999a: xxii–xxv; Niedzielski and Preston 2003: 26). Preston describes perceptual dialectology as '*a sub-branch*' of folk linguistics (Preston 1999b: xxiv, our italics), which focuses on non-linguists' beliefs and perceptions. He proposes the following research questions (Preston 1988: 475–6):

a. How different from (or similar to) their own do respondents find the speech of other areas?
b. What do respondents believe the dialect areas of a region to be?
c. What do respondents believe about the characteristics of regional speech?
d. Where do respondents believe taped voices to be from?
e. What anecdotal evidence do respondents provide concerning their perception of language variety?

There have been many attempts to investigate these five questions. Although in the past perceptual dialectology has been neglected as an area of research in countries such as the UK, more recently several studies have specifically examined perception in this country (Inoue, 1999a, 1999b; Montgomery 2006). The development of perceptual study in the UK could be seen as a logical extension of Preston's interest in the discipline, which in turn could be viewed as a revival of 'traditional' perceptual dialectology research pioneered in Japan and Holland.

6.3 Traditional perceptual dialectology

The pioneering study in traditional perceptual dialectology was undertaken in the Netherlands. This saw the first systematic attempt to investigate perceptual dialect boundaries following a Dutch dialect survey undertaken in 1939. In what has become known as Questionnaire #8, two questions were present that asked informants first to state where people spoke *the same* dialect as them, then to answer a subsidiary question about dialect difference (Rensink [1955] 1999: 3). The resulting data were analysed by Weijnen (1946) who devised the 'little-arrow method' (Preston 1999b: xxvi). This involved the use of a map with a network of arrows to show the relationships between villages and towns where there was a perceived dialect link. These showed the extent of the perceived dialect similarity, and the final maps illustrate in a relatively clear way how the informants viewed relationships between language varieties in the Netherlands. This interest in the perception of dialect similarity was sustained in the Netherlands,

and led to Kremer's (1999) investigation into the Netherlands–Germany border as a perceptual dialect boundary. Also resulting in a little-arrow map, this study identified a number of phenomena that would be of interest to future linguists investigating perceptual dialectology and folk linguistics. Investigations of such things as 'barriers in the mind' have been the primary focus of subsequent perceptual surveys in Korea (Long and Yim 2002) and along the former east–west border of re-unified Germany (Dailey-O'Cain 1999).

The 1950s also saw the study of perceptions in Japan, although there had been an interest in the area since Tôjô (1927). The resurgent interest in perceptions of dialects by researchers such as Sibata ([1959] 1999) was inspired in part by perceptual work on dialects in the Netherlands, as well as Tôjô's (1927) investigation of the perception of dialect boundaries in Japan. The methodology of Japanese perceptual dialectology was fundamentally different from that used in the Netherlands in that informants were asked about 'grades' of *difference* along a continuum (from 'not different' to 'incomprehensible'). This being the case, the little-arrow method could not be used in the Japanese studies; instead a system of drawing lines between areas to indicate a scale of difference was implemented (Mase [1964] 1999). This system of drawing lines was the first method of 'calculating' perceptual boundaries. Finding that these subjective difference boundaries did not correlate to production isoglosses led some linguists in Japan to dismiss the findings as irrelevant (Weijnen 1999).

Despite these Japanese linguists' views, linguists in other countries (along with some in Japan) became interested in how non-linguists actually distinguish between language varieties, a question still of importance today. Indeed Butters (1991) supports a view that perceptual dialectology has value in that '[it raises] the question of just how much dialectologists' supposedly scientific determination of dialect areas may be artifacts of the dialectologists' own cultural bias' (Butters 1991: 296). For example, the Japanese studies seemed to illustrate the importance of school districts, as well as natural and political boundaries (Nomoto [1963] 1999) in the perception of dialect boundaries. Nomoto also investigated perceptions of grammar, vocabulary, and pitch accent and compared perceptual maps with these (Nomoto [1963] 1999: 88–96). Although vocabulary, segmental phonetics/phonology, and grammar (to an extent) are the 'traditional' measured components of dialectologists, the investigation of pitch accent as one of the ways in which non-linguists distinguish between different varieties was an interesting development, illustrating that perceptual dialectology can provide alternative explanations to those supplied by 'mainstream' dialectology and sociolinguistics.

6.4 Language attitude studies

The development of research into language attitudes in the early 1960s played a major role in shaping contemporary perceptual dialectology. Language

attitudes, or 'the attitude which speakers of different languages or language varieties have towards each others' languages or to their own language' (Richards, Platt and Platt 1992: 199) are of obvious importance in the study of perception as well as being useful for any investigation of dialect variation (Clopper and Pisoni 2002: 271–6). It was quickly discovered that language attitudes are real, can be tested, and are worth testing (Agheyisi and Fishman 1970: 139). These tests have typically examined two types of language attitudes: conscious and unconscious. Conscious attitudes are investigated when the informant knows that the questioner is asking about language attitudes, whilst unconscious attitudes can be measured when the informant is unaware of this. Fasold (1984: 149) describes the methods used for gaining access to the two attitude types as 'direct and indirect', and gives examples of a 'totally direct method' (simply asking informants their views on a language/variety), and a 'totally indirect method' (not letting informants know their attitudes are being investigated).

The most successful and enduring methodologies for investigating language attitudes stemmed from studies performed by social psychologists (Lambert *et al*. 1960). In an experiment investigating listeners' 'evaluational reactions to English and French', recordings of four bilingual men were made. Each man read a passage of French prose, and then read the same passage translated into English. In addition to these four participants, two more speakers were recorded (one reading in English, and the other in French) in order to offer "filler" voices and for practice for listeners. Lambert *et al*.'s experiment thus had ten voice samples for listeners to evaluate, eight of which were 'matched', with 'each speaker using both languages' (all citations Lambert *et al*. 1960: 44). The method in Lambert *et al*.'s landmark study has subsequently become known as the 'matched-guise' technique, and it was quickly seized upon as particularly productive for investigating language attitudes and became a mainstay of the field. The use of speakers who assume 'guises'[1] allows controllability, ensuring that the researcher can eliminate any attitudes that the listener may have towards voice quality or other variables inherent with different speakers. Of course, for the matched-guise technique to be successful, the speaker must be particularly competent in the guises he or she assumes. Concerns about this methodology have been expressed due to the 'alleged artificiality' (Edwards 1982: 22) of the technique (Agheyisi and Fishman 1970: 139). These concerns find support from Labov, who has expressed doubts about whether a speaker can master more than one dialect (1972b: 215). Preston has also expressed reservations about the effectiveness of the matched-guise technique, arguing against the 'gross, stereotypical imitations of varieties' used in such studies (Preston 1999c: 369).

Despite these concerns about the matched-guise methodology, it has been used by many linguists since it was first adopted. Although, in view of the possible problems with the methodology, we may want to be cautious about some

of the results, the matched-guise method has uncovered some points of interest. Principal findings have revealed divergent attitudes to perceived 'standard' and 'non-standard' varieties, along with a 'general tendency to relate linguistic standardness with intelligence' (Ryan and Giles 1982; Clopper and Pisoni 2002: 273). Despite this, however,

> [a]ccents judged as showing high speaker *competence* need not always have greater influence upon listeners than regional varieties seen to reflect more speaker *integrity* and *attractiveness*. (Edwards 1982: 24, italics in original)

Studies of this type therefore demonstrate that speakers 'can and do make a number of attitudinal judgments about a talker based on his or her speech' (Clopper and Pisoni 2002: 273).

Further studies have adapted the matched-guise method in order to investigate how effective listeners are at perceiving different accents, and how good speakers are at imitating accents. Markham (1999) performed such a study, asking eight native Swedish speakers to read an unfamiliar passage in a number of different accents (such a technique, moving on from single speakers assuming guises, will be called the 'subjective reactions test'). The results were then played to linguistically trained listeners who were asked to rate the reading on 'naturalness and purity'. The results showed that in some cases talkers could convincingly imitate accents of Swedish, which perhaps goes some way to reassuring those concerned about the matched-guise test as a methodological approach. Subjective reaction tests have perhaps been most widely used by Giles and Bourhis (1976; Paltridge and Giles 1984). The first of these studies (reported in Giles 1977) involved informants listening to tape recordings of different speakers in Cardiff. The study examined racial categorisation with twenty-four listeners hearing tape recordings of local, 21-year-old, working-class speakers. The listeners were again required to judge the speakers 'on a number of measures, one of which was a racial categorisation item' (Giles 1977: 9). The study produced interesting results, with the major finding that 'second generation West Indian adults in Cardiff [were] misattributed as Whites 75% of the time' (Giles 1977: 10). Many other studies have been carried out using matched-guise or subjective reaction tests, mostly with great success.

6.5 The response to language attitude research: contemporary perceptual dialectology

The methodological developments in the field of language attitudes led directly to renewed interest in perceptual dialectology, partly in response to the perceived failings of language attitude research. These failings were noted by Preston, who claimed that 'language attitude research did not determine where

informants thought regional voices were from' (Preston 2002a: 51). He claimed that traditional language attitude research did not assess whether informants had a 'mental construct of a "place"', or allow determination of 'their mental maps of regional speech areas' (Preston 2002a: 51). Thus, as linguists were examining non-linguists' evaluative responses to various voices (in matched-guise or matched-guise-type tests), they were not examining where informants thought the voices came from. This criticism is what we might expect from those committed to investigating folk linguistics – it may be obvious to the trained linguist where (geographically or socially) voice samples have been taken from. What traditional language attitude research failed to do, however, was to assess whether this was also obvious to the non-linguist listener.

In recent years some attempts have been made to rectify this specific shortcoming of traditional language attitude research, notably by Kerswill and Williams (2002), who asked informants to name which town or city they believed voice samples came from, Diercks (2002), who included a ranking task in order to assess dialect difference, and Clopper and Pisoni (2005), who requested informants to make a completely free choice and group voice samples in whatever order they wished.

Despite the undoubted value of the studies listed in the previous paragraph, we will focus here on Preston's (1999a) approach to perceptual dialectology. Preston's early research in the field was designed to address the shortcomings of language attitude research discussed above and was performed in ignorance of the perceptual studies carried out in Japan and the Netherlands. As such, although Preston's recent work acknowledges traditional perceptual work, there is little similarity between the approaches. Wales (2006a) describes Preston as 'the major proponent' of perceptual dialectology and the procedures he has proposed have become a methodological benchmark from which subsequent work in the field has advanced. Some of the approaches Preston advocates arise from an interest in perceptual geography, which was the inspiration for many of his techniques.

6.5.1 Perceptual geography

Preston stated that 'it has long been the case that maps may not represent the physical or political reality of the terrain' (Preston 1989: 13), and in this sense the academic discipline of perceptual and cultural geography is not a new one. Some of the earliest academic interest in environmental (geographical) perception came with the publication of Lynch's *The Image of the City* (1960). Lynch asked people their feelings for major American cities' landmarks and the routes they used to travel around them (Gould and White 1986: 12). In doing this, he was able to build up an image of the city held by his informants; an idea that was used in the 'City-Scene' project in the English city of Birmingham

(Goodey 1971a). In the Birmingham exercise, readers of the *Birmingham Post* were invited to draw maps of the city. The aim was to obtain 'quick, unaided impression[s]' (Gould and White 1986: 12) of informants' images of the city which were then combined. The results of the 'City-Scene' projects demonstrated the importance of lay-persons' views in an applied subject, supporting one of Preston's central justifications for this type of study.

Research by the cultural geographer Orleans (1967, 1973) used a mental mapping technique and focused on a wide range of informants in Los Angeles. He found that knowledge of the 'imagery of the urban environment might vary amongst distinctive groupings of urban residents as well as from one site (or location) in the city to another' (Orleans 1973: 118). Orleans' 'distinctive groupings' referred to the different ethnicity of residents in various parts of the city. In this study white, upper-class informants from a well-off suburb of Los Angeles (Westwood) had a 'very rich and detailed knowledge' (Gould and White 1986: 17) and were the only informants who could provide 'a well-formed, and generalised image of the entire Los Angeles Basin' (Orleans 1973: 118). This detailed knowledge was in stark contrast to Spanish-speaking informants from the centre of the city whose knowledge was 'confined to a few city blocks' (Orleans 1973: 118).

The methodological approaches taken by Goodey (1971a) and Orleans (1967, 1973), who asked non-experts to draw maps, directly influenced Preston's approach to creating some of the components for the study of perceptual dialectology. In addition to this, further findings of cultural geographers such as Gould and White (1986) are particularly helpful when studying non-specialists' perceptions, including discussions of the relationship between social interactions, place in society and perception of the local area.

6.5.2 Preston's methodology

Having studied the literature on perceptual and cultural geography, Preston found that many of the techniques used in this discipline could be translated to the field of perceptual dialectology. In his early work, Preston utilised a modified version of the techniques used by Ladd (1970) and Orleans (1973) and asked his informants to construct a hand-drawn map of where they believed dialect boundaries to exist. Preston claims that the value of these hand-drawn maps is not simply the profit to be gained by examining individual maps. The real value is in the ability to generalise the findings of many maps in a single composite map (after Goodey 1971a), thus creating 'perceptual isoglosses' (Preston 1999c: 361). The creation of these perceptual isoglosses mirrors the early perceptual dialectological work carried out in the Netherlands and Japan, and can then be analysed in a similar way (e.g. by examining the correlation (or lack of it) between perception boundaries and production boundaries).

However, Preston decided to analyse his composite perceptual isogloss maps in a different way, computerising them and introducing methodological components modified from cultural geography (Gould and White 1986), such as rank ordering (Preston 1999c: 363).

Preston believes that the investigation of language attitudes in addition to boundary perception is particularly worthwhile as '[u]nlike classic matched-guise attitude studies, this research provides informants with the category name and mapped outline of the region rather than actual voice samples' (Preston 1999c: 368). Advantages of this include the fact that the rated regions are 'cognitively real' (Preston 1999c: 368) to informants. This does, however, depend to an extent on an informant knowing what a specific space's dialect (as defined by the previous perceptual survey) sounds like in order to give a judgement on it. Preston claims that 'there is little or no difference in evaluations where the stimulus is a category name or … speech sample' (Preston 1999c: 369). Preston does concede that this method does not answer one particularly pertinent question: whether or not informants can actually identify varieties, although Montgomery (2006) has attempted to provide a method for doing just this.

Preston refined his complete methodology for the investigation of non-linguists' perceptions of regional varieties over many studies carried out between 1981 and the present day. Preston's first volume of the *Handbook of Perceptual Dialectology* (1999a) contains in its introductory chapter a five-point approach to the study of perceptual dialectology (Preston 1999b: xxxiv):

1. *Draw-a-map*. Informants draw boundaries on a blank (or minimally detailed) map around areas where they believe regional speech zones exist.
2. *Degree of difference*. Informants rank regions on a scale of one to four (1 = same, 2 = a little different, 3 = different, 4 = unintelligibly different) for the perceived degree of dialect difference from the home area.
3. *'Correct' and 'pleasant'*. Informants rank regions for correct and pleasant speech.
4. *Dialect identification*. Informants listen to voices on a 'dialect continuum', voices are presented in a scrambled order, and informants are instructed to assign voices to an area.
5. *Qualitative data*. Informants are questioned about the tasks they have completed and engaged in open-ended conversations about language.

This five-stage approach contains two important methodological additions to the study of perception the first of these being the inclusion of the fourth 'dialect identification' task. This, as mentioned above, must be of value when investigating perception. It allows the researcher to ask questions about *how* informants perceive variation, and not simply whether they do. The innovation of asking informants to identify dialects also attempts to address the shortcoming

of language attitude studies mentioned above. Also added is the fifth component, which looks very sensibly at qualitative data, exploring informants' perceptions and attitudes in a less formal way as well as providing production data if conversations are recorded.

6.5.3 Perceptual study in the UK

Historically there has been little interest in the study of perceptual dialectology in Great Britain (although there has been a good deal of folk linguistics study, in the shape of language attitude research). This is now changing, with modifications to the Survey of Regional English (SuRE) methodology (Burbano-Elizondo 2006: 116; Llamas 1999), and Montgomery's (2006) research (discussed below). There are exceptions to this lack of prior interest, starting with Inoue, whose two mid-1990s investigations on perception in England and Great Britain (collected in Preston, 1999a) follow research in Japan which examined perceptual dialect boundaries and regions based on 'dialect image'.

Inoue (1999a) uses multi-dimensional scaling analysis to plot and group dialects together, producing from this a dialect image map of Great Britain. The country is revisited by Inoue in later research, in an investigation of the effectiveness of hand-drawn maps for an examination of perceptual dialect boundaries. Inoue again uses this technique alongside multi-dimensional scaling analysis. Inoue urges caution in drawing conclusions from subjective perception maps due to the gap in the layperson's knowledge of the dialects of a language, and states that 'people often form dialect images even without listening to the actual dialect' (Inoue 1999b: 174), although this is surely an interesting phenomenon.

The study undertaken in England by Kerswill and Williams (2002) into dialect recognition by three speech communities is also of interest to the perceptual dialectologist. The study's focus was the process of dialect levelling, or 'the loss of localised features … to be replaced with features found over a wider region' (Kerswill 2003: 223), and when and where this occurred. Kerswill and Williams used a modified version of Preston's fourth dialect recognition task. The main aim was to investigate whether informants listening to recordings of speech from Hull, Reading, and Milton Keynes (plus four control locations) (Kerswill and Williams 2002: 181) could place where the voices were located geographically. The investigation concluded that dialect levelling plays a significant role in the recognition of dialects even where there are strong local networks. A strikingly frequent identification of older Reading speakers as rural 'West Country' suggests a 'perceptual dislocation' of the town's older accent as 'a consequence of rapid social changes in the town' (Kerswill and Williams 2002: 202). Although Kerswill and Williams do not include a draw-a-map task,

their discovery of a problem of identification in 'dialect-levelled' areas could have implications for studies that rely on informants' ability to draw an effective map. The phenomenon of levelling might also play a role in the type of maps informants draw, and perhaps reflects the convergent nature of certain varieties.

In comparison to other countries, then, the UK has not been well served by the field of perceptual dialectology, with other areas of linguistic study predominating. As mentioned above, this situation is slowly changing with apparent acknowledgements that perceptual study has the ability to answer questions which pose problems for mainstream research.

6.6 Studying perception in northern England

Montgomery's (2006) study built upon the growing international interest in contemporary perceptual dialectology as proposed by Preston (1999a), and situated it in a geographical area (the north of England) which has received a good deal of attention from sociolinguists, dialectologists, and scholars in the fields of cultural studies and human geography (see, for instance, Russell (2004), Dorling *et al*. (2005), and Wales (2006b)).

6.6.1 Methods

Montgomery (2006) adapted Preston's five-point approach for his investigation of perception in the north of England. The study involved participants from three locations in the north of England: Carlisle in the far north-west, Hull in the east and Crewe in the very south of the region. His research took place in two distinct stages: the first was the draw-a-map task, which gathered data from participants both on where they believed dialect areas to exist as well as where they believed a north–south divide to exist (see Wales (1999) for a discussion of this important 'divide' in England). After this first stage, the results were analysed and composite maps were produced. These composites were used in the second stage of research, which involved the rating and placement of voices from eight locations within and outside the north: Barnsley, Newcastle upon Tyne, Warrington, (West) London, Liverpool, Hull, Preston, and Carlisle. The locations of participants in the tests and those who provided voice samples are marked on the map in Figure 6.1.

The draw-a-map task was the primary method used to gain access to non-linguists' perceptions. Preston's approach was adapted for subsequent perceptual tasks, such as the ratings element of the methodology, which was also map based and incorporated Preston's fourth task. Informants were presented with a composite map of their responses to the draw-a-map task (echoing work performed by perceptual geographers) and were requested to rate each

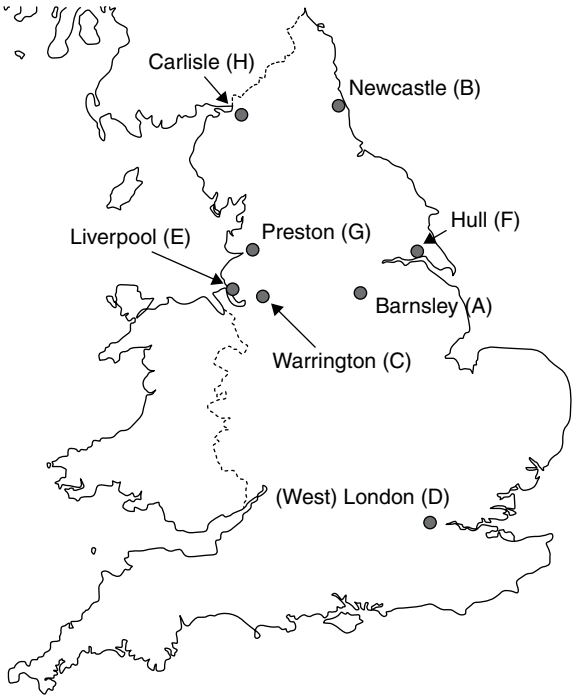


Figure 6.1 Provenance of voice samples

area along the scales suggested by Preston. Figure 6.2 shows one informant's hand-drawn map.

In order to answer one of the key questions of perceptual dialectology, participants were also requested to listen to recordings of eight voice samples and rate them along the same scales. Participants were further requested to use a blank map to indicate where they believed the voices came from. In this way, participants' ratings of 'cognitively real' (Preston 1999c: 368) dialect areas (in the form of the composite perceptual maps) along with their ratings of voice samples, could be examined side by side in order to establish similarities and differences. The map-based voice placement task permitted an examination of whether the freely drawn dialect areas corresponded with non-linguist informants' ability to recognise and place dialects from voice recordings.

6.6.2 Data processing

The two stages of Montgomery's (2006) research were processed in different ways, each technique developed to produce data that could be statistically

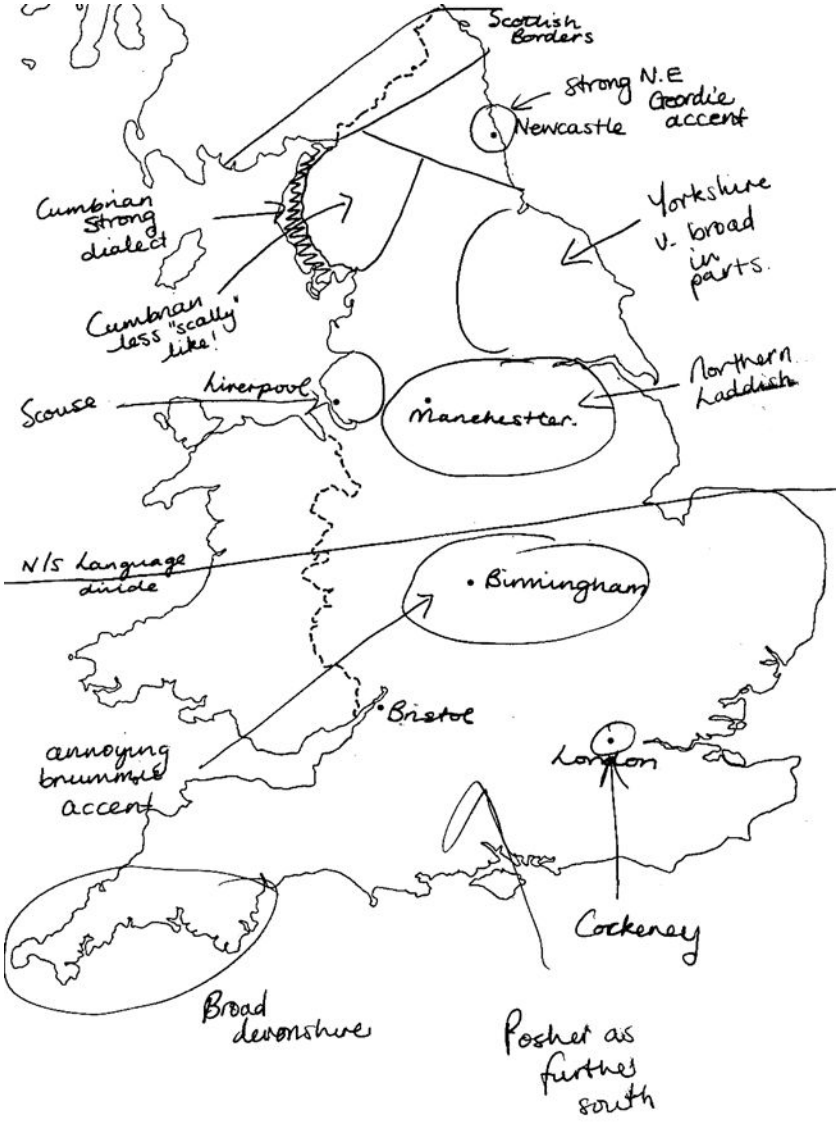


Figure 6.2 Carlisle informant's hand-drawn map (Female, 24)

analysed. The first stage involved the production of composite 'gradient' (Long 1999a: 181) perceptual maps. These maps gave a visual representation of where informants drew dialect areas, along with the amount of agreement amongst informants over the extent of the areas. One of these composite maps

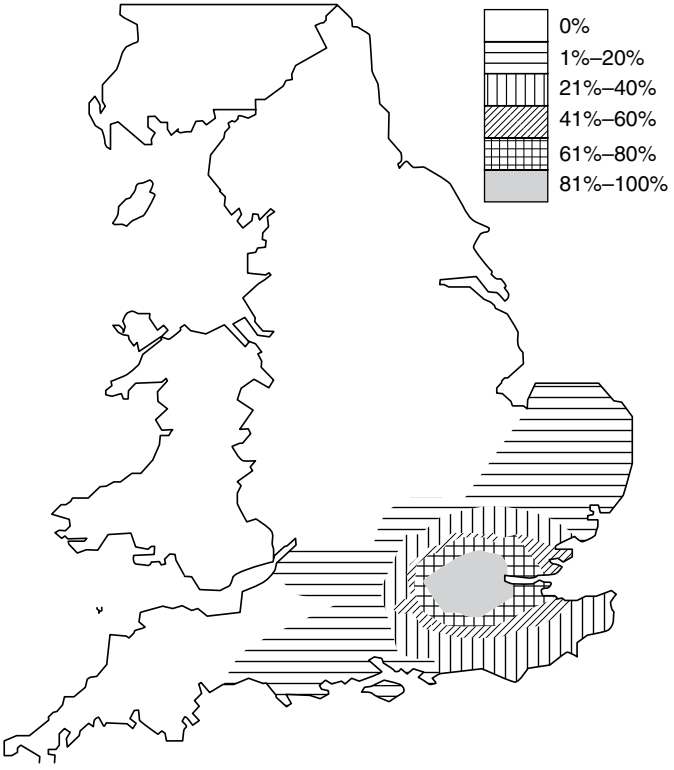


Figure 6.3 Hull informants' 'Cockney' dialect area ($n = 21$) (from Montgomery, 2006: 211)

is shown in Figure 6.3: the percentage shading relates to the percentage agreement amongst the informants drawing the area on the map.

Montgomery's (2006) percentage shaded maps, after Preston (1999c: 362) and Long (1999a: 188), were combined into an overall composite map. The map, which can be seen in Figure 6.4, provided a visual representation of the perception of dialect areas, and also how they interacted with each other. The two steps of the data processing for hand-drawn maps enabled analysis of not only the visual patterns but also the numerical make-up of the perceptual data, some of which is discussed below.

The data from the second fieldwork stage (voice sample location) were processed using the *starburst* method, which drew inspiration from techniques used in the Romanian Online Dialect Atlas (Embleton *et al*. 2007). This method enabled the collation of all the voice sample placements by informants for a specific voice sample, and the placement of these onto a chart that clearly

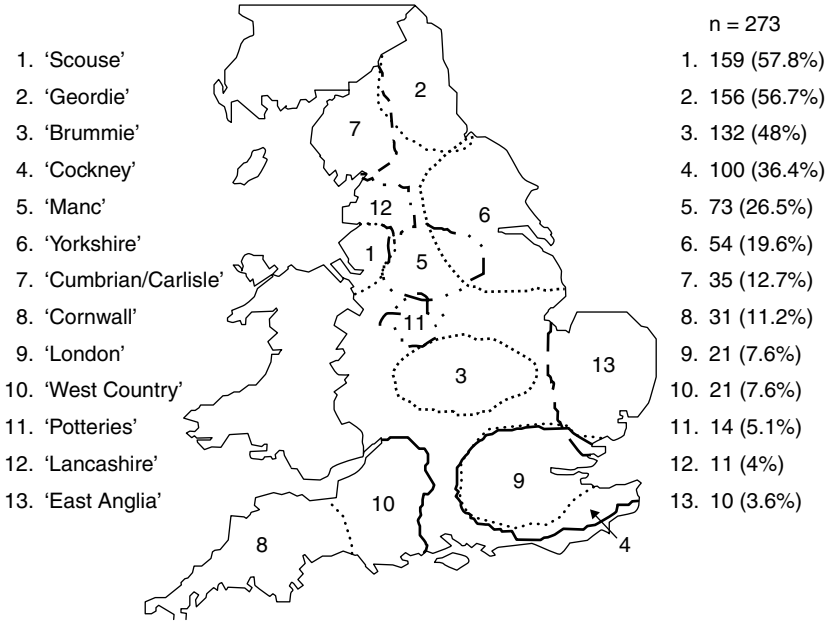


Figure 6.4 Overall composite map, indicating thirteen most recognised dialect areas by informants from all survey locations[2]

displays the relationship between actual provenance and perceived provenance. A starburst chart can be seen in Figure 6.5. Each concentric circle on the chart represents twenty-five miles (40 km) (except the lighter/dotted lines at the very centre of the chart, which represent five-mile intervals). The chart is organised according to points of the compass, also indicated. The actual provenance of the voice sample is the centre of the chart (marked with a dark dot), and the ends of lines radiating from the centre indicate where informants placed the sample. At the bottom right of each starburst chart is a 'mean error' value, which indicates the mean error of a perceptual task involving placing cities on a blank map. This value can help the reader of the chart establish how accurate it might be. Two further dots are shown on the charts (labelled 'A' and 'B' in Figure 6.5), dot A (which is also pointed to by the arrow in the chart) is the actual home location of the informants who attempted to place the sample. The dot labelled 'B' indicates the mean co-ordinate error of placements (calculated by taking the co-ordinates of each sample placement and averaging this).

As well as providing a visual guide to the extent and direction of voice sample placements, the starburst chart can also be used to calculate the degree of placement error. This again permits the use of statistical methods to further interrogate the data. The ability to gain statistical data about voice sample

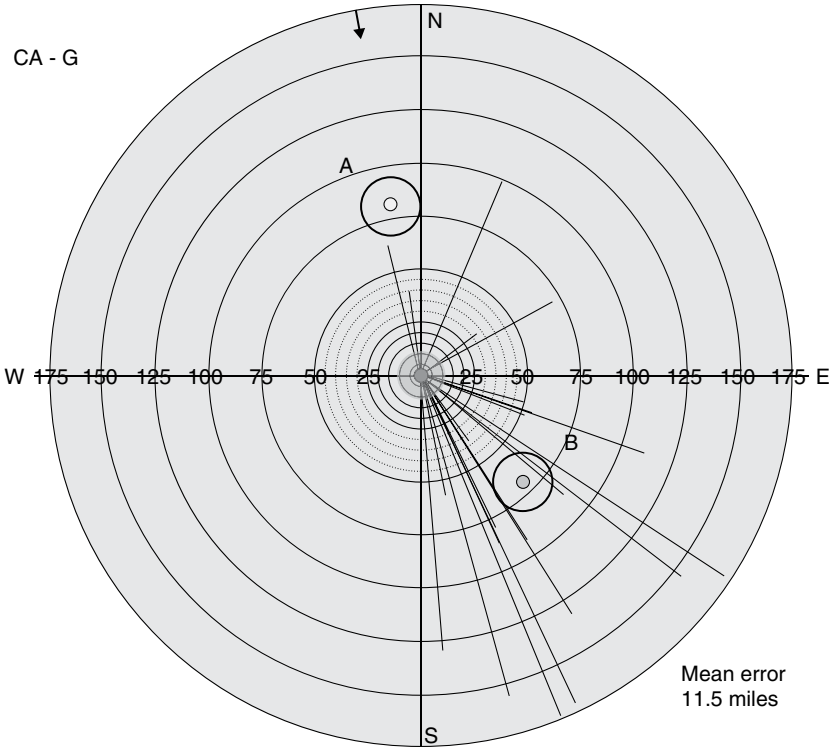


Figure 6.5 Starburst chart of Carlisle informants' placements of voice sample taken from Preston

placements was also useful in terms of comparisons between the ratings of samples and their placement, as discussed below.

6.6.3 Findings

We will concentrate here on overall findings from Montgomery's (2006) research, focusing on broad findings that indicate the presence of specific constraints on the way in which non-linguists perceive dialectal variation in England. These constraints are 'cultural prominence', 'claiming and denial', and 'proximity'.

Cultural prominence Cultural prominence, the salience of certain population centres in the national consciousness, played a major role in perception in Montgomery's (2006) study. There was no relationship between the size of population of a city and the level of recognition by informants, as demonstrated

Table 6.1 *Population of cities on which dialect areas were based by informants (population data from www.visionofbritain.org.uk) (from Montgomery 2006: 204)*

City-based dialect area	Total lines drawn	Population in 2001
Liverpool (Scouse)	159	439,476
Newcastle (Geordie)	156	259,573
Birmingham (Brummie)	132	977,091
London (Cockney)	100	7,172,036
Manchester (Manc)	73	2,482,352
Carlisle (Cumbria/Carlisle)	35	100,734
Stoke (Potteries)	14	240,643

Table 6.2 *The ten most frequently identified dialect areas by survey location (from Montgomery 2006: 196)*

Carlisle (*n*=98)		Crewe (*n*=85)		Hull (*n*=93)	
Area	Number	Area	Number	Area	Number
Geordie	52 (53.1%)	Scouse	67 (78.8%)	Scouse	44 (47.3%)
Scouse	48 (49%)	Geordie	61 (71.8%)	Geordie	43 (46.2%)
Brummie	34 (34.7%)	Brummie	61 (71.8%)	Brummie	37 (39.8%)
Cumbria	33 (33.7%)	Cockney	46 (54.1%)	Yorkshire	33 (35.5%)
Cockney	33 (33.7%)	Manchester	33 (38.8%)	Cockney	21 (22.6%)
Manchester	26 (26.5%)	Cornwall	16 (18.8%)	Manchester	14 (15.1%)
Cornwall	10 (10.2%)	Potteries	13 (15.3%)	London	10 (9.3%)
Yorkshire	9 (9.2%)	Yorkshire	12 (14.1%)	South West	9 (9.7%)
West Country	9 (9.2%)	London	9 (10.6%)	Hull	6 (6.5%)
Lancashire	8 (8.6%)	West Country	7 (8.2%)	East Anglia	6 (6.5%)

by Table 6.1. Table 6.2 (giving details of the most frequently recognised perceptual areas) provides some context for these figures.

Support for the phenomenon of cultural salience can be found in the data for the 'Manc' dialect area, which displayed a relatively high level of recognition, as Tables 6.1 and 6.2 demonstrate. This contrasts with the lack of recognition of a Manchester-based subjective area in the only other perceptual study in the UK (Inoue 1999b: 167). This is despite the presence of other city-based areas in Inoue's study, whose draw-a-map task was based on a map with county boundaries marked (1999b: 168). Montgomery concluded that the high recognition level for the 'Manc' dialect area could be due to an increase in Manchester's popular cultural prominence since Inoue's

study took place. We believe that this is a persuasive explanation for the increase in recognition from seemingly nothing (or nearly nothing) in 1989 (when Inoue's data were collected) to a recognition level of 26.5 per cent in 2005 (when Montgomery's data were collected). Cultural prominence also helps to explain the high recognition of the 'Geordie' dialect area, as despite the relatively small size of Newcastle in terms of population, its recognition level was particularly high (56.7 per cent), second highest amongst informants. These data demonstrate that population size is of no matter to informants when drawing salient dialect areas; what is of issue is the prominence of the area. This is not to minimise the importance of proximity (see discussion below) in the perception of dialect areas; however, salience is something that can be shown to be a major contributing factor in overall perception.

Claiming and denial Drawing on theories proposed in Hogg (1992), Williams, Garrett and Coupland (1999) discuss *claiming* and *denial* in relation to the recognition (and misrecognition, respectively) of a number of voice samples taken (mostly) from locations in Wales. Williams *et al*. (1999) found that a higher 'likeability' score could result in a particular sample being placed in the informants' home area, seemingly despite other factors. In Williams *et al*.'s (1999) research, one Cardiff speaker was *claimed* by many informants from different locations around Wales.

> Although he is correctly identified as a Cardiff speaker by the majority [of informants] … he is also the most claimed: 26.7% of the northeast listeners claim that he is from northeast Wales, 18.2% of the southwest Wales judges say he is from southwest Wales, 16% of Mid-Wales listeners think he is from Mid-Wales, and 10.3% of Valleys listeners claim that he is from their dialect community (Williams *et al*. 1999: 356)

The *claiming* and *denial* phenomena appeared to be in evidence in the voice sample placements in Montgomery's study. Although Montgomery (2006) differed from Williams *et al*. (1999) in the methods used in order to elicit placement information (a free map-based task versus an 'allocation' task, respectively), there still appeared to be a relationship between ratings and placement.

Figures 6.6 to 6.8 show the placements of the Warrington voice sample, and Table 6.3 demonstrates the mean rating of this voice sample by survey location. To aid in reading the charts, the additional dots indicating mean co-ordinate error and home location of listeners have been labelled as in Figure 6.5.

The Warrington voice sample was the most accurately placed voice sample of the eight that were played to informants, and was placed with an overall mean error of just 42.9 miles. There were no significant differences between

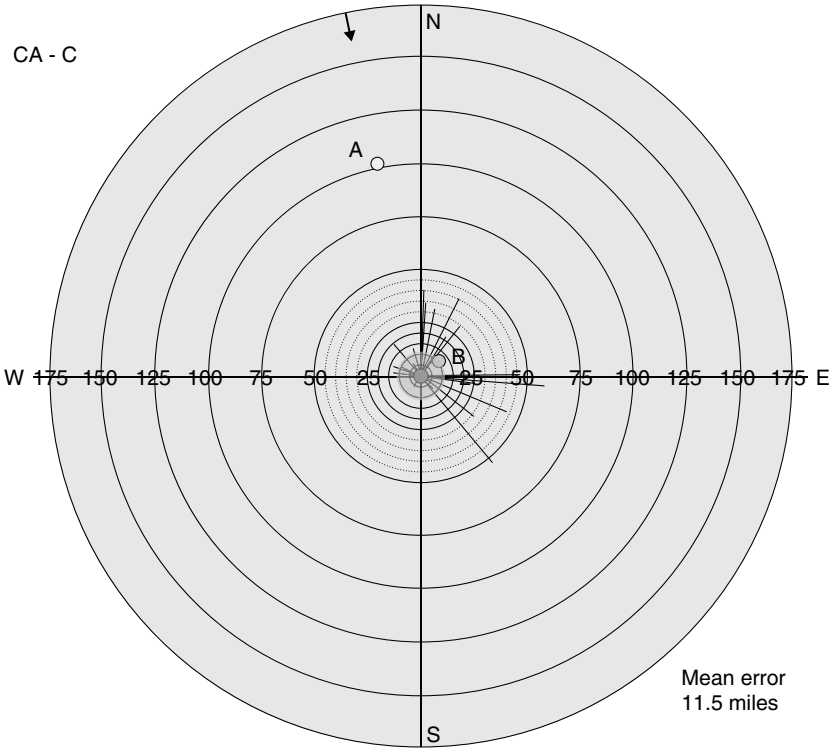


Figure 6.6 Carlisle informants' placement of Warrington voice sample at ≥21% agreement level (n = 27, mean error = 40.9 miles)[3]

the mean errors of placements due to this low overall mean, and starburst charts for Carlisle and Hull share similar placement distributions, with no 'lightbulb effect' (a skewing of placements towards a particular location, indicating *claiming*). The placements by Crewe-based informants, however, did seem to be affected by the towns' relative locations, as did the mean ratings, displayed in Table 6.3.

Examination of Figure 6.7 reveals a partial 'lightbulb effect' for the Crewe-based informants, with some of the placements of the voice sample skewed towards the location and the mean co-ordinate error of placement within the mean error factor. Taken with the ratings data, this could be a case of *claiming*, whereby Crewe-based informants perceive the sample as an example of a 'home' variety due to the direction of the placement errors, many of which run past Crewe or terminate around the approximate location of the town (21 miles south-east of Warrington). The ratings data in Table 6.3 also provided support

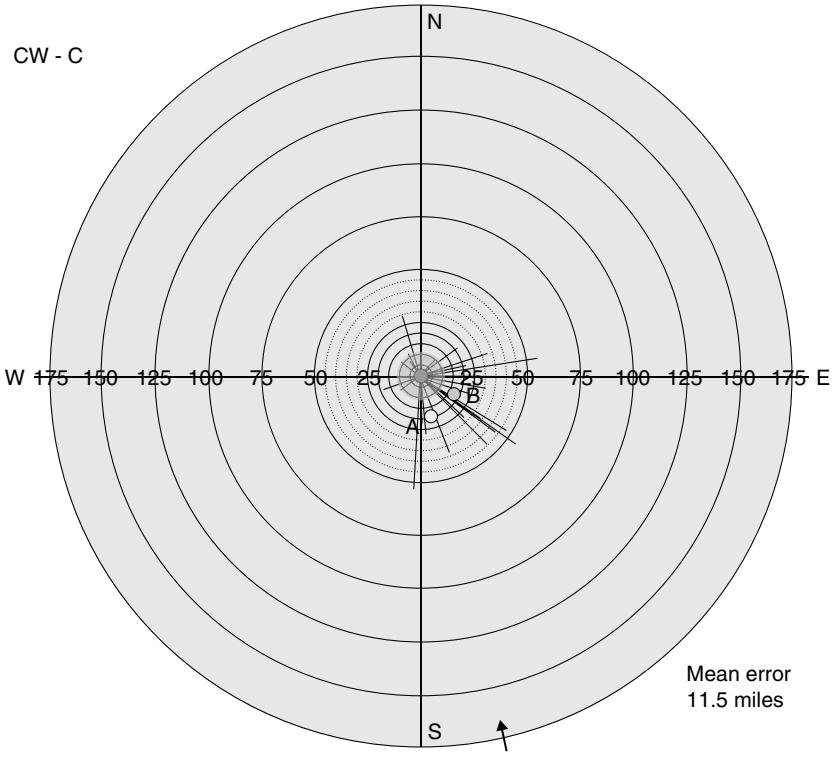


Figure 6.7 Crewe informants' placement of Warrington voice sample at ≥21% agreement level (n = 33, mean error = 39 miles)

for the *claiming* of the Warrington voice sample. There were no significant differences between the mean ratings from Carlisle and Crewe, which could perhaps be due to the location of the sample west of the Pennines. However, when comparing the mean ratings from Crewe-based informants with those from Hull, there are significant differences for all ratings with the exception of 'Correctness'. The mean for all ratings (excluding 'Difference') is also shown to be significantly different between ratings from Crewe and Hull ($p < 0.05$). This indicates that the voice sample from Warrington was judged significantly more favourably by informants from Crewe than those based in Hull. Not only was the sample viewed more favourably by Crewe-based informants but they also judged it to be significantly more similar (less different) to their own variety ($p < 0.05$). These factors, taken together, create a strong case for the *claiming* of the Warrington voice sample by informants from Crewe.

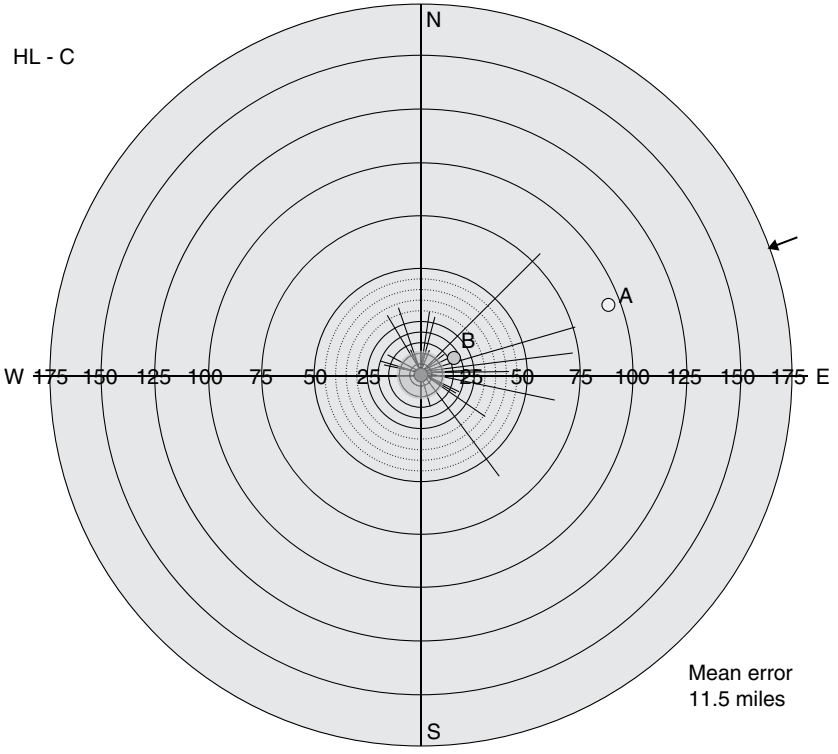


Figure 6.8 Hull informants' placement of Warrington voice sample at ≥21% agreement level (n = 29, mean error = 49.1 miles)

Montgomery's (2006) research can perhaps shed more light on the phenomenon of *denial* than of *claiming*, and in this way can build upon Williams *et al*.'s (1999) findings. Figures 6.9 to 6.11 show the placement data for informants' placements of the voice sample taken from Carlisle. Table 6.4 displays the ratings of the voice sample by survey location.

The Carlisle voice sample might only be notable for the sheer inaccuracy of its placement by informants from all of the survey locations. However, if one also examines Table 6.4 it can be observed that the Carlisle sample had low scores for most of the ratings scales. Indeed, the sample was the lowest rated of all voice samples played. These low ratings may have been due to informants' difficulty in placing the sample; the results are so inaccurate that a discussion of the placement errors can be restricted to noting that informants from all survey locations show a south and east skew in their placements of the sample. This is a relatively universal pattern and was the clearest placement

Table 6.3 *Mean ratings for the Warrington voice sample along ratings scales for each survey location, with significant differences flagged*

	Carlisle	Crewe	Hull
Correctness	5.37	5.83	5.47
	CA-CW	CW-HL	CA-HL
P	NS	NS	NS
Pleasantness	5.29	5.47	4.58
	CA-CW	CW-HL	CA-HL
P	NS	**0.05**	NS
Difference	5.29	4.53	5.6
	CA-CW	CW-HL	CA-HL
P	NS	**0.05**	NS
Friendliness	5.17	5.51	4.67
	CA-CW	CW-HL	CA-HL
P	NS	NS	NS
Trustworthiness	4.85	5.23	4.41
	CA-CW	CW-HL	CA-HL
P	NS	NS	NS
All (w/o diff.)	5.17	5.51	4.79
	CA-CW	CW-HL	CA-HL
P	NS	**0.05**	NS

Significant differences are the result of one-way ANOVA tests with Tukey's HSD post hoc tests run on all voice sample placement data in SPSS 14 for Windows. CA-CW compares Carlisle and Crewe scores; CW-HL compares Crewe and Hull scores; CA-HL compares Carlisle and Hull scores.

skewing witnessed in Montgomery's (2006) investigation. Of particular interest was that the Carlisle voice sample was the 'home area' sample for informants from Carlisle. Even so, the pattern of wide inaccuracy in placement could simply have been due to these informants not recognising their home sample. However, this did not appear to be the case.

Taking the 'Difference' scale in this instance, it can be seen that Carlisle informants recognised that the Carlisle voice sample was the least different from their own variety (and therefore the most similar). The mean rating is significantly different from that of both Hull- and Crewe-based informants ($p < 0.05$). Despite this clear and significant acknowledgement of least difference, Carlisle-based informants exhibit the greatest mean error in voice placement (170 miles). This appears to be a clear case of *denial*; whilst simultaneously identifying the sample's similarity to their own accent, the informants from Carlisle attempted to place it as far away as possible. It is difficult to account for this, other than by the fact that the Carlisle voice sample has such low overall ratings that even its home informants do not want to be associated with it. Williams *et al*. offer a potential explanation for the *denial* phenomenon

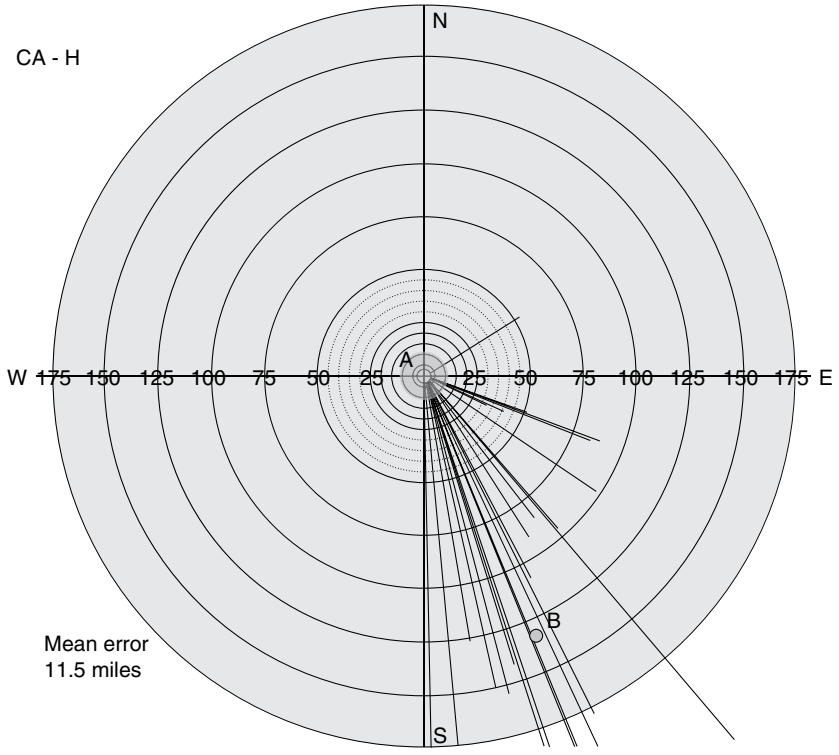


Figure 6.9 Carlisle informants' placement of Carlisle voice sample at ≥21% agreement level (n = 26, mean error = 170 miles)

here in terms of 'social attraction … [being] the set of prototypical properties of the group' (1999: 356). Competition and innovation, they claim, 'establish and maintain a relatively positive evaluation of one's own group'. As a result, 'in-group prototypes are generally evaluated positively' (1999: 356). In Montgomery's (2006) ratings and placement task, then, informants were (as in Williams *et al*.'s study) 'able to decide for themselves whether the speakers they heard were in-group or out-group members' (Williams *et al*. 1999: 357). So, for informants from Carlisle, they could recognise the voice sample as being similar to their own (in a city-wide sense), but did not wish to be associated with the speaker.

Williams *et al*. claim that 'processes such as claiming and disavowing are an intrinsic part of dialect recognition processes' (1999: 358), something which Montgomery's study supports. The inclusion of a free-choice voice-placement task alongside the more traditional ratings task allows the researcher to

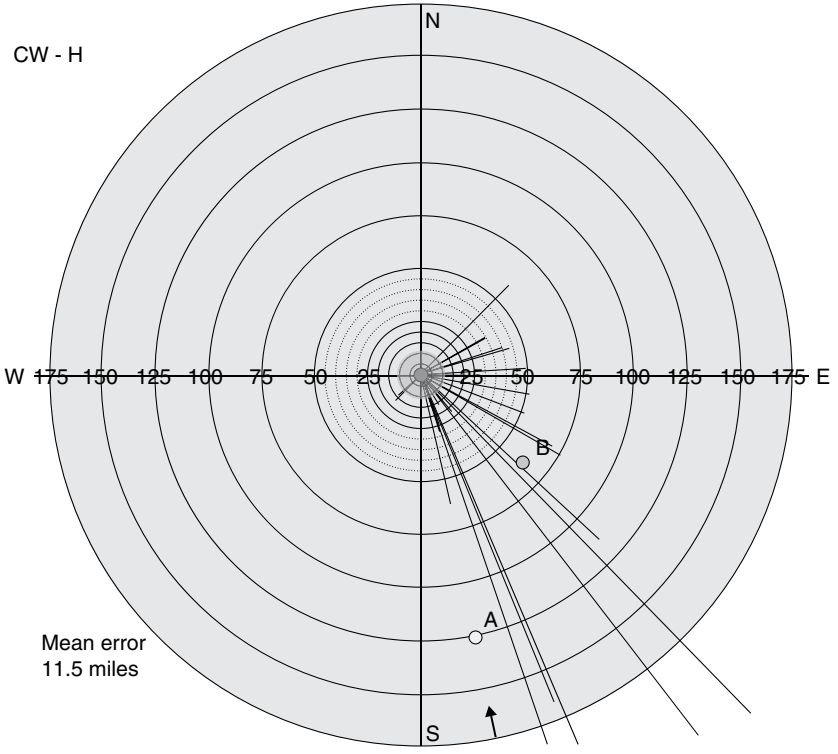


Figure 6.10 Crewe informants' placement of Carlisle voice sample at ≥21% agreement level (n = 26, mean error = 110.6 miles)

investigate these phenomena and assess not only geographical cognition but also '*social* cognition' (Williams *et al*. 1999: 357, italics in original).

Proximity In Montgomery's (2006) study, *proximity* was taken to mean the 'closeness' to an area. The idea is that closer proximity would enable informants to distinguish a greater number of dialect areas, or be more accurate in their recognition of boundaries. Following Preston's findings that after drawing stigmatised areas, informants would draw 'local areas more frequently' (1999b: xxxiv) in draw-a-map tasks, proximity did indeed seem to be of great importance to informants in this study. Its effects were noted in the draw-a-map task (both in the country-divisions element and in the area recognition/delimitation element) and, to an extent, in the dialect recognition task. There is some interaction in the case of proximity with cultural salience, discussed

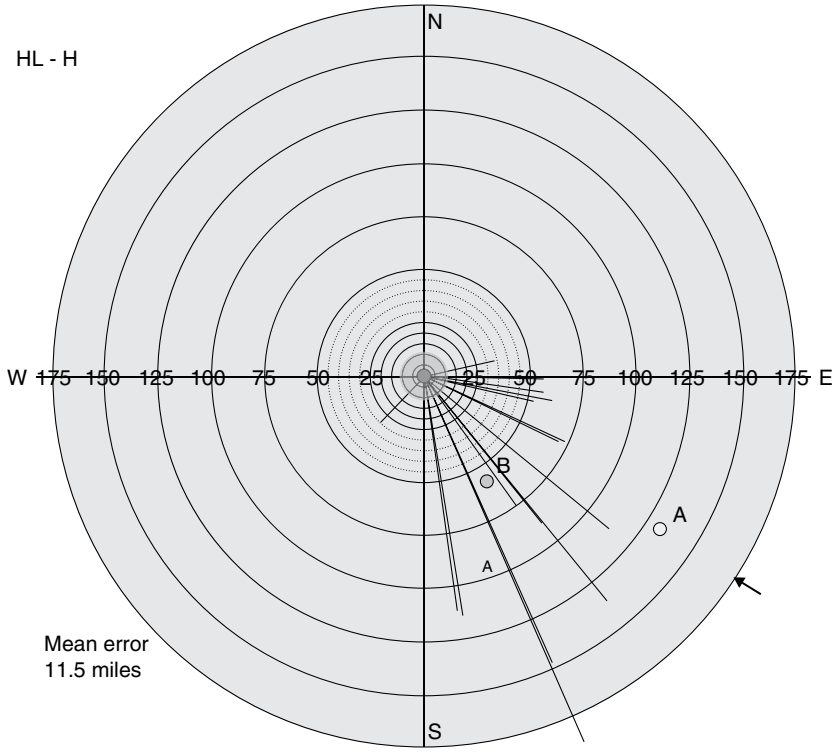


Figure 6.11 Hull informants' placement of Carlisle voice sample at ≥21% agreement level (n = 22, mean error = 109.6 miles)

above, and the ideas underlying proximity effects find support from perceptual geographers such as Goodey (1971b).

In a diagram, Goodey (1971b: 7) demonstrates the importance of 'near to' places, which are shown as 'personal space'. 'Far places' are beyond the limits of personal experience, only registering through the mechanisms of 'radio places', 'talked of places', 'TV and film places' and 'printed places'. If we add 'Internet places' to the mechanisms of experience, Goodey's schema provides a useful way of understanding the roles of proximity and cultural salience in perception.

The ability to 'experience' far places is of great importance in the perception of language as in many cases it is differences rather than similarities which are noted by language users, although Preston (1999b: xxxv) does note that close proximity will allow informants to make more detailed distinctions, with fewer detailed distinctions made in the case of 'far-off' varieties. This finding carries with it an

Table 6.4 *Mean ratings for Carlisle voice sample along ratings scales for each survey location, with significant differences flagged*

	Carlisle	Crewe	Hull
Correctness	4.19	4.32	4.60
	CA-CW	CW-HL	CA-HL
P	NS	NS	NS
Pleasantness	3.86	3.92	3.51
	CA-CW	CW-HL	CA-HL
P	NS	NS	NS
Difference	4.86	6.45	6.38
	CA-CW	CW-HL	CA-HL
P	**0.05**	NS	**0.05**
Friendliness	3.56	3.69	3.47
	CA-CW	CW-HL	CA-HL
P	NS	NS	NS
Trustworthiness	3.74	4	3.48
	CA-CW	CW-HL	CA-HL
P	NS	NS	NS
All	3.84	3.98	3.76
	CA-CW	CW-HL	CA-HL
P	NS	NS	NS

acknowledgement that 'far-off' varieties will be perceived as 'different' or 'very different' and a fine reading of the systematic way in which the varieties differ will be difficult. This will not be the case for 'near-to' varieties in which a much finer reading will be enabled through the listeners' prolonged exposure to them. Of course, the perception of 'far-off' varieties will only be enabled if, through the mechanisms listed above, the varieties are present on the listener's 'radar' (i.e. if they are culturally salient). If the varieties are not present on the 'radar' then the informant will most likely not draw any areas.[4] Goodey's (1971b: 7) diagrammatic representation of perception offers a good way of understanding the relationship between proximity and cultural salience, and is of great use in explaining why certain varieties are more prominent than others.

There is a good deal of evidence from Montgomery's study that supports Goodey's inclusion of personal space at the centre of human perception. Table 6.5 shows the total number of lines drawn representing each 'home' area, and the striking effect of survey locations on the 'near-to' dialect areas.

The table shows the number of lines drawn representing each area, the bracketed figures indicating the percentage of the total that the lines drawn represent. It shows the clear influence of survey location on the identification of the 'home' area, and thus the importance of proximity on the perception

Table 6.5 *Number of lines drawn representing 'home' dialect areas, by survey location*

Dialect area	Carlisle (% of total)	Crewe (% of total)	Hull (% of total)
Cumbria/Carlisle	33 (94.3%)	1 (2.9%)	1 (2.9%)
Potteries	1 (7.7%)	13 (92.3%)	0 (0%)
Yorkshire	9 (16.7%)	12 (22.2%)	33 (61.1%)

(or at least, the recognition) of dialect areas. Even for the 'Yorkshire' dialect area, which is recognised by informants in both Carlisle and Crewe, the Hull-based informants account for nearly two-thirds of the lines drawn indicating the area.

Montgomery (2006) also demonstrated that proximity played a role in the perception and placement of voice samples in the ratings task. Significant relationships were found between some of the mean placements and survey location which showed that 'near-to' samples would in some cases be more correctly identified than 'far-off' samples.

6.7 Where next?

We began this chapter with a discussion of the opposition between linguists, such as Bloomfield, who have tended to discount non-linguists' views of language and those who, like Hoenigswald (1966) see the value of systematic study of language attitudes and perceptions. Within sociolinguistics, this opposition has been less marked: Labov included subjective reaction tests in his pioneering study of New York City (1966) and quantitative sociophonetic studies often use attitudinal data from interviews, albeit in a rather post hoc manner, in the discussion of their results. More recently, Llamas (2007) and Burbano-Elizondo (2006, 2008) have brought issues of place and local identity to the forefront of their research, the latter including perceptual dialectology methods as a complement to other data-elicitation techniques in order to provide a more nuanced account of language variation, change, and identity than has hitherto been provided by quantitative studies of variation and change. The innovations introduced by Montgomery (2006) and outlined above will allow researchers to investigate more systematically the attitudes and perceptions that might influence language variation and change.

The most important and comprehensive accounts of perceptual dialectology to date are Preston (1999a) and Niedzielski and Preston (2003), and any research involving perceptual methods should use these as a starting point. However, the innovative methods introduced by Montgomery (2006) and

outlined in this chapter should also be taken into account. This is a field which we expect to develop further as technical innovations in mapping are developed and as interdisciplinary investigations of language, place, and identity emerge. Research projects may be conducted on a small scale, investigating perceptions of linguistic and social divisions within a single city, as Finnegan's (forthcoming) study of Sheffield has done, or may involve the study of larger areas. The map tasks described in this chapter enable researchers to collect data fairly quickly from large groups of informants, and have been used successfully with classes of undergraduates and with school students of various ages in Britain, but they could be adapted for use in other countries and for investigations of other languages.

Part II

Why does it matter? Variation and other fields

7 Variation and linguistic theory

Patrick Honeybone

7.1 Introduction

It may seem surprising, but linguistic variation is often seen as a 'problem' for linguistic theory. The models that formalist, theoretically minded linguists work with typically assume that linguistic behaviour is categorical and idealise away from the variation that is found in speech. The justification for this, following Chomsky (1965), is that much of the variation found in utterances is due to non-linguistic factors, and thus idealisation is necessary in order to see the underlying patterns behind speakers' linguistic performance. A number of strands of work in theoretical linguistics have, however, sought to take linguistic variation seriously, and they form the topic of this chapter, along with the argumentation that arises when linguistic theorists talk about (or refuse to talk about) linguistic variation.

It's no secret that languages like English are full of variation. If illustration is needed, let us consider a simple sentence like (1), which might describe a woman giving her coat to her brother.

(1) Betty took off her coat and gave him it.

If we limit ourselves to syntactic and phonological variation (as I do throughout this chapter), we could imagine a number of ways in which speakers of English might utter (1), or something very close. As a speaker of English born and raised in the English East Midlands, I could easily utter (1), but I could also utter (2).

(2) Betty took her coat off and gave him it.

Is (2) the same sentence as (1)? It would be true under the same set of circumstances and it features the same set of words, so let's assume that it is. This means that a speaker from the north-west of England, for example, would also be uttering the same sentence if they said (3), which is how they, among others, might prefer it (as Siewierska and Hollmann 2007 explain).

(3) Betty took her coat off and gave it him.

The fact that (1), (2) and (3) are all possible in English plainly shows that syntactic variation exists, and variation at the phonological level is also

unavoidable. That same speaker from England's north-west might well pronounce *Betty* as [bɛtɪ], *her* as [ə] and *off* as [ɒf]. Some speakers from that area, however, particularly if they came from Lancashire, might pronounce *her* as [əɹ]. Indeed, the same speaker might sometimes pronounce it as [ə] and sometimes as [əɹ]. A speaker from the north-west of the USA, on the other hand, would certainly have a rhotic pronunciation of *her*, like the Lancashire speaker, but would likely pronounce *Betty* as [bɛɾi] and *off* as [ɑf].

Such examples can be multiplied manifold, as any speaker of English knows. *Betty* would likely be [bɛʔi] for a speaker of London English, and many speakers from other parts of the UK might now vary between [bɛti] and [bɛʔi], with different types of populations favouring either the oral stop over the glottal or vice versa. Speakers from Liverpool, on the other hand, may realise the /t/ as a slit alveolar fricative, in a case of lenition, which we can represent as [bɛθ̠i], and speakers from Newcastle upon Tyne might pronounce the name as [bɛtʔi].

It is often said that linguistic variation occurs when one meaning can be attached to more than one form. This is clearly the case for *Betty*: it doesn't change the meaning if a speaker says [bɛti] one minute and [bɛʔi] the next, and it also seems right to say that 'gave him it' and 'gave it him' *mean* the same, but involve different linguistic forms. These are two cases of linguistic variables – single linguistic items ('meanings') which have multiple identifiable variants ('forms'). The variable (t) has all the variants described above, including [*t*] and [ʔ] and the variable (pronoun-object-order) has the two variants given here. In this piece, I focus on variation of this type, where one referent has more than one form, and where some sort of geographical, social or at least stylistic effect is associated with the different forms. I leave aside other ways in which language can vary (such as when a phoneme varies categorically in its allophones or a form varies diachronically over time).

How can this abundant variation be a problem for theoretical linguistics? In part, this derives from deep-reaching disagreements about what we mean by 'language'. There is an everyday meaning for that word, and it might seem to be obvious what we mean when we talk about 'English'. However, as we will see, neither of these notions is as straightforward as its everyday meaning might imply when we view it through the lens of linguistic theory.

There are two fundamental types of variation that confront us when we consider the notion 'a language like English', and both of them were exemplified above. I investigate this point further in §7.2. Sections 7.3 and 7.4 take these two types of variation in turn and discuss why they should matter to theoretical linguists, considering some of the methods that are used to analyse such cases of variation, and showing their relevance for linguistic theory. Section 7.5 concludes.

I use 'linguistic theory' here with its standard, restricted reference: that approach which aims to provide formal, concise statements concerning the

structural generalisations that can be made about language in general, or about individual languages. This 'theoretical linguistics' takes at least some impetus from the body of ideas associated with the generative linguistics of Chomsky. It can be contrasted with 'variationist linguistics', in the tradition of Labov (1966), which explicitly focuses on the ways in which speakers vary in their utterances, in terms of the number of variants that they produce for particular linguistic variables. This has shown that all languages are *inherently variable* (including cases of stable variation which can persist in a language for centuries) and that this involves *orderly heterogeneity* – speakers of similar backgrounds tend to consistently use the same proportion of variants of a variable: variation is not haphazard.

In what is to come, we will both see why much of theoretical linguistics does *not* really pay much heed to linguistic variation (seeing it as a problem which can reasonably be ignored, because the problem actually belongs to someone else – variationists), and consider some work which aims to integrate accounts of linguistic variation into formal linguistic theory (seeing variation as a problem to be solved).

7.2 Linguistic theory and the two types of variation that it needs to deal with

Language varies in a number of ways, but there are arguably two types of variation which are fundamentally distinct from each other, and which could have different implications for linguistic theory. Some of the variation in (1), (2), (3), and the realisations of *Betty*, *her* and *off* above, compares forms that are possible in *different* dialects of English, while other aspects of this variation refer to how a *single speaker* (of a single dialect) might realise the forms. The term 'variation' is thus ambiguous, and either (4) or (5) can be intended by it:

(4) variation between speakers = inter-speaker variation
(5) variation within a speaker = intra-speaker variation

Inter-speaker variation was illustrated above by the comparison between those speakers who might prefer the order of 'gave it him' and those who might prefer 'gave him it', and by the comparison between those speakers who might tap the /t/ in *Betty* ([bɛɾi]) and those who might glottal it ([bɛʔi]). Intra-speaker variation is involved in the cases where the same speaker might order the particle before the direct object in 'took off her coat' one moment, but might use the other order the next, and in the case where the same speaker might glottal the /t/ in *Betty* in one utterance, but might realise it as a plain [t] in the next.

If our aim is to investigate 'the linguistics of a language' such as English (as it surely is in a volume such as this) we need to consider *both* types of variation as they both exist in the phenomenon that we call 'English'. As we

will see when we consider them individually, however – (4) in section 7.3, and (5) in section 7.4 – their implications and the responses of linguists who have considered them are very different.

Theoretical linguistics is thus faced with two questions (two problems?): how should it deal with inter-speaker variation and how should it deal with intra-speaker variation? Theoretical linguists of different persuasions have argued that both or neither or only one of these two should be taken into account as they work to figure out the nature of speakers' grammars. It is probably fair to say that most work in linguistic theory does not see inter-speaker variation as relevant to theory construction. This position depends on answers to two questions, given here in (6) and (7), which go to the heart of a fundamental issue: the nature of the proper object of linguistic study.

(6) What do we mean by *language*?
(7) What do we mean by *languages*?

The broadly generative tradition in theoretical linguistics has clear answers to these questions. This tradition includes syntactic work in the narrowly Chomskyan frameworks from the standard theory through to minimalism, and also work in commensurable frameworks like lexical functional grammar and head-driven phrase structure grammar (HPSG). It also includes work in the generative strands of phonology, from standard generative phonology through to lexical phonology and government phonology, and mainstream work in optimality theory (OT) in both phonology and syntax. From within this tradition, Chomsky has long argued that it is crucial to recognise a distinction between different perspectives on (6) and (7). Building on an early distinction between *competence* and *performance*, which is itself reminiscent of Saussure's (1916) distinction between *langue* and *parole*, Chomsky (1986) proposed a distinction between *I-Language* and *E-Language*. I-Language is an aspect of *knowledge* – it is the set of linguistic generalisations that we know in order to be able to speak our language (or languages). For example, we know that in English, the plural is formed by adding an alveolar fricative which takes its voicing from the final segment in the morpheme that it attaches to, and we know that *wh*-phrases like *what* and *how many* occur at the start of clauses in questions. Like all types of knowledge, I-language is *i*-nternal to each *i*-ndividual. It forms the grammar that exists in the mind/brain of speakers. 'E-language' is the term Chomsky uses to refer to all the utterances that speakers make when they talk or write. This connects with everything else that the 'everyday sense' of *language* involves, including those things that are connected to language which are *e*-xternal to the speaker, such as the community that they belong to.

The standard generative opinion is that theoretical linguists should only focus on I-language, and thus the generativists' answer to question (6) is that 'language' only makes sense and can only be a coherently investigable entity

if we see it as a cognitive entity that exists in a speaker's mind. E-language is subject to influence from many different areas: physical pressures affecting linguistic performance, social and political pressures affecting interactions, and historical contingency in terms of what speakers happen to talk about. The claim is that the set of utterances that speakers make is not a well-formed or coherent object. It can't be enough simply to wait for occurrences of everything that is possible in a language in a collection of data: corpora of utterances cannot tell us what is impossible in a language, and so E-language is not a suitable object of study if we hope to figure out the generative system that a speaker knows. On this approach, the answer to question (7) is that 'languages' are I-languages, the grammars which exist in speakers' minds. The implications of this are quite considerable. Each speaker is an individual, of course, and no individuals share a mind. Each speaker's knowledge of language may be slightly different from every other speaker, as their experience in acquisition is different. Thus, while such grammars will need to overlap considerably in order for speakers to communicate with each other, we should not expect them to be the same. Each person has their own language.

All of this shows three things: (i) generative linguistics takes an inherently mentalist approach to language, situating it in the mind; hence, (ii) given that such linguists are required to focus on the knowledge in individual speakers' minds, it cannot make sense to focus on inter-speaker variation; and (iii) the everyday notion of what 'a language' like English is, is seen as incoherent. Regarding point (ii), Chomsky (1965) has famously written that '[l]inguistic theory is concerned primarily with an ideal speaker-listener in a completely homogenous speech community'. This deals with the 'problem' of (4) by identifying it as irrelevant to theory building, so that it should be ignored. The homogeneity that is idealised in order to focus on an individual's grammar might even be taken to exclude intra-speaker variation but, as we will see below, this is not necessary. To consider point (iii), if *languages* are I-languages, which are states of knowledge, there is no room for entities which correspond to the everyday notion of 'English' – something which is shared by millions of speakers, with both small and major variations between speakers or between groups of speakers who speak different English dialects. On this approach '[i]t is very doubtful that one can give any clear or useful meaning to the "everyday sense" of the term "language"' (Chomsky 1980), and the same is true of the term 'dialect'. English is an E-linguistic notion, and there is no point in doing 'the linguistics of a language', because languages do not exist. The only true linguistic object, on this approach, is the idiolect – each speaker's own grammar.

All of these points are challengeable, however, and the answers given above have been rejected by other linguists. While mentalism is the mainstream in theoretical linguistics, other approaches have situated language outside of the

mind by giving different answers to questions (6) and (7). An *instrumentalist* approach might argue that language does not exist in the mind because linguists' abstractions and constructs do not (need to) literally exist – theoretical linguists should simply produce the best analysis of the data that they are presented with and should not care where their generalisations exist, as long as they succeed in characterising the data in the most insightful way and allowing for prediction. Few linguists have overtly adopted instrumentalism, but 'post-Bloomfieldians' in the USA, and those following Firth in the UK have done so (see, for example, Honeybone 2005 and Carr 2006). However, some work on linguistic theory can only *seem* to claim a commitment to mentalism, by situating itself in a generative tradition, while the actual analyses proposed seem to hint at a covert instrumentalism. This issue is discussed briefly in §7.3.

More importantly, variationist linguistics has issues with the I/E-language distinction, because variation in production – 'E-language' – is the main focus of study. Labov and others who work in this field aim to discover the patterns that exist in the use of variants that we find in a whole speech community. On this approach, language – the object of linguistic study – exists in the speech community, rather than simply in the I-linguistic mental grammar of each speaker, and some variationist work therefore rejects the focus on I-language, focusing instead on inter-speaker variation, and answering questions such as: in those communities where /t/ can be glottalled, do females glottal more or less than males, or the young more than the old? There *may* also be a focus on intra-speaker variation in Labovian work, however, as we shall see in §7.4. Such work aims to model the variable performance of individual speakers.

In the next section, we shall see that, despite protestations to the contrary, inter-speaker variation can and has affected analyses within linguistic theory. This may rely on an overt commitment to a fundamental approach to the conception of language, or it may rely on an analyst simply aiming to come up with the best analysis of variable data.

7.3 Inter-speaker variation and linguistic theory

As §7.2 has shown, most work in theoretical linguistics sets aside inter-speaker variation, as it is not seen to provide relevant evidence for linguists in their quest to work out a model of an I-language. So can linguistic theory entirely ignore the potential problem of inter-speaker variation? One strand of work within generative linguistics has claimed, to the contrary, that linguistic theory should seek to accommodate variation of this type within the modelling of speakers' grammars. This is *generative dialectology*, which developed in the 1960s and 70s in such work as Newton (1972). While explicitly identified generative dialectology petered out by the 1980s, we shall see that the fundamental ideas that generative dialectology espoused have been and are adopted

in other linguistic work. While they are often rejected out of hand in their full-blown, explicitly stated version, aspects of them can creep into linguistic analyses, and there are resonances of them in a number of areas of theoretical work on English.

Although we have seen that languages such as 'English' have a problematic status in theoretical linguistics, most speakers have the perception that they typically understand other speakers of the thing that we call 'English', even though their speech patterns show considerable variation. Indeed, it could even be argued that languages exist mentally in speakers' perceptions of their close and extended speech community – who it is that they would at least try to speak to if they met. Given this, it is not unreasonable to wonder how speakers store the necessary knowledge of varieties of English that are different from their own. Is this knowledge stored as part of their own I-language? That might be the most economical option, so it is worth entertaining as an idea, and this is what generative dialectology and related approaches do: the grammars that are proposed in such models are meant to account for more than one dialect of a language, with only minimal changes to the analysis to account for the variation. Such approaches explicitly aim to account for inter-speaker variation.

This is a *polylectal* approach: one grammar should be able to generate the surface forms of more than one *lect* (that is, variety, dialect, sociolect, etc.) of a language. Some work from this perspective (such as Bailey 1973, 1996) argues that the best analysis should be *panlectal*, that is, it should account for every variety of a language (using the word in its perceptual, everyday sense). Polylectal approaches need to combine the fact that something among all the lects that are to be collected in a grammar must be the same (so that all speakers of all the lects have something in common) with the obvious fact of variation in utterances. The approach that has typically been adopted, in generative dialectology and in conceptually related approaches such as those of Bailey (1973, 1996) and Agard (1971), relies on the widespread analytical assumption that there is more than one linguistically important level. It is often claimed that there is an underlying and a surface level (or a deep and surface structure, or pre- and post-movement stages of a derivation, or an input and output) to any linguistic item, such as a sentence or phonological string. Linguists who adopt a polylectal approach to linguistic analysis have cohered around the idea that the lects of a language like English may vary considerably at the surface level, but are expected to be more similar at the underlying level. This makes sense, as it gives a common grammatical 'core' at the underlying heart of all dialects that are perceived to belong to a single language.

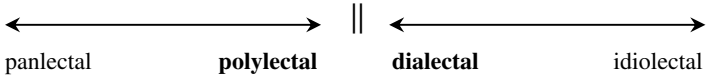
Most of the arguments in this area have been made from phonology, though, as Harris (1984b) shows in some detail, the same arguments can be had in syntax. This kind of argumentation fits easily into rule-based phonology, as there is often expected to be something of a gulf between the underlying

representation (UR) and the surface representation (SR). All that a polylectal approach needs to assume is that lects of a language differ in terms of the rules that they have (or perhaps in their ordering). In syntax, this type of analysis fits more clearly into early generative models, where surface forms were directly generated from deep structures, but it can be transferred to any syntactic (or phonological) model with 'earlier' and 'later' stages of a derivation. It fits less easily with optimality theoretic models, because the input is less important or firmly fixed than is the UR in rule-based phonology, for instance, but it is possible to imagine a model in which the input is expected to remain identical across lects, and only the ranking of constraints should be expected to differ.

Such polylectal approaches to linguistic theory clearly differ from the 'orthodox' approach to variation sketched out in §7.2, which emphasises the separateness of linguistic varieties to the extent that each speaker has their own dialect (indeed, their own language). Because of this, I propose to call the approach which contrasts with polylectalism a *dialectal* approach to linguistic analysis, because each dialect/language is analysed entirely on its own, such that dialects may be expected to vary at the underlying level just as much as at the surface level.

There are thus two *fundamental linguistic approaches* that theorists need to choose between in the face of inter-speaker variation: will they take a dialectal or polylectal approach? In fact, we can differentiate further within these fundamental approaches: both are really clines from extreme to less extreme positions, although there is a conceptual gulf between them. All this is schematised in (8).

(8)



An approach which is polylectal to any extent will write a grammar which in some way makes reference to or aims to account for more than one variety, but this could be only a very minor part of an analysis. Panlectalism, the extreme of this position, will seek to produce one underlying analysis to account for all dialects or varieties of a language, with only superficial surface variation, as we saw above. A dialectal approach will stress a need to ignore all varieties other than the precise one that a linguist is focusing on in any particular analysis. Much work with a commitment of this type does indeed claim to focus on the analysis of one single dialect of a language (in the everyday meanings of those words), analysing the phonology of Liverpool English, for example (as in Honeybone 2001), or the syntax of Belfast English (as in Henry 1995). Most of

the time this is not noticed because the variety in question is 'Standard English' or a reference phonological variety, such as General American (GA) or RP, but it is nonetheless only a step on the way to the extreme of this fundamental approach, which is an *idiolectal* approach. This would take the rhetoric of the I-language-only approach seriously, and focus on the grammar of only one speaker. Placement on one of these clines of fundamental linguistic approach does not have to be an all or nothing affair: it would be perfectly possible to adopt a polylectal approach to one aspect of phonology or syntax, for example, and a dialectal approach to another.

Consideration of these kinds of distinctions is not accorded all that much importance in most theoretical linguistics, and while the distinctions are discussed in some work, they are often quickly dismissed. However, it is arguably crucial to bear them in mind when thinking about the treatment of variation in English (or any other language), because it is otherwise easy to fall into the trap of taking one approach while intending to take another.

We might wonder whether all these types of approach are in any way defensible. Clearly we all have *some* knowledge of lects other than our own. But should this knowledge inform a speaker's grammar, or should we merely assume that it is 'tacked on' to our I-language as an additional aspect of language-related knowledge? On a mentalist perspective, panlectalism seems impossible to defend – a language like English has so many speakers, and so many varieties, that it is absurd to assume that any speaker's grammar could be informed by all varieties. It may be defensible under instrumentalism – if it allows for the best account of the regularities of all the data from English to assume a common underlying analysis for one aspect of its linguistic system, some linguists may be tempted to adopt that analysis (as we will see below when we discuss the English monophthong system). Most theoretical linguists will only entertain mentalism, however. A steady stream of work, such as Harris (1984b, 1985), McMahon (1992, 2000) and Lodge (2009), argues against panlectalism, and indeed, for the most part, against even the slightest hint of polylectalism.

Certain analyses that are common in the phonology of English have a somewhat polylectal flavour, however, and it seems likely that a similar situation holds for syntax. Take the English segment which is commonly transcribed as /r/, for example. This is subject to massive inter-speaker variation in SRs, as (9) shows.

(9)

- English English speakers typically use [ɹ]
- North American speakers may use [ɹ] or [ɻ]
- Scottish speakers typically use [ɾ] and [ɹ]
- (some) Northumbrian speakers (still) use [ʁ] (see Påhlsson 1972)
- an increasing number of speakers in (especially South-Eastern) England use [ʋ] (see Foulkes and Docherty 2000)

Nonetheless, all of these are typically assumed to be 'the same thing', by both phonologists and speakers: a realisation of English /r/. The very notion 'English /r/' is a polylectal, if not panlectal concept, and the fricative, tap and approximant articulations by which is it realised have little in common phonetically. This variation does not typically cause problems with comprehension between dialects. It could be argued that a polylectal analysis of /r/ in English makes most linguistic sense – all varieties have underlying /r/, which is realised differently in different dialects. This also shows that the URs assumed in a polylectal approach do not need to be taken from any one dialect: [r] is, of course, an alveolar trill, which is almost only ever used as an uncommon stylistic variant in English. Nonetheless, /r/ is widely seen as the formalisation of the underlying segment, with realisation rules required to map it onto the forms shown in (9). This type of situation is not so unusual. For example, although he is generally anti-polylectal, Lodge (2009) perhaps unintentionally assumes a similar situation for the 'English low front lax vowel' thus: '/æ/, realized as [æ] (RP), [*a*] (Manchester) or [ɑ] (Belfast) can be treated as the same vowel'.

Furthermore, the standard rule-based analysis of the kind of inter-speaker variation considered in §7.1 for the realisation of /t/ in *Betty* fits in well with a generative dialectological polylectal approach, as shown in (10), where the dialects involved are the same at the underlying level and only differ at the surface, due to the application of different processes.

(10)

Process	GA	London	Liverpool
underlying	/bɛti/	/bɛti/	/bɛti/
tapping	[bɛɾi]	—	—
glottalling	—	[bɛʔi]	—
lenition	—	—	[bɛθ̲i]
surface	[bɛɾi]	[bɛʔi]	[bɛθ̲i]

Other phonological models might implement this without rules, but could still retain a polylectal applicability (thus the essence of (10) is assumed in Harris's (1994) Government Phonology framework). This analysis also coincides with that of Bailey (1980), who uses precisely this type of data to argue for a panlectal approach. This is perhaps a coincidence of analysis, but it might also indicate a general acceptance of, or at least lack of surprise at, the idea that the UR for a linguistic item should be the same across dialects,

in the face of surface variation. It's surely true that we expect to find /t/ in every variety of English, just as we expect to find full series of stops, tense and lax vowels, modal verbs and *wh*-fronting, so when we encounter something that seems to correspond to any of these in a dialect which has not been well analysed, we rush to transfer these categories across to the new variety. This is a polylectal approach to analysis. Knowles (1973) bemoans this in his analysis of Liverpool English, and argues, on a truly dialectal approach (but likely incorrectly) that there are no underlying oral stops in that variety, but rather a different category corresponding to the obstruent in *Betty*, on the basis that there is massive lenition in the variety, affecting segments at all places of articulation; thus the fricative in *Betty* is typically found alongside fricatives in the final and medial obstruents in words such as *back*, *dock*, *tiger*, *lad*, *leader*.

The approach illustrated in (10) shows that an analysis which would be expected on a polylectal approach can appear far from absurd. Other such analyses are less easy to accept. What about the variation in *off* and *her* discussed in §7.1? A GA speaker might say [ɑ*f*] and [hɜɹ], while a London English (LE) speaker might say [ɒ*f*] and [ɜ], and a speaker of Scottish Standard English (SSE) might say [ɔ*f*] and [hɛɾ]. The analysis to derive all of these forms from common underliers is much more complicated. For instance, /h/ and /ɾ/ would need to be deleted from the London English form, and either [ɜ] derived from /ɛ/ or [ɛ] from /ɜ/ (all this can be done in powerful phonological models, of course). The vowel in *off* is even less straightforward. Because GA lacks a vowel contrast that is present in LE, certain occurrences of GA [ɑ] correspond with LE [ɒ], as in *off*, *plot* while others correspond with LE [ɑ], as in *father*, *smart*. This distribution is not predictable so *plot* and *father* cannot both have polylectal /ɑ/ as there would be no way of deriving the LE forms by rule. The contrast would need to be stored in the URs of all varieties, and merged on the surface by an absolute neutralisation rule in GA.

The same issue would arise for all cases where any variety of English has a contrast that other varieties lack, meaning that all contrasts in any form of English, even very dialectologically restricted ones such as /u~ɪu/ in Welsh English (as in *through* [θɹuː] vs. *threw* [θɹɪ*u*]), would need to be stored in URs for all speakers and removed on the surface by neutralisation rules. This is less compelling as an analysis than was (10), and this is tied to the fact that (10) is a quite concrete analysis, with little distance between URs and SRs, and with a small number of 'natural' processes in the derivation. Typically, the polylectal approach involves an increase in the abstractness of the URs that are needed, and a greater difference between URs and SRs, because the URs have to be mappable onto the SRs of several varieties. Dialectal approaches tend to be phonologically more concrete (and thus more appealing for those who view abstractness in analysis as a problem).

It may be that different types of dialect differences seem reasonable to different degrees under a polylectal analysis. One influential typology of dialect differences is that discussed in Wells (1982) (following others, including Trubetzkoy 1931). A major distinction is made here between *systemic* and *realisational* differences between varieties. Systemic differences occur when dialects differ in terms of the number of contrasting underlying segments that they have, so that there is a difference in the segmental systems of the dialects. Realisational differences occur when dialects have the same type of contrasting underlying segments, but realise them differently, so that there is a difference in terms of the number or nature of the phonological processes in the dialects. Those dialect differences which submit most reasonably to a polylectal analysis are realisational differences. This is the case with (10), where flapping, glottalling and Liverpool lenition are realisational differences. Systemic differences make much less sense under polylectalism: should we really assume that speakers have a large number of underlying contrasts which never surface? A polylectal account of systemic differences, such as English English /u~ʊ/ vs. SSE /ʉ/ (as in *pool* and *pull*), and Welsh English /u~ɪu/ vs. English English /u/ (as in *through* and *threw*) assumes that absolute neutralisation is a common phenomenon, and that we should allow it in one grammar on the basis of argumentation from a different grammar. This suffers from all the criticism levelled at absolute neutralisation (see, for example, Kiparsky 1968; Hooper 1976; Carr 1993; McMahon 2000), and given that phonology is increasingly surface-orientated (as in OT, for example, see Prince and Smolensky 1993/2004), it will persuade few.

The very notion that we should compare dialects (and/or accents) of one language, and that it is worth distinguishing typologies of dialect differences is a polylectal notion. It is common in work on the phonology of English, however – why compare lects of English, rather that varieties of English and varieties of German, if not because we assume that it will help us understand the structure of any one variety by *seeing it as* a variety of English? Wells (1982) is a towering example of this, which has been immensely influential in work on phonological variation in English, especially in Europe. Wells developed a basic framework for discussing the vowels of accents of English (consonants are dealt with somewhat differently) which relies on *standard lexical sets*, referred to by keywords, as in (11), which includes all the sets for the vowels that can occur in stressed syllables, in the order that Wells gives them (he numbers them, so that KIT = 1, DRESS = 2 *etc*).

(11) KIT, DRESS, TRAP, LOT, STRUT, FOOT, BATH, CLOTH, NURSE, FLEECE, FACE, PALM, THOUGHT, GOAT, GOOSE, PRICE, CHOICE, MOUTH, NEAR, SQUARE, START, NORTH, FORCE, CURE

The keywords stand for '(i) any or all of the words belonging to the standard lexical set in question; and (ii) the vowel sound used for the standard lexical set in the accent under discussion' (Wells 1982: 124). While they are intended to be supplementable by other sets if other contrasts are needed for a variety, this set of words can also be seen (and can sometimes be used in analysis) as a panlectal analysis of the English stressed vowel system – that is, the set of contrasts of which we expect to find at least a subset in any lect of English. Forms of the language involved are expected to fit into these predefined slots. Indeed, if we set out those sets which typically involve lax vowels (the first six in Wells' list), as in (12a), it looks very much like the symmetrical set of lax vowels in the basic English vowel system in Giegerich (1992), given in (12b).

(12)

(a)		(b)	
KIT	FOOT	ɪ	ʊ
DRESS	STRUT	ɛ	ʌ
TRAP	LOT	a	ɒ

In some sense, both of these systems provide an analytical framework which is English-specific, and which is applicable, with a small set of moderations (which are themselves built in to the system) to any and all varieties of English. The analysis of non-low tense vowels in English is another area where polylectal argumentation has played a role in well-known analyses. This is shown in (13), which includes the analysis from Trager and Smith (1951), which has been very influential in America. This proposes an 'overall pattern' for the phonology of English, into which multiple varieties should fit. All three systems in (13) expect all or most forms of English to have contrasts which will fit into their slots.

(13) Representations of the non-low tense vowels in English

Giegerich (1992)		Trager and Smith (1951)		Wells (1982)	
i	u	iy	uw	FLEECE	GOOSE
e	o	ey	ow	FACE	GOAT

The slots assumed in all three systems allow for considerable surface variation. Thus, to take the high back part of the vowel space, it is no problem to reconcile the fact that RP might have [uː] in words such as *goose*, *loot*, whereas SSE might have [ʉ]. This can be implemented by realisation rules: for RP,

all tense vowels are long, and for SSE, the high back tense vowel is fronted, and the Scottish Vowel Length Rule (see Aitken 1981; McMahon 2000) will account for its length. Again, this analysis fits with the polylectal approach rather well, and it would fit fine in the clothes of generative dialectology. It succeeds in showing that RP [uː] and SSE [ʉ] count phonologically as the same thing (the high, back tense vowel), and that we should expect speakers of the different dialects to be able to recognise the forms as realisations of what is phonologically functioning as the same thing. It may be that an analysis of this type imports a covert instrumentalism into linguistics: it is the best analysis to fit in with an ideal phonological form, which is symmetrical and economical. These are considerations which make for the neatest analysis, but we cannot be sure that they reflect the cognitive reality that mentalism aims for.

The discussion here has focused on phonology, but similar arguments could be imagined for syntax, as Harris (1984b) explores. Klima (1964) gives an early generative syntactic analysis (from the period when phonology and syntax were modelled in essentially the same way, on the basis of linguistic rules) which explicitly adopts polylectalism, dealing with a number of cases of variation including the differences in case-marking between varieties which allow (14) and those which allow (15). The difference is accounted for through a basic grammar, which generates (14), and which is supplemented by additional 'extension' rules in lects which allow for (15).

(14) He and I left.
(15) Him and me left.

We could also imagine a movement analysis for object order in ditransitives, which assumes an underlying or initial order such as that in (16), with the direct object closer to the verb, which can surface unchanged as in (3), and also the derivation of (17) in some dialects through movement (leaving a trace), deriving the surface spelt-out form as in (1) and (2), producing a polylectal analysis which shares the approach of those seen above for phonology (see Haddican, 2010).

(16) VP[V[give] DO[it] IO[him]]
(17) VP[V[give] IO[him]$_i$ DO[it] t_i]

It is thus certainly possible to conceive of polylectal analyses in syntax, but little work which overtly identifies with this position exists. A similar methodological approach to that identified for phonology above can certainly be expected, however, whereby an analyst simply assumes that the categories established on the basis of one set of varieties will be relevant in unresearched varieties of a language.

The above discussion shows that a broadly polylectal approach has been adopted in some work in theoretical linguistics, incorporating knowledge of more than one variety of a language into a single I-linguistic grammar, despite

the argumentation in §7.2. As we have seen, this has occurred both consciously and sometimes without intention, so that we need to be sure to interrogate an analysis to work out if there are polylectal assumptions behind it.

To sum up this section: how should theoretical linguistics deal with inter-speaker variation? Is a polylectal approach possible? Or should we embrace dialectalism? Or, indeed, idiolectalism? Should linguists allow data and evidence from one variety of a language to weigh on the analysis of another variety in any way at all? Consistent polylectalism (let alone panlectalism) seems problematic, and is widely derided in most work which considers it. There are a number of ways in which arguments stack up against the idea that one I-linguistic grammar can be fully shared across individuals. However, grammatical analysis need not take on or reject every aspect of polylectalism (or dialectalism) wholesale. As we have seen, several widely accepted analyses tend towards polylectal argumentation to some degree. It might yet prove to be a theoretically coherent position to argue for potential partial polylectalism in phonology: with realisational differences between accent candidates for a polylectal analysis, and systemic differences analysed dialectally.

Although it can have an impact on which particular analysis is seen as best for a particular linguistic phenomenon, little discussion exists of the fundamental linguistic approaches of analysts to the polylectal/dialectal distinction and the question of how to deal with inter-speaker variation. The role of intra-speaker variation is typically seen as more important in linguistic theory. We turn to this now.

7.4 Intra-speaker variation and linguistic theory

Like inter-speaker variation, intra-speaker variation in production is a fact of linguistic life. How should linguistic theory deal with it? Some theoretical linguists insist that grammar should only account for what's *possible* in language (such as a glottal stop or plain [t] as a realisation of /t/) and that something else must account for any variation that occurs in the realisation of a category (Hale and Reiss 2008: 139, for example, call this a 'post-grammatical processor'). This would mean that grammatical theory *should not* account for intra-speaker variation, and that it can be excluded from the argumentation used to work out an I-linguistic grammar. Most theorists who investigate such variation in detail disagree, however, arguing that the same theories that account for the categorical aspects of phonology or syntactic structure should also account for cases where there is intra-speaker variation.

Complex issues can arise when intra-speaker variation and linguistic theory are confronted. As Henry (2002), among others, explains, intra-speaker variation has a very uncertain place in the minimalist programme, the most recent incarnation of Chomskyan generative syntax: 'Chomsky (1995) proposed an

economy principle under which movement occurs only when it is forced to do so. Such a grammar explicitly excludes optionality. If something moves only if forced, it will be impossible in principle for there to be an internalized grammar in which any movement operation is optional. If the option not to move exists, movement will not take place, since it will not be forced to do so.' This means that, on a purely formal perspective, intra-speaker variation makes no sense. Minimalism assumes that language is economical, and that its design features should be as minimal as possible for a system that links sound and meaning. 'The question arises as to why such a system should have variation – this simply seems to add complications, both in terms of the syntax itself, and of the learnability of the syntax' (Henry 2002).

Speakers do vary in their performance, however. A single speaker may produce (1) or (2), or even (1), (2) and (3), and a single speaker might call the same person both [bɛti] and [bɛʔi] in different utterances. Such variation is crucial socially – it helps people mark out their identity, indicate their sense of belonging to particular places and social groups, and to accommodate to people who they want to identify with; but these are not structural linguistic issues, so they are not easily integrated with the purely formal entities that theoretical linguistics typically deals with, such as segments, syllables, syntactic phrases and functional heads, and some would argue that they absolutely should not be integrated in this way because of this ontological difference.

In what follows we shall (i) briefly consider the observations that have been made about intra-speaker variation, (ii) consider whether such observations should be integrated into formal linguistic analysis and be modelled in the same way as categorical linguistic phenomena, and then (iii) see how this has been done by those who answer (ii) in the affirmative.

Variationist linguistics has shown that, alongside inter-speaker variation, intra-speaker variation is normal in individuals' linguistic performance. As one example among many, Sankoff (2004) shows that one thirty-five-year-old male speaker (originally from Yorkshire, but who had moved to the USA at age twenty-six) realised words from the STRUT lexical set, which might be described as the variable (ʊ), in the way shown in (18) in a single passage of speech. This speaker clearly realises one category variably.

(18) Intra-speaker variation of the vowel in the STRUT lexical set

ʊ	ɤ	ʌ	phonetic realisation
4	16	14	number of words

Quantification of the number of tokens of each variant like this is central to variationist sociolinguistics, and it can allow us to measure intra-speaker variation precisely. Such work usually involves more than one speaker, however,

producing results which amalgamate the scores of speakers. We can see this in an example of variation at the syntactic level, taken from work by Smith (2000). She discusses variation in the agreement of past tense-forms of *be* in Buckie, in north-eastern Scotland, the variable here being (*was*/*were*). Some of Smith's results, as presented in Hudson (2007), are shown in (19).

(19) Variation in agreement of past-tense forms of *be* in Buckie

	You (singular)		We	
Speaker age	You was	You were	We was	We were
Old	45	5	113	36
Middle	23	12	32	41
Young	43	33	101	45
All	111 (= 69%)	50	246 (= 67%)	122

This shows that, within one age group of speakers, both *you was* and *you were* are found, as are *we was* and *we were*. The numbers in (19) show the total number of times individual forms occurred in the corpus of data that Smith collected, grouped together in three age groups. In ascending age order, the groups contained nine, fourteen and sixteen speakers, with (nearly) equal numbers of males and females in each group. Quantitative investigations into variation typically present results for groups of speakers in this way, using categories such as age, gender and class (see Guy, this volume). Smith presents her data in a more fine-grained way than (19), but it has been discussed in terms of group scores in the theoretical and variationist literature on variation (for example, Adger 2006, 2007 and Hudson 2007).

What should linguistic theory make of such variation? Although groups of speakers are involved, the approach *need* not be polylectal. While some researchers in this field do aim to present a picture of the variation found in a community of individuals, others do not. As we saw above, some research explicitly counts the tokens of individuals as well as groups. Other work assumes explicitly that the group scores of quantitative sociolinguistics can be taken to represent the output of individual I-languages. Guy (2005) writes that 'extrapolating within communities from group data to individual grammars *is* justified. Granted, it's an approximation, but no more so than any research based on less than perfect sampling.' This type of approach aims to propose I-linguistic analyses of variable data, but it may veer into polylectalism as it does so. For the rest of this section, I assume that the theorists are aiming to model I-language when they grapple with variable data, and that they assume that individuals' I-languages *are* involved in the generation of variation.

One position in the theoretical literature which aims to deal with variation with little extra machinery is the claim that all cases of intra-speaker variation occur when a speaker has more than one grammar in their mind. This approach assumes that one grammar produces one variant, such as *you was*, or realising /t/ as [t], and this competes with one or more other grammars, which produces the other variants, such as *you were*, or realising /t/ as [θ̪]. Variation occurs on this model because a speaker sometimes uses one grammar and sometimes another. This approach is found in work such as Kroch (1989b), and is described by Lightfoot (1999: 94), discussing variation between OV and VO order in the history of English, which was subject to intra-speaker variation at some points in recorded Old English: '[w]here a language has such an alternation, we say that this manifests diglossia, and that speakers have access to two grammars. Certain speakers have access only to one grammar; others have access to the other grammar; and others have access to both grammars in an internalised diglossia.'

At its simplest, this approach means that intra-speaker variation need also not be relevant for the analysis of individual grammars – if it is always simply *grammar competition*, each individual grammar can be analysed as an invariable system, fitting in with the impetus within minimalism to banish grammatical optionality. This position means that intra-speaker variation is actually the same thing as inter-speaker variation on an idiolectal approach: the variation involved simply occurs as the result of different grammars (simply within one speaker, rather than within more than one speaker). For there to be a real conceptual difference between the subjects of this section and of section 7.3, we can now see that we should not really oppose (4) and (5), but, rather, (20) and (21), where (20) can occur either within or between speakers. Now (21) is the contentious issue: a model which assumes that intra-speaker variation is explained only by grammar competition, or a 'post-grammatical processor' will claim that (21) does not exist.

(20) variation between systems = inter-grammar variation
(21) variation within a system = intra-grammar variation

It is uncontroversial that a single speaker can possess more than one grammar – this is what we can assume occurs in cases of multilingualism. But is variation of the type shown in (18) and (19) really the same kind of thing as multilingualism? There may well be cases of intra-speaker variation that involve grammar competition, but this seems difficult to defend in many cases of variation. The patterns that are found in Liverpool lenition, for example (as mentioned above), seem to be clearly one system. The stops can be realised (i) as stops in all phonological environments, and (ii) as stops or affricates in a restricted set of environments, such as a post-nasal context, and (iii) as stops, affricates or fricatives in a further set of contexts, such as a postvocalic context (see Honeybone 2001; Watson 2007). This is a classic lenition trajectory (see Honeybone 2008), as shown for /t/ in (22), realising two 'stages' of one

lenition process, implying that occurrences of /t/ in the right environment (e.g. word finally in a word such as *dot*) can be realised as [t], [t͡θ] or [θ̲].

(22) t → t͡θ → θ̲

The realisations as affricates and fricatives seem so closely related, with similar but slightly different constraints on their realisation, that it is difficult to believe that [t͡θ] is derived by a completely different grammar to [θ̲]. The relevance of this approach is reinforced by cases where the same structural phenomenon can be seen to play both a categorical role and variable role in the same language. Guy and Boberg (1997) argue, for example, that the Obligatory Contour Principle (OCP, which forbids identical adjacent phonological elements, or forbids them from sharing a specific phonological property, such as a particular constellation of phonological features) plays a role in both categorical and variable phenomena in English phonology. Categorically, it forbids lexical geminates and enforces the avoidance of sequences of similar segments in past tense and plural suffixation (with *–ed* and *–es*). They argue that it also plays a role in determining when Coronal Stop Deletion (CSD) is likely to occur in a number of American varieties of English. CSD is one of the best-studied variable phonological processes in English. It involves the deletion of [t] and [d] when they occur in word-final consonant clusters, so that words like *rift*, *west*, *bold* and *find* can be pronounced with final [f] or [ft], [s] or [st], [l] or [ld], or [n] or [nd], respectively. Intensive study has shown that the likelihood of CSD applying varies according to a number of factors, including the nature of the previous segment (as shown in (23), for Philadelphia English, from Guy and Boberg, 1997) and the nature of the following phonological environment (as shown in the percentage figures in (24) for a range of varieties, from Coetzee and Pater (to appear), who also give references for the studies of each variety). Thus, for example, in terms of preceding segments, the stops are most likely to be deleted if they are preceded by a sibilant fricative or stop (which will be non-coronal, as in *act*), and are least likely to be deleted if they are preceded by a lateral (when they are still likely to be deleted, but could also perfectly well be pronounced).

(23) Probability of CSD in Philadelphia English

Preceding segment	Probability of deletion
Sibilant fricative	1.00
Stop	0.84
Nasal	0.78
Non-sibilant fricative	0.69
Lateral	0.66
N = 1,860	

(24) Percentage of CSD in different contexts in a range of varieties of English

	Pre-V	Pre-Pause	Pre-C
	west end	west	west side
AAVE (Washington, DC)	29	73	76
Chicano English	45	37	62
Jamaican English	63	71	85
New York City English	66	83	100
Tejano English	25	46	62
Trinidadian English	21	31	81
Philadelphia English	38	12	100

Guy and Boberg (1997) argue that the generalisations in (23) can be insightfully accounted for if we assume that a probabilistic OCP favours deletion of the following coronal stop if the preceding segment shares the features that characterise the stops (which they assume to be [−son, −cont, +cor]). They further argue that this effect is cumulative: the more features that are shared, the more likely it is that the stops delete. Their full analysis is intricate, but a simple form of it will show its basics: the segments most favouring deletion share two features ([−son, +cor] in the case of sibilants, [−son, −cont] in the case of stops), the nasals are mixed in terms of number of features shared, and the segments least favouring deletion share one feature ([−son] in non-sibilant fricatives and [+cor] in laterals). Guy and Boberg argue that the fact that structural linguistic constraints such as OCP can be seen to play a role in both categorical and variable phonology means that I-linguistic grammars should be expected to model categorical and variable phenomena in the same way, with the latter accounting for intra-grammar variation, which is seen as an authentic phenomenon.

A number of approaches have been developed to deal with intra-grammar variation within linguistic theory. The precise implementation of this naturally depends on the characteristics of the theoretical framework, and we cannot consider all linguistic frameworks here, for reasons of space. I narrow the discussion to the most popular models: rule-based phonology, optimality theoretical phonology and syntax, and principles and parameters syntax (although it should be noted that other models have also been proposed as appropriate to model variability, such as word grammar, see Hudson (1997) and HPSG, see Bender (2007)).

The simplest way to model variation is to assume that a phenomenon, such as the realisation of /t/ as [ʔ] or [θ̪], or the deletion of /t/ and /d/ in CSD is

controlled by a linguistic rule which is marked as optional in the grammar. The phonological use of optional rules stretches back to Chomsky and Halle (1968), and has recently been defended by Vaux (2008). However, most of the work in rule-based phonology which takes variation seriously has replaced optional rules with variable rules, which aim to write factors into the formulation of a rule which governs how often particular variants of a variable occur, so that 'the predicted relative frequency of a rule's operation is, in effect, an integral part of its structural description' (Cedergren and Sankoff 1974). This is in part based on the claim that variation is not simply a matter of optionality, but is subject to orderly heterogeneity.

Labov (1972b) discusses CSD, presenting a number of ever-refined variable rules to account for it. One of these is given in (25), accounting for the observation shown in (24), that the stop is less often deleted if it precedes a vowel than if it does not.

(25) [−cont] → <Ø> / [+cons] __ # # <−syl>

The angled brackets in (25) indicate aspects of the rule that place variable constraints on it, so (25) does not claim that the whole process is optional – rather, it indicates that the process is more likely to apply if a vowel does not follow the word-final stop, producing a probabilistic grammar. The variable rule approach has been taken up in variationist linguistics, but has not been all that widely used, probably because most variationist work focuses on quantifying the variation found in particular linguistic phenomena, rather than modelling it grammatically, and most phonological work does not focus on variation, either dealing only with categorical processes, or ignoring the fact that the processes considered are, in fact, variable. Most of the work on phonological theory which *does* engage with variation is now conducted in OT. Some work on variation and syntactic theory also uses OT, but the principles and parameters framework is also well represented in theoretical discussions of syntactic variation.

Optimality theory replaces linguistic rules with violable constraints on linguistic forms which are ranked in terms of their relative order of importance in determining the surface form (called the output). The earliest variable OT work was similar to the optional rule format in simply allowing more than one output without giving any indication of the pattern in the variation, thus an analysis could allow two candidates to tie in terms of constraint violations, either because neither violates a set of fully ranked constraints, or because particular constraints are tied in the ranking. This is illustrated in (26), from Pesetsky (1997), which accounts for variability in the realisation of the declarative complementiser *that*. Both forms of the sentence in (26) are possible, with the complementiser pronounced (26a) or not (26b), and Pesetsky accounts for this

by giving them the same number of constraint violations, of equally ranked constraints (LE(CP) and TEL are 'tied constraints'). The constraints are:

- RECOVERABILITY: a syntactic unit with semantic content must be pronounced unless it has a sufficiently local antecedent
- LEFTEDGE(CP): the first pronounced word in CP is the complementiser that heads it
- TELEGRAPH: do not pronounce function words

(26)

Candidates	REC	LE(CP)	TEL
a. I believe [$_{CP}$ that Peter is hungry].			*
☞ [on ranking: LE(CP)»TEL]			
b. I believe [$_{CP}$ ~~that~~ Peter is hungry].		*	
☞[on ranking: TEL > > LE(CP)]			

This approach has not found widespread favour within OT, however, because there will likely always be some constraint, even if it is very low ranked, which would be able to decide between the candidates which seem to be tied in the necessarily abbreviated constraint rankings that are considered in analyses such as (26). A more finessed approach to variation in OT assumes that constraints can be variably ranked. Such approaches always deliver a winning candidate, avoiding the problem of tied constraint violations. They assume that, while most constraints are ranked normally, the constraints which account for variation are only partially ordered – they 'float' in a block in the hierarchy and are not ranked with reference to each other. Each time that the grammar is used to evaluate a candidate set, the partially ordered constraints are fully ranked in a way which is consistent with the partial ordering – the ranking is randomly chosen, by the grammar, delivering surface variation.

Unlike the tied-constraint-violation model, this approach aims to say something about the probability with which particular variants occur (thus improving on the 'tied-violation' model, as variable rules aimed to improve on optional rules). It aims to model the type of quantification of variation shown in the numbers given in (18), (19), (23) and (24).

The number of grammars which can be created through the random ranking of a set of partially ordered constraints is limited, and an analyst can work out how many of each of the possible grammars allows for each possible variant

output. As the ranking is entirely random, the prediction is that the number of possible grammars which gives a variant output should align with the proportion of variants actually produced by speakers. Coetzee and Pater (to appear) describe Coetzee's (2004) model (following work by Kiparsky and Anttila) of the aspects of CSD shown in (24). The constraints used are:

- *CT = consonant clusters may not end in a coronal stop
- MAX = input consonants must be present in the output
- MAX-PRE-V = input consonants in pre-vocalic position must be present in the output
- MAX-FINAL = input consonants in phrase final position must be present in the output

If these constraints are variably ranked, they can give rise to twenty-four different rankings, which fall into five sets in terms of the outputs they can produce, shown in (27). If MAX is ranked highest, as in (27a), there can be no deletion, no matter how the other constraints are ranked, and if *CT is ranked highest, (27e), the stops will delete no matter whether they are pre-vocalic, pre-consonantal or pre-pausal (phrase final). Other rankings show different patterns of deletability.

(27)

			Deletion produced?		
	Crucial rankings	Total #rankings	Pre-V	Phrase-final	Pre-C
a.	MAX »*CT	12	No	No	No
b	MAX-PRE-V » *CT » {MAX, MAX-FINAL}	2	No	Yes	Yes
c.	MAX-FINAL » *CT » {MAX, MAX-PRE-V}	2	Yes	No	Yes
d.	{MAX-PRE-V, MAX-FINAL} » *CT » MAX	2	No	No	Yes
e.	*CT » {MAX, MAX-PRE-V, MAX-FINAL}	6	Yes	Yes	Yes

The prediction here (shown by comparing the second two columns) is that pre-consonantal deletion is much more likely than deletion in a pre-vocalic or phrase-final environment because many more grammars allow it, and this matches the observed results in (24) for all dialects. However, the precise percentages of the observed variation are not modelled, nor could they be using the same set of constraints for all dialects, as the proportions are different. (27) produces the same number of rankings to allow deletion in pre-vocalic and

phrase-final environment. (24) shows that dialects vary as to which of these two environments is most likely to allow deletion, so this may be the right result, but it means that it is not clear how far the model can or should go in modelling the probability of particular variants occurring. It would be too optimistic to expect that such modelling will always account for the number of variants of each variable, and the variation that is captured here cannot model differences between individuals in a speech community – it only models broad relative probabilities in the overall surface variation, relying on fixed numbers of specific constraints to derive the numerical effects. Problems with this approach are evident: if there were also a MAX-PRE-C, the numbers would not fit so well. Also, some work in this approach goes further, attempting to model precise percentages of variation in terms of the likelihood of particular rankings, relying on the existence of precise numbers of constraints for each case of variation.

The other main OT model of variation retains the idea that there is always one full constraint ranking for each 'use' of the grammar but models variation more simply, by essentially weighting constraints with numbers which determine how likely a particular constraint ranking is. Constraints can be allowed to overlap in their ranking (to different, numerically specified degrees), as in Stochastic OT (e.g. Boersma and Hayes 2001), such that one constraint can overpower another most of the time, but not all of the time, in line with their ranking value and some random numerical 'noise', which is generated each time a linguistic form is processed, ready for utterance. This approach is quite radical, as it incorporates numbers directly into the grammar, meaning that language acquirers must abstract the ranking value for constraints from the speech that they are confronted with, and attach it to a particular constraint, and, perhaps for this reason, has met with some considerable objections.

Stochastic OT was developed in phonology, but has also been applied to syntax, as in Bresnan, Deo and Sharma (2007). Most work in theoretical syntax works with non-violable principles which apply in all languages in the same way, however, rather than ranked constraints. Cross-linguistic variation is derived largely through the use of parameters, which allow a linguistic system to choose from a restricted range of structural possibilities. When syntax was modelled using construction-specific rules, it is easy to imagine how optional or variable rules could be used in the same way as they were in phonology, as discussed above in (25), and, although little variationist work on syntax was being carried out at that point, obvious cases of variation, such as in verb-particle constructions, like the difference between (1) and (2), could simply be modelled by an optional rule (as could passivisation, for example).

In frameworks which work with principles and parameters, intra-speaker variation has often been modelled as grammar competition involving multiple

grammars within one speaker, as in work by Kroch and Lightfoot, discussed above. Not all work has assumed this position, however. Henry (1995), and considerable subsequent work, has argued that the grammar competition model seems implausible when confronted with multiple cases of variation. Thus, Henry (2002) writes:

> [W]ithin Belfast English, there is variability in relation to agreement patterns, word order in imperatives, inversion in embedded questions, and the use of the relative pronoun in subject contact relative clauses. By no means every speaker has variation in all of these, so that there is a range of possible grammars with and without variation for a range of structures; if there is grammar competition, then it is between a wide range of grammars, not just two, and a better characterization seems to be that individual structures/parameter settings are variable, rather than that there are actually separate grammars.

Henry (1995) argues that the approach later adopted by minimalism – that intra-grammar variation is impossible – is problematic. She argues explicitly that there can be optionality in syntactic movement. For example, certain dialects of Belfast English allow overt subjects in imperatives (a clear inter-speaker difference to Standard English), and they furthermore allow weak object pronouns to precede or follow the subject, thus both (28a) and (28b) are possible as imperatives.

(28)
(a) Give it you to the teacher.
(b) Give you it to the teacher.

Henry accounts for this difference as optional subject raising, from SPEC/VP to SPEC/AGR$_S$P. She argues that similar cases of syntactic optionality should be modelled in similar ways, typically assuming that differences in parameter settings drive the variation. This leaves the likelihood of either form occurring fully outside of the grammar, drawing back to a position where grammar should only account for what is possible, leaving sociolinguistic factors (a 'post-grammatical processor'?) to account for which form is used when.

Other syntactic work has tried to retain the notion that syntactic structure does not allow for optionality by exploiting the reinterpretation of parameters that has occurred in principles and parameters syntax, so that they are no longer seen as 'switches' but as lexical items. As Adger and Trousdale (2007) write, this 'doesn't seem like a major change; however, while it is impossible to have a switch being in both positions at once, it is certainly possible to have two lexical items in a language with contradictory specification'. This means that it is possible to model variation such as the variation in agreement described in (19), as in Adger and Smith (2005) and Adger (2006), by allowing a language to contain both agreement options as functional elements, which a speaker

tacitly chooses from when a linguistic form is processed. Adger (2006) also aims to model aspects of the probability of occurrence of particular forms.

This section has shown that linguistic theorists have developed a range of ways of dealing with the problem of intra-speaker variation, even if it is seen as true intra-grammar variation. The question as to whether theoretical linguistics *should* deal with this type of variation remains unanswered in many phonologists' and syntacticians' minds, however. Linguistic theory has had a clear degree of success in modelling the patterns of variation that can occur, and that have been observed in variationist linguistics, as we have seen. Many would still argue, however, that, while linguistic theory might allow for and model variation, it should not predict how frequently each variant of a variable will occur, because that is seen as purely the job of sociolinguistics. The status of variation in syntax has also led to some concern – it is less easy to see two different surface forms as being derived from the same underlying form in syntactic theory than it is in phonology. It is straightforward to see a statement that 'the variable (t) has the variants [t] and [ʔ]' is equivalent in status to a rule such as /t/ → [ʔ], but the equivalent does not sit so easily with modern models of syntax. We have assumed in this chapter that essentially the same issues arise when theoretical phonology and syntax are confronted with variation, but some syntacticians would object that this is not so.

Intra-speaker variation is more widely seen as something which linguistic theoretical linguistics should deal with than is inter-speaker variation. The issues that the two raise are rather different. As we have seen, intra-speaker variation only really forces linguists to develop new theoretical machinery if it is accepted that it involves intra-grammar variation. If it is, the models discussed in this section become important extensions of linguistic theory.

7.5 Conclusion

This chapter has considered quite a wide range of issues which arise when linguistic theory is confronted with the observations that variationist linguistics and other studies of linguistic variation, such as dialectology, have established. Both inter-speaker and intra-speaker variation can be seen to pose problems for linguistic theory, and, as we have seen, some theorists consciously decide to ignore these problems, believing that theoretical linguistics should deal only with categorical phenomena, or at most should describe what is possible, not what is likely or unlikely (for example, Newmeyer 2003, 2005 summarises a range of arguments as to why theoretical linguistics should ignore variation). We have also seen that some considerable thought has been devoted to tackling these problems head-on, however. While linguistic theory wrestles with these issues, and there is substantial disagreement in the field about how linguists of a theoretical bent should deal with them, I hope to have shown that variation is

not really a *problem*, but rather an opportunity for innovative thinking, to test the boundaries of linguistic theory and for far-reaching argument concerning the nature of language.

7.6 Where next?

For empirical details of phonological variation between speakers (and varieties) in English, Wells (1982) is still the place to start, but this should be read in conjunction with more detailed work, such as the chapters in Foulkes and Docherty (1999) for UK varieties and Labov, Ash and Boberg (2006) for US varieties. Variation in both phonology and syntax are covered in the chapters in the two-volume Kortmann *et al*. (2004). There is a vast amount of variationist quantitative work on English, which is relevant to understanding variation within as well as between speakers, starting with that by Labov (Labov 1972b, for example, is still well worth reading). Any textbook on sociolinguistics (such as Chambers 2003) will likely also discuss the basics. Most relevant is work applying linguistic theory to data which involves variation in some way. For syntax, the articles in Trousdale and Adger (2007) and Cornips and Corrigan (2005) represent a good cross-section of work, and the introductions to the volumes discuss fundamental issues and details, as does the overview in Henry (2002). For phonology, the chapters in Hinskens, van Hout and Wetzels (1997) are similarly worth consulting, as are overviews like Coetzee and Pater (to appear) and Anttila (2002b). There is a considerable strand of work attacking the polylectal approach to inter-speaker variation: parts of Harris (1984b, 1985), McMahon (1992, 2000) and Lodge (2009) all consider some aspects of the issues. Little work explicitly advocates polylectalism (the work of Bailey, such as Bailey 1996, is one often-cited exception, but is unfortunately difficult to read and contains few analyses). However, as mentioned in the chapter, traces of this approach can be discerned in quite a range of work.

8 Variation and change

Gregory R. Guy

8.1 Introduction

Like most human activity, language does not fit neatly into the analytic boxes that observers often use to segment, categorise, and theorise about the subject. Whether those boxes are called features, phonemes, or syntactic structures, or rules, constraints, or principles, the facts of language always slop over the edges or ooze from one into another. The customary approach in linguistics is to treat this mismatch between categories and facts as 'linguistic variation' – but we should be clear that doing so effectively privileges the analytical categories over the empirical substance. Variation, as traditionally understood, involves single categories being mapped onto variable realisations, as if the categories were primary and given – platonic ideals existing on a higher, purer, plane, that are only imperfectly reflected in the muddy reality of speech. An alternative view, in which natural language in all its richly variegated glory is primary, and the analytical categories are as yet imperfect theoretical constructs that provide only a crude model of reality, is rarely considered. As a healthy terminological corrective, perhaps linguists should consider thinking about variation as highlighting the problem of 'theoretical inadequacy'.

Nowhere is this lousy fit between theoretical models and variable facts more evident than in the treatment of language change. Since Saussure, linguistic theory has for the most part assumed the irrelevance of diachrony in the construction of formal theory, producing as a consequence static models that not only fail to accommodate change, but actually appear to exclude it as a logical possibility. Theoretical models are designed to be self-contained systems, supported by their internal structure and logic, covering a strictly defined terrain (from which diversity is excluded). Such theories are like buildings and, as we know from experience, buildings do not evolve organically; rather, they change by getting completely or partially demolished and replaced. Consequently, such theories make change seem anomalous, or impossible, and in any case, located outside of theory. And yet, linguistic reality obstinately refuses to accommodate to these models, and all languages go on changing continuously all the time. What's a linguist to do?

The resolution of these contradictions lies in abandoning the theoretical assumptions that inhibit a proper treatment of linguistic variation and change. Since the Neogrammarians, the main stream of theoretical development in linguistics has been enchanted with the idea that 'exceptionlessness' (Neogrammarian *Ausnahmlosigheit*) is an essential trait of valid linguistic generalisations; since Saussure, variability has been defined as lying outside the linguistic system, external to *langue*, competence, and grammar. But an alternative model exists which avoids these anti-empirical assumptions, in which valid generalisations may be non-categorical, and variation may be seen as systematic and internal to grammar. A path-breaking formulation of this position is found in Weinreich, Labov and Herzog 1968. These scholars enunciate two principles that are foundational to this alternative: orderly heterogeneity is the principle that variation is not equal to chaos, but may still contain system and order, and inherent variability is the principle that variability is intrinsic to language, effectively perceived, processed, and produced by all speakers, and therefore lies within competence, and hence within grammar.

With this change of assumptions, it becomes clear that variation and change are essentially one and the same phenomenon. The speech community, and the mental grammars of speakers, encompass and manipulate linguistic differences at all times. No two speakers have identical grammars and linguistic repertoires, and no single speaker has a completely homogeneous and invariant grammar. Therefore, to say or understand anything at all, a language user must be able to deal with difference, with 'variation'. Unsurprisingly, the particular patterns of difference fluctuate across time, just as they fluctuate across speakers and social situations. Therefore, variation is the synchronic face of change, and change is nothing more than diachronic variation. Indeed, the historical record, along with studies of change in progress, make it clear that there is no such thing as change without variation: all changes pass through periods of time during which outgoing and incoming forms coexist in variation in the speech community. However, the evidence also suggests that change is not an inevitable outcome of variation; certain sociolinguistic variables, such as the *–in/–ing* alternation in English (cf. alternations like *running ~ runnin'*), appear to have existed for many centuries without one form completely supplanting the other. But this asymmetry between the two is not unexpected in an adequate dynamic model of language. In expanding our view of grammar to incorporate variability, we do not preclude stability; synchronically some features of language do not vary, and diachronically, some features of language do not change, at least within certain time horizons. Hence diachronically stable variation is a possible characteristic of an adequate model of language.

What has conventionally been treated as two topics – 'linguistic variation' and 'language change' – is thus really one topic differentiated only by time scale: change is long-term variation. Consequently, each of these topics

illuminates the other. Studies of variation in a short time frame (i.e. 'synchronically') implicitly contain information about what, from the perspective of a longer time frame (i.e. 'diachronically'), may be seen as change. This prospect has inspired a great deal of work within sociolinguistics and variation studies, addressing issues such as:

- what does linguistic variation today tell us about recent and future change? (e.g. how can change be read off from the synchronic record of diversity?)
- what does the study of language change tell us about variation today? (e.g. how does knowledge about change influence the interpretation of synchronic variation?)
- how does change proceed and progress?
- how can variation and change studies address traditional questions of historical linguistics?
- how does the social embedding of variation play out in diachrony?

This chapter presents a survey of contemporary issues in the study of variation and change, along with a reflection on the relationship of this work to the traditional approaches to language change embodied in the field of historical linguistics. We conclude with a consideration of the implications of this work for linguistic theory.

8.2 The study of change in progress

The earliest work on change in progress, by Labov (1963, 1966), made a basic distinction between two types of change that differed according to their social and psychological properties, what Labov called change from above and change from below. In this model, changes from above are effected consciously (hence Labov's elaboration as 'above the level of conscious awareness') and involve imitations of external models, while 'changes from below' are 'below the level of conscious awareness', and involve spontaneous innovations that are not based on an external model. Subsequent work confirms the need to recognise distinct types of change, based on different social mechanisms, but the specific criterion of consciousness is of doubtful utility in making the distinction, since there are changes involving spontaneous innovations of which there is considerable conscious awareness (e.g. the spread of high rising terminal intonations in declaratives in Australian English, Guy *et al.* 1986), and changes involving accommodations to external models that speakers show little awareness of. Instead, the literature suggests a convergence on a three-way distinction between **spontaneous** innovations, arising from within the speech community (subsuming 'change from below'), **borrowings** involving language or dialect contact but conducted by native speakers of the variety undergoing change (including Labov's 'change from above'), and **impositions**, arising in contact

situations but conducted by speakers involved in language shift (cf. Thomason and Kaufman 1988; Van Coetsem 1988; Guy 1990). This third type has no equivalent in Labov's dichotomy; it includes the transferences from L1 to L2 that underlie 'foreign accents' and substratum effects.

Since each of these types involves a distinct social and psychological mechanism, it is expected that they display different social and linguistic distributions. Impositions should, at least initially, reflect systematic features of the L1, including patterns of variation; to the extent that progress in acquisition of the target language tends to suppress transference, one might expect gradual convergence on the variable patterns of the L2, led by speakers who have greater access to the target. Changes of the borrowing type most typically involve borrowing of prestige norms from a source external to the speech community, meaning that they are led by higher status speakers, and speakers in the age and class groups that have greater mobility, investment in, and/or access to the external prestige norm. Thus Labov (1966) finds that the reintroduction of coda /r/ in New York City is led by the upper middle class, and by young adults rather than adolescents. But note that other motivations for borrowing exist, and imply other social distributions; thus Cutler (2002) describes the adoption of features from African American English by white youth affiliated with hip-hop music and culture, and Stuart-Smith *et al*. (2007) describe the spread of historically non-local features like TH-fronting (i.e. replacement by /f/) among working-class Glasgow speakers, apparently motivated by the construction of a distinctive identity differentiating them from 'posh' (i.e. middle-class) speakers.

Spontaneous innovations do not involve contact with external sources; hence their social distribution reflects internal social dynamics of the speech community. Since they diverge from existing usage, rather than converging on a target, they reflect social processes of differentiation – contrastive processes of identity formation in which groups or individuals, by advancing the change, distinguish themselves linguistically from some reference point, rather than accommodating to it.

8.2.1 The social distribution of change in progress

Much of the research on the social distribution of change in progress has focused on spontaneous innovation ('change from below'). More than forty years of research has revealed a set of social tendencies that are well validated, at least for the types of societies in which these studies have been done – which, admittedly, are predominantly advanced industrial societies in the western world, although there are studies from Latin America (e.g. Cedergren 1973 on Panama, and numerous studies of Brazilian Portuguese), from Japan (e.g. Hibiya 1996), Egypt (Haeri 1996a), Iran (Modaressi 1978), and elsewhere.

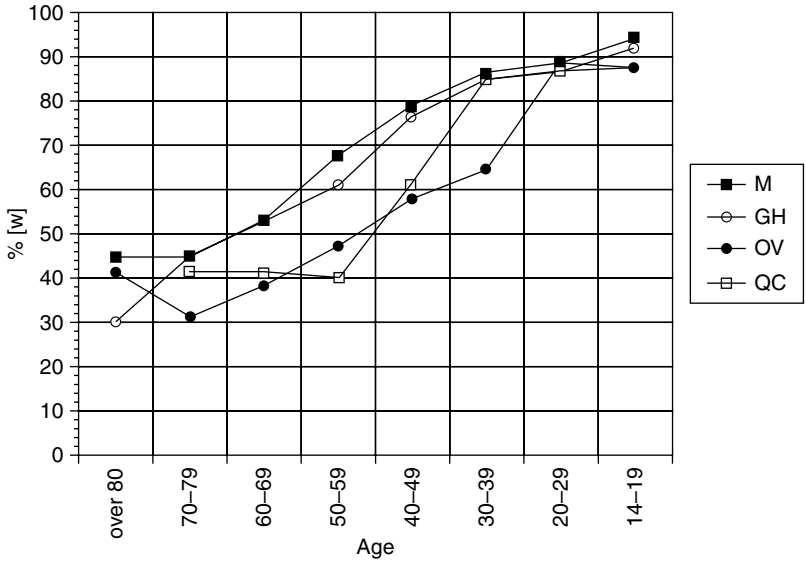


Figure 8.1 Percentage of speakers with [w] not [hw] in words like *which* and *whine* in four Canadian regions: Montreal (M), Golden Horseshoe (GH), Ottawa Valley (OV), Quebec City (QC) (from Chambers 2002: 63)

This work identifies age, class, and gender as the principal social dimensions reflecting ongoing spontaneous linguistic change in a community. The most commonly observed empirical patterns are as follows.

Age: The synchronic age distribution of a variable is considered the most crucial evidence for spontaneous change in progress. Older speakers are conservative, while younger speakers lead in the use of an innovation. The age distribution typically follows an S-shaped curve, as seen in Figure 8.1, from Chambers' research on the loss of /h/ in /hw/ clusters in Canadian English (Chambers 1998, 2002).

Such a distribution, recurring in many studies, is the typical synchronic face of ongoing change. It should be noted that where information is available on younger age-groups (children and younger adolescents), the data almost invariably show that the highest rate of innovation is not found among the youngest speakers, but rather among older adolescents and young adults, that is, somewhere in the age range 15–24. This no doubt reflects the development of social autonomy and the formation of a distinct social identity; the youngest speakers live in their parents' homes, and lead lives strongly governed by adults who, according to the age distribution seen in Figure 8.1, are relatively conservative. It is only in late adolescence or young adulthood that speakers construct an independent social and linguistic identity, achieve social autonomy, and

minimise parental linguistic influence. This, evidently, is the point in the lifespan when speakers advance the use of linguistic innovations, going beyond what the next older age cohort has done to a still higher level of usage.

Social class: The social-class distribution of a spontaneous innovation has been argued by Labov and others also to display a distinctive pattern which is absent in cases of social stratification without ongoing change. This position is not entirely uncontroversial, but there are numerous studies reflecting the distribution that Labov considers decisive: the so-called 'curvilinear pattern', in which the peak use of an innovation is found towards the middle, or lower middle, of a social spectrum, while both the lowest and highest status groups lag in adopting the innovative form. Some classic examples are found in Figure 8.2, showing the distribution of vocalic changes in Philadelphia English. In both figures, the most advanced forms (in these cases, those raised the farthest along the front vowel diagonal) are found in the upper working class, as defined by a composite scale of socioeconomic status based on measures of occupation, education, and income.

The social motivation for the curvilinear pattern has been much debated. Labov's (2001) explanation relates the phenomenon to the differential importance of 'local identity' (i.e. solidarity with one's friends, family, neighbours, and community) across social classes. This is low both in highest-status groups (cf. concepts like the 'jet-set', people who are not strongly tied to one place, but derive their social position from supralocal affluence and influence), and lowest-status groups (cf. groups like the homeless, who also lack strong ties to a specific neighbourhood). For Labov, this aspect of social and psychological identity peaks in the upper working and lower middle classes, who have strong community ties and relatively low mobility. Hence these are the people with the greatest motivation to adopt and extend the distinctive characteristics of the communities that they belong to, and to demonstrate community membership contrastively by differentiating themselves from other individuals who do not belong. This view is reinforced by Milroy's work, showing strongest use of local forms by speakers with the strongest local community ties (1987).

Gender: Studies of the gender distribution of spontaneous innovations are distinctly skewed: substantially more of them show female speakers in the lead. This topic has attracted a great deal of interest in the field. But there are some studies showing males in the lead (e.g. the centralisation of the nucleus of /ay/ (in PRICE words) before voiceless segments in Philadelphia), and studies with no significant gender differentiation. The empirical findings are thus more mixed than those for age and class, and the explanations that have been proposed are more diverse.

Some of the major lines of explanation that have been advanced for gender differences in change are as follows. One approach refers to networking and socialisation patterns. Labov (2001) finds that the leaders of change are people,

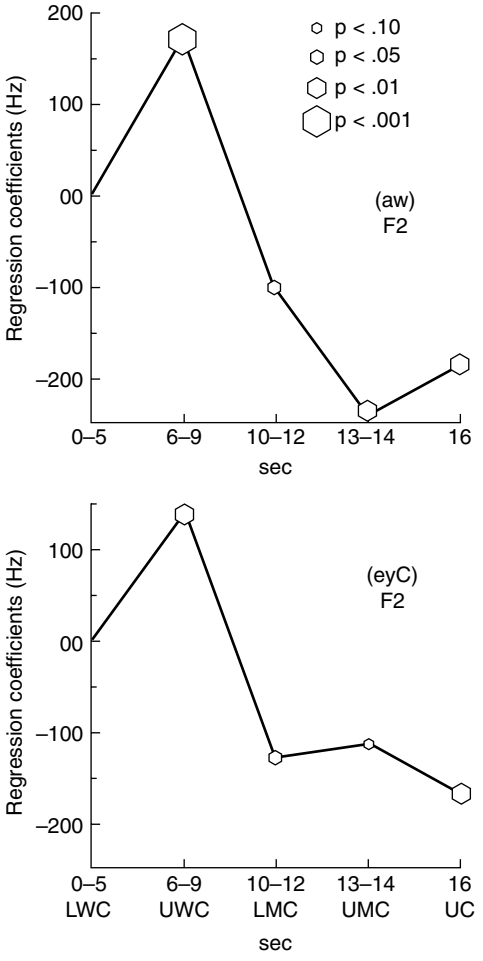


Figure 8.2 Curvilinear socioeconomic class distribution of vowel changes in Philadelphia English (from Labov 1980)

often women, who have both strong local ties and broader networks. In communities where women are more socially connected with a range of interlocutors, they would be more likely to have access to innovations as they develop and advance, and to participate in the construction of a local community identity via language. Another explanation appeals to gender differentiation in contact with younger children: if most of the adult caregivers for young children in a community are female (mothers, childcare workers, primary school teachers, etc.), then gender differentiated innovations favoured by females are more

likely to be transmitted to the next generation of language acquirers (cf. Labov 2001). Male-led changes face a transmission problem if men don't talk much to the children.

An interesting account of sound changes in terms of acoustic differences between men and women is found in Haeri (1996b). Haeri suggests that acoustic iconicity is involved arising from the size differences between men and women. Vowel-like sounds are acoustically defined by their formants, which are resonances of the vocal tract. Like all resonances, formants vary with the size of the resonating space (compare the high notes of a little piccolo with the deep bass notes of a big tuba). Hence larger speakers (like adult males) have lower formant frequencies, and smaller speakers (like adult women, and more extremely, children) have higher formants. In the front–back dimension of the vowel space, which is acoustically signalled by the second formant (F2), a higher formant value means that a female speaker's vowels sound relatively more fronted. In normal speech perception, hearers compensate for this difference by 'normalising': interpreting the formant values with reference to the apparent size of the speaker's vocal tract. But hearers presumably retain access to the raw acoustic difference at some level. In changes on the front–back dimension, hearers might then systematically interpret female productions as being marginally fronter. Surveying nineteen different changes involving this dimension, Haeri notes that twelve out of thirteen fronting changes reveal a female lead, while five of six changes involving backing show males in the lead.

8.2.2 The linguistic distribution of innovations

Another productive area of research considers the question of directionality of linguistic change: do certain changes proceed only in one direction, and never reverse? If so, they permit the direction of change to be read off from the mere fact of variation. For example, grammaticalisations are changes that involve content words evolving into function words; the English indefinite articles *a/an* developed from the word *one*, for example. In Brazilian Portuguese *a gente*, historically a noun phrase meaning 'the people', is currently becoming a pronoun meaning 'we' (Zilles 2005). The reverse direction of change, of function word into content word, is virtually unknown. Hence, in the Brazilian case, the fact that content-word and function-word usages of *a gente* were in variation until recently (and may still be varying for some speakers) immediately implies that the content form is prior, and the pronominal usage is the innovation. In this particular case, a written historical record exists in which we can trace this development, but if grammaticalisation is unidirectional, the conclusion would still be valid even without any data from earlier times.

There are a number of claims in the diachronic linguistic literature that particular changes are unidirectional. As we have noted, grammaticalisations

appear to be irreversible, perhaps because they typically involve a cluster of related changes in morphosyntax, phonology, and meaning leading to the evolution of function words and affixes, which would be impossible to unpick. In phonology, cases of deletion are perhaps the most obvious candidate for unidirectionality; hence it is extremely likely that any cases of variation between the presence and absence of some element derive historically from the full form, via deletion, rather than the zero form, via insertion (barring cases of excrescent insertion such as Spanish *homre* > *hombre*, English *thunre* > *thundre* > *thunder*). Hence when we look at English final coronal stop deletion, encountering variable realisations of words like *just~jus'*, *old~ol'*, etc., we can be confident that *just*, *old*, etc. are closer to the historical sources. Other phonological processes that are much more common in one direction than the reverse include lenition, assimilation, and merger. Fortition (the reverse of lenition) is typically limited to specific prosodic conditions, when it occurs at all, and dissimilations are rare. Complete mergers are essentially irreversible, which is why they tend to spread across the dialectological and sociolinguistic landscape; however, there are attested cases of near-merger in which speakers retain some capacity to distinguish the merged phonemes, which occasionally leads to subsequent re-differentiation (see Thomas, this volume, also Labov 1994).

One well-known case where unidirectionality has been claimed is the theory of vocalic chain-shifts advanced by Labov, Yaeger and Steiner (1972) (LYS; also Labov 1994; for further discussion of vowel shifts, see Thomas, this volume). Based on extensive empirical studies, these scholars propose three principles governing chain shifts in vowel systems (such as the English Great Vowel Shift):

1. Tense vowels raise.
2. Lax vowels fall.
3. Back vowels move to the front.

Tense vowels, for these scholars, are those that are relatively peripheral, articulated near the perimeter of the vowel space. An example of the first two of these principles is the English Great Vowel Shift. The non-high long vowels of Middle English (tense and peripheral) were all raised – for example, ME [eː] and [oː] raised to Modern English [iː] and [uː] respectively – while the high long vowels of ME first diphthongised and acquired centralised (i.e. non-peripheral and therefore lax) nuclei, which then fell down the central vowel space, so that ME [iː, uː] yield ModE [ay, aw].

These unidirectional principles are illuminating when we examine cases involving vocalic variability. Canadian English, for example, has systematic variation in realisation of the lax front vowels [ɪ, ɛ, æ], each of which varies along a range from higher and fronter to lower and backer (Clarke *et al*. 1995; De Decker 2002). LYS's principle II predicts that the direction of the change

is towards the lower realisations, and indeed, other evidence, such as the age distribution of the variants in the population, confirms this prediction. This Canadian Shift involves lax vowels lowering in a chain shift, possibly triggered by the prior merger of /a/ and /oh/ (i.e. the vowel classes of COT and CAUGHT), which created room in the low central region of the vowel space for /æ/ to lower and back, generating a pull-chain shift.

In striking contrast to the Canadian Shift, the front lax vowels of English in Australia and New Zealand are raising, which appears at first blush to contradict LYS principle II. However, upon closer inspection, it turns out that these vowels are the most peripheral front vowels in these dialects. The front 'long' vowels (/ey / and /i/ as in FACE and FLEECE), have acquired centralised, non-peripheral nuclei in antipodean English, thereby abandoning the periphery to /ɪ, ɛ/, which are raising following LYS Principle I.

8.3 Real and apparent time

The underlying unity of linguistic variation and change is perhaps clearest in the analysis of the temporal extension of linguistic variables. There are two traditional perspectives on this question. One viewpoint has been described in the previous section: we can examine the age distribution of a variable at one point in time. This is customarily referred to as **apparent time** evidence, and it typically shows that the innovation is used more by younger speakers (Guy *et al*. 1986; Bailey *et al*. 1991). The alternative is akin to the perspective of traditional historical linguistics: examining data from different points in time, to see how the usage of variants has shifted during the interval between the samples. Such an approach looks at evidence from **real time**; for all innovations that are continuing to advance, real-time evidence will show an increase in the occurrence of the newer forms across time. (In language, as in genetics, not all changes are successful; some innovations appear, advance, and then recede. Thus Blake and Josey (2003) found that the vocalic innovations – centralisation of the nuclei of /ay, aw/ (in PRICE and MOUTH words) – that Labov had described in Martha's Vineyard in 1963 were disappearing forty years later, as the island economy was re-oriented away from fishing toward more integration with the mainland.)

The relationship between these two kinds of evidence, real and apparent time, has received much attention. The two basic findings – spontaneous innovations show greater use by younger speakers in apparent time, and greater occurrence at later points in real time, suggest an obvious social mechanism for the spread of linguistic change. The community is not changing as a whole – with every speaker moving in the same direction, but rather, the membership of the community is changing, as new generations arrive and older ones depart, and different generations speak differently. The time course of a change

spreading through a community thus involves two separate principles. First is incrementation: in a linguistic change that advances continually to completion, each successive age cohort uses on average a higher frequency of the new form than the cohorts that preceded it (their older siblings, in effect). Second is individual constancy: each cohort, and for the most part, each individual, remains mostly stable in their usage once they reach some point in late adolescence or young adulthood.

Real-time evidence supplements apparent-time evidence in an important way: it rules out an alternative hypothesis that might explain apparent-time data by a different mechanism, namely, age-grading – a situation in which individuals regularly alter their behaviour as they get older, but the community is not changing. We have interpreted Figure 8.1 as showing that [hw] clusters in *which, whine*, etc. are being lost in Canadian English. But why do we not believe that every Canadian speaker starts out using primarily [w] in these words when they are young, and proceeds gradually across their life-span to prefer more [hw] usage, peaking in old age? This is a logical alternative, which would not imply any change in the community as a whole; rather, it would imply that if we repeated Chambers' study at multi-year intervals, we would get essentially the same graph for age distribution, but each individual, were we able to track them, would have increased their [hw] usage as they aged. There are various reasons to be dubious of such an interpretation, including our understanding of normal language acquisition, but there is much real-time evidence in the literature that convincingly refutes such explanations. An example appears in Hibiya's (1996) study of the change of the velar nasal to [g] in Tokyo Japanese, shown in Figure 8.3.

This graph combines real-time and apparent-time data. The individuals to the left of the vertical line in Figure 8.3 are speakers interviewed by Hibiya in 1986. They follow a standard S-curve with younger speakers using more [g]. The speakers to the right of the line are people recorded by Japanese national radio in the 1940s who were born in the late nineteenth century. They are plotted according to the age they would have been in 1986 when Hibiya interviewed the other speakers, and they show the extension of the lower end of the curve into the past. But when they were interviewed, their actual ages were sixty–seventy-five years; a comparison of these subjects with Hibiya's subjects of comparable ages recorded forty years later rules out the possibility that all speakers start with high [g] use and decline as they get older. Rather, it is birth year and generational cohort that is associated with rate of [g] use, not age at any given point in time.

There are two approaches to collecting real-time evidence that permit the most detailed picture of a change in progress, and maximise the comparability of data across time. These are **panel studies** and **trend studies**. A panel study follows a specific group of individuals and resamples them at various points in

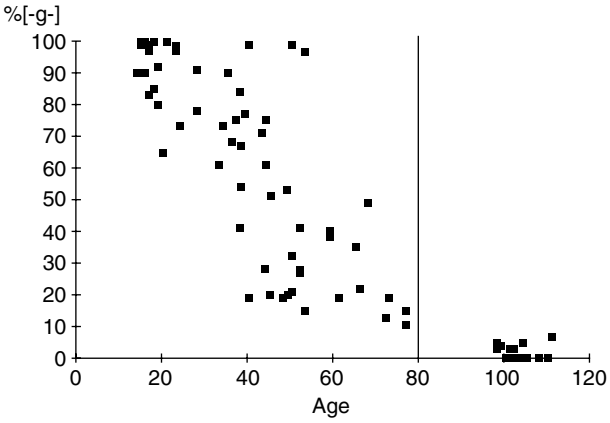


Figure 8.3 Apparent-time and real-time data: denasalisation of the velar nasal in Tokyo Japanese (from Hibiya 1996)

time. This makes it possible to test whether individuals indeed remain constant in their usage of variables across time. However, a panel study does not show us what subsequent generations do, to examine the progress of incrementation. The method that addresses this problem is the trend study, which examines successive cross-sections of the population at different points in real time. With reasonable comparability of the successive samples, especially with respect to the major social dimensions involved in linguistic change (age, class, and gender), a good trend study provides a moving picture of the change in progress, showing the generational advance of an innovative form.

Real-time studies, both trend and panel, have recently been a major area of research in language variation, as a result of the maturity of the field. Although it is rare to encounter a study in linguistics that is planned in advance to last for decades, what many linguists have done is to opportunistically replicate earlier research that indicated change in progress, in order to examine what has occurred in the community after the passage of a decade or more. In some cases this has involved new speaker samples, yielding a trend study, and in others, some original subjects have been recontacted, yielding a panel study.

One of the earliest (and still best) panel studies in sociolinguistic research builds on the Montreal French corpus, initiated in 1971 by D. Sankoff and G. Sankoff. Sixty speakers from the original sample were recontacted in 1984 by Thibault and Vincent (1990). Sankoff and Blondeau (2007) analysed a panel of thirty-two speakers who were recorded in both 1971 and 1984 for the use of the /r/ variable, which has been undergoing a change in Quebec French from apical to dorsal pronunciations. Of particular interest are the data in Figure 8.4, which shows the personal trajectories with respect

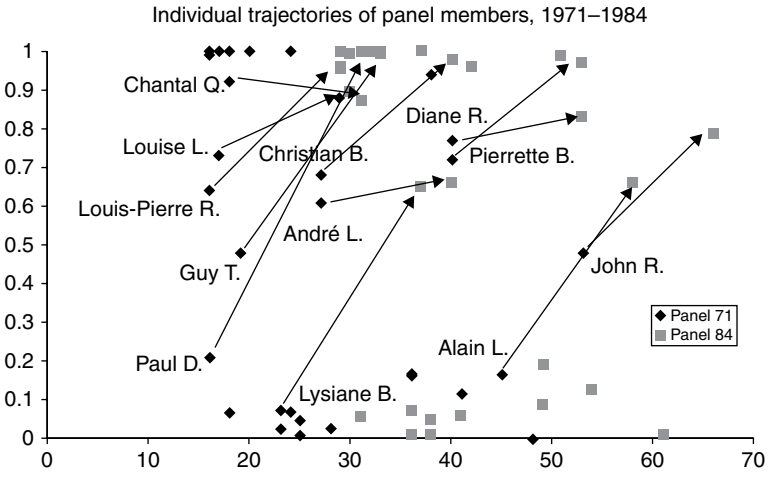


Figure 8.4 Real-time panel study of Montreal French speakers' use of [R] for /r/ (from Sankoff and Blondeau 2007)

to the change of thirty-two speakers across thirteen years. The figure shows that individual constancy is indeed the norm for speakers who were beyond the age of twenty-five when first recorded. Of speakers younger than this, those who markedly favoured one or the other variant when first recorded also show constancy of usage in later years (with one exception, Lysiane). The speakers who shift markedly in their usage are those who had intermediate rates of usage in the earlier sample, and who were young when first recorded. Only two significant exceptions to these generalisations are evident (Alain and John).

One striking consequence of the mechanism of change by cohort incrementation is that the S-curve of the age distribution of change in apparent time is also replicated in the real-time advance of change. Since trend studies with socially stratified samples are a recent methodological development, few of them exist with more than a few decades of time depth to illustrate this point, but it appears clearly in longer-term studies using written documents. One example is found in Kroch's work on the rise of English periphrastic *do* in questions and negative declaratives (1989a, 1989b, 2000). The modern form of this construction first appears in late Middle English in variation with earlier constructions involving subject–verb inversion in questions (e.g. *Do you eat fish?* varies with older inverted construction *Eat you fish?*) and postverbal negation in negative declaratives (*I don't eat fish ~ I eat not fish*). This was a spontaneous innovation in English which has no equivalent in any of the neighbouring languages with which English had contact in the Middle English period. After *do*-periphrasis

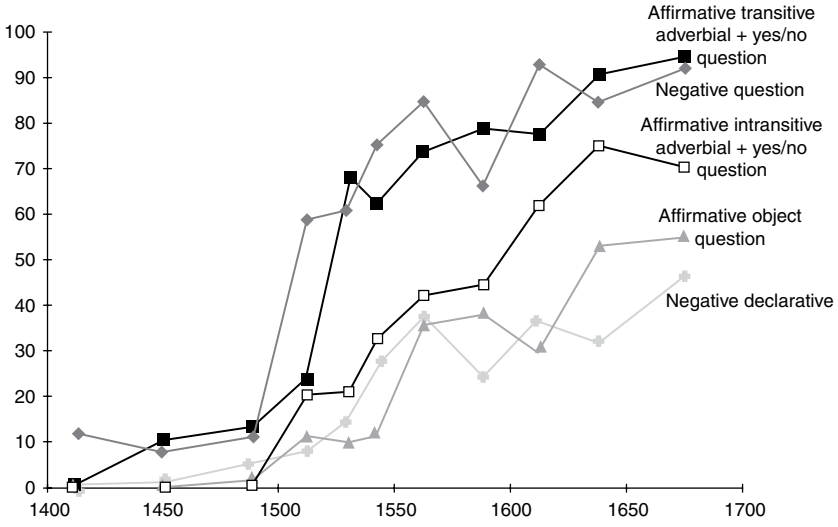


Figure 8.5 The rise of periphrastic *do* in Middle and Early Modern English (from Kroch 2000)

appears, it continues to vary with the older forms for about 400 years. Across that time span, the use of *do* rises along an S-shaped curve, as can be seen in Kroch's (2000) figure, reproduced here as Figure 8.5.

Five different contexts for the use of *do* are each plotted with a separate line on this graph. Although the data are somewhat noisy, causing a certain amount of jitter in the lines, each context progresses on a basic S-curve. Indeed, Kroch (1989b) shows that these contexts are all statistically equivalent to S-curves defined by a logistic equation. The noisiness of the data is to be expected in a study based on historical documents, which do not afford us controlled and stratified samples of community usage. The original data supporting this figure were collected by Ellegård (1953) from documents and manuscripts whose date of provenance could be reasonably well established. But what is not often known for such data is the author's sociolinguistic identity – age, sex, class, dialect background, and residential history, for example – dimensions along which usage of this innovation most likely varied. Since at any given point in real time, document writers in England included people who differed on all these dimensions, each data point in this graph is a partly random selection from a cloud of possible values that might have been obtained from other writers in the same community at the same time. Given generational incrementation of change in adolescence followed by individual constancy in adulthood, not knowing the age of the authors is a clear source of noise: if the source

documents in 1525 were written by young people aged about twenty-five, and the documents from 1575 were the product of elderly people of about seventy-five, the curve would appear to be flat in this time period, because both samples were drawn from the same generation, who did not change their usage in the interim. Consequently, the curves in Figure 8.5 track shifts in the limits and tendencies of the change, rather than the values associated with some specific reference point or community mean. Since, as we have noted, changes sometimes reverse direction, it is possible that not every change follows a smooth S-curve anyway, but the available evidence indicates that this is the dominant temporal pattern.

The accumulated evidence combining real- and apparent-time data thus confirms that changes advance in a community by incrementation among adolescents and young adults, and that after this age, individuals mostly stabilise their usage. But there are still many open questions, especially those involving causation and motivation. Why increment and then stabilise? We have suggested that incrementation is associated with adolescent identity formation: an incoming innovation has a sociosymbolic value as new and youthful, and serves to differentiate the innovators from their elders. But then why do individuals stabilise their usage in adulthood? This is subject to various interpretations. It could have something to do with linguistic maturation, much as is argued for so-called 'critical-period' effects (the decline in the ability to achieve native-like competence in languages learned in adult life). This account appeals to neuro-biological factors. But there is also a recognisable psycho-social element: individuals are purposeful agents, who always command a range of styles and registers, and always vary their usage, including use of innovations, for purposes of accommodation, contrastive differentiation, identity construction and performance, and so on. Hence it is also possible that individuals associate their rate of use of innovative forms with a generational identity they wish to preserve across their lifespan. This explanation would link age stratification of language with the age stratification that is evident in clothing and hairstyles, music preferences, personal adornment, and other social behaviours that express generational identity.

Recent work on language and identity draws attention to the complexity of the social meanings and motivations of linguistic innovations, and their use by individuals in identity construction (see Moore, this volume). Individuals make personal choices about the use of variables to show affiliations with groups, to express personal stances towards hearers or situations, and to reference social interpretations and evaluations that may be used to index identity traits. Linguistic innovations of all sorts provide rich material for this elaborate orchestration of personal identity; it is therefore perhaps remarkable that broad social trends of the kinds we have identified (e.g. the temporal S-curve, the curvilinear class pattern) consistently emerge.

8.4 Variation and change and historical linguistics

The deepening engagement of sociolinguistics and variation studies with language change has brought these fields into close contact with the traditional discipline of historical linguistics. Initially, these were complementary approaches to diachronic questions. Historical linguistics focused on real-time data, used written evidence (necessarily, since no sound recordings of speech existed prior to the late nineteenth century), and dealt with completed changes across large-scale time spans – change across centuries and millennia. The focus on completed changes typically implied invariant, categorical models and descriptions. Variationist studies introduced the use of apparent-time evidence, focused on speech, and dealt with changes in progress, over shorter time spans – decades and generations – using quantitative models and descriptions. But more recently, this neat division of labour and focus has eroded. Variationist analyses have been conducted of documentary evidence from times long past. Quantitative approaches have been brought to bear on completed long-term changes using written materials to study their time course and the variation that occurred while they were in progress. Sociolinguistic models of the mechanisms of change have illuminated historical questions. Whereas traditional historical linguistics sought, in effect, to use the past to explain the present, the addition of the variationist perspective to diachronic research has also, in Labov's words (1994: 9), made it possible to 'use the present to explain the past'.

Particularly noteworthy examples of the fruitfulness of this fusion of variation and diachrony have occurred in research on historical syntax, such as the previously mentioned work of Kroch on English, and other studies on languages as diverse as Yiddish (Santorini 1993), Greek (Ann Taylor 1994), and Portuguese (Tarallo 1996). Phonological change is less amenable to this kind of approach, because of the limitations of orthographic evidence; nevertheless, some fruitful work has been undertaken, such as Toon's study of 'the politics of early Old English sound change' (1983).

The principal impact of variation research on historical linguistics, however, may be less methodological and empirical, and more theoretical. The variationist perspective has had little impact on the comparative method and the reconstruction of proto-languages, but understanding that variation is an essential way station in the course of change has substantial implications for evaluating the plausibility of the changes that are postulated and their social settings. For example, change-in-progress studies show that the period of variation can last for a long time, and multiple changes may be underway simultaneously. This has implications for reconstructing the sequencing of events (e.g. chain shifts). Contemporary variables sometimes exhibit lexical conditioning or irregularity, with implications for the regularity of sound change. Studies of

the sociolinguistic types of change have implications for evaluating prior language contact. The transmission of lexical items, for example, implies borrowing as a primary mechanism of language change; hence in England after the Norman conquest, the huge inventory of French loanwords in English implies that native speakers of English were borrowing from French, while the paucity of phonological and syntactic effects suggests that Norman accents had very little impact on the English of descendants of the conquerors who underwent language shift. By comparison, English in India, which shows phonological characteristics common in Indian languages, such as retroflex consonants, and syntactic phenomena such as an invariant tag question '*isn't it*', is a case in which the main mechanism of change was imposition.

A significant consequence of the variationist perspective in historical studies has been the development of quantified models of change. Yang (2001) is a noteworthy example of this trend. Treating syntactic change as the product of the interaction between the distribution of syntactic structures in the input and the choices that child language learners face in their construction of a mental grammar, Yang proposes a probabilistic model of grammar competition that drives change forward along an S-shaped time course. The model crucially depends on variation: child language learners do not construct a single, static, invariant grammar to account for all the facts they encounter; rather, they entertain multiple alternatives, and select among them probabilistically.

Another significant contribution of variationist studies to historical linguistics is the refined view they permit of the stages of change. Conventional historical studies, relying on reconstruction from fragmentary evidence, typically account only for the endpoints of a change; a diachronic statement like $x \rightarrow y$ tells us that an early form x is realised centuries later as y, but gives no perspective on what happens during the intervening years. But synchronic studies of changes in progress make it possible to investigate triggering events and onsets of change (the actuation phase), and subsequent expansion of the innovation (the implementation phase).

Actuation appears to involve both social and linguistic factors; thus Labov (2010) attributes the original generalised tensing of /æ/ in the Inland North of American English to a social event: the early nineteenth-century mixture in north central New York state of speakers coming from several different dialect regions (including New England and southern New York), during the construction of the Erie Canal. These source dialects had different contexts for /æ/ tensing. The new communities that emerged from this mixture koinéised these conflicting patterns by tensing /æ/ in all contexts. The completion of the Erie Canal provided the pathway to settlement of the Upper Midwest, disseminating the new vowel phonology across a wide area. The tensed /æ/ vowel subsequently raised, vacating the low front corner of the vowel space; this provided a linguistic trigger for the fronting of the other low vowels. The implementation

of this change had far-reaching effects on the vowels of this region, following the linguistic principles of vocalic chain shifting discussed in section 8.2.2, and ultimately yielding the complex vowel rotation known as the Northern Cities Shift (LYS, also Labov, Ash and Boberg 2006).

8.5 Change and linguistic theory

Why does language change at all? Accounting for language change in linguistic theory is a long-standing problem in linguistics, dating from at least the Neogrammarians. As we have noted, Saussure famously denied the relevance of diachrony to synchronic linguistic theory, and his position has been widely emulated for a century. Nevertheless, linguists of all theoretical camps have been unhappy with this drastic division of the field, and many have attempted (even if uneasily) to model change within whatever theoretical framework they favoured. Thus in the structuralist framework, sound changes were characterised as involving phonemic mergers, allophonic splits, alterations in phonetic values, and the like (cf. Hoenigswald 1960). In the generative period, changes were described in terms of rule additions, losses, reorderings, and so on (cf. King 1969). Recent theoretical developments such as optimality theory (e.g. Anttila 1997) and exemplar theory (e.g. Bybee 2001) have often sought to explicitly incorporate accounts of linguistic change within their models.

Optimality theory has proven to be an exceptionally flexible framework for modelling change. The theory postulates a universal inventory of constraints, each stating some desirable phonological state of affairs; where languages differ is merely in the hierarchical rankings of these constraints (plus, of course, differing lexical inventories). Since any change in ranking defines a different grammar, and a different potential 'language', both variation and change can be subsumed into the OT account of language typology. Variable realisation of a final consonant, such as final –s and –r deletion in Caribbean Spanish and Brazilian Portuguese, and final –t deletion in English and Dutch, are modelled as variable rankings of constraints that militate against syllabic codas and those that favour faithful surface realisations of underlying segments. When faithfulness constraints are more highly ranked, the segment surfaces, but when the 'no coda' or 'simple coda' constraints prevail, surface realisations without the final segments are preferred. This is the typological difference between languages with open syllables (e.g. Yoruba), and those with closed syllables (English, Spanish, etc.), so the same mechanism can be pressed into service to account for variation and change. Variation is modelled by postulating variable ordering between the relevant constraints, and change across time is modelled by postulating a diachronic reordering of the relevant constraints (see, for example, Anttila 1997, 2002a, 2002b; Kiparsky to appear).

Exemplar theory is a recent development that places variation and change at the core of the model, relying on the naturally occurring variation in the input as the driving force (cf. Bybee 2001; Pierrehumbert 2002). This theory eschews abstract representations, postulating instead that speakers remember, in rich phonetic detail, the tokens of words that they hear pronounced, or produce themselves. Therefore, speakers have memories of the full range of variants they have encountered, and use these memories (the 'exemplar cloud') as targets for their own production, which then necessarily varies as well. The theory emphasises natural phonetic processes such as lenition and assimilation as the driving force in phonological change; words that are often repeated are more subject to these processes, altering the exemplar clouds of speakers in the direction of the change produced by the process. This model emphasises the importance of lexical identity and lexical frequency in variation and change, predicting that lexical items may differ (i.e. lexical diffusion), and that frequent words should lead sound change. Lexical diffusion has long been advocated in historical linguistics as an alternative to the exceptionless sound change model of the Neogrammarians (see Wang 1977; Labov 1981; Phillips 2006), based on a number of empirical cases where lexical irregularities are found in historical changes. Exemplar theory provides a formal model to account for such facts.

The most widely used theoretical framework in studies of variation and change, growing out of Weinreich, Labov and Herzog 1968, is the 'variable rule' (VR) model, a broadly generativist model in which optional elements in a grammar are probabilistically quantified (see Cedergren and Sankoff 1974 and Sankoff 1978 for further discussion). This is the dominant model in variation studies, and its extension to modelling change is straightforward, but has subtle and substantive implications.

The VR model postulates that any variable process may be subject to two conceptually different quantitative forces. First are contextual conditions: most variables have a lumpy distribution across the language, occurring often in some context and rarely in others: tensed and raised variants of /æ/ in English dialects, for example, are more common in pre-nasal contexts, and rarer or less advanced before voiceless stops. Deletion of final –t, and –d in English is more frequent before a following word beginning with a consonant, and rare before a following vowel. But second, there are also overall differences in the rate of use of any given variant, in different speakers, social-class groupings, speech styles, age cohorts, and so on. A particular dialect may have tensed /æ/ more frequently or more advanced phonetically than another dialect, even though both favour tensing in the pre-nasal context. A working-class speaker may delete /–t,–d/ more than a middle-class speaker, even while both delete more before consonants than before vowels.

This distinction between overall rate of use and contextual effects is captured in the VR model by two kinds of factors. Each process is associated with

an 'input' probability, or p_0, which captures overall rate of use. In addition, a process may be associated with multiple contextual constraints, capturing the quantitative effects of favouring and disfavouring environments that promote or retard the selection of a particular variant. These are factor weights or partial probabilities associated with contexts, p_i, p_j, p_k, etc.

Given this model of variation, what changes across time is typically the overall rate of use of the innovative variant. Just as speakers and social groups differ in overall use while preserving the same constraint effects, and vary their overall rate in different speech styles while leaving contextual effects unaltered, successive age cohorts across the course of a change will increment the overall rate of use, leaving context effects unchanged. Change is change in the value of p_0, while the constraints on variable selection (p_i, p_j, p_k …) do not change.

The constancy of contextual effects across time has been demonstrated in a number of empirical studies, beginning with the work by Kroch illustrated above in Figure 8.5. Kroch formulates this observation as his 'constant rate hypothesis' – the claim that the rate of change in all contexts is the same. Kroch shows that the rate rises in English periphrastic *do* in all the contexts investigated in the figure are mathematically equivalent; that is, the logistic transform of each of the curves is a straight line with an essentially identical slope. Therefore, the most plausible interpretation is not that each context represents a separate change proceeding at an independent pace, but rather that there is only one change, following a single time course, governed in variable rule terms by a single p_0.

Syntactically, this single change can be described as a loss of V-to-I (verb to INFL) raising; in old and early Middle English, in sentences without auxiliaries, a main verb could move up to the high position (a.k.a. INFL) in the clause an auxiliary would occupy, thus preceding a negative, for example (e.g. *They know not what they do*, with main verb *know* preceding *not*, parallel to *They must not know*, where auxiliary *must* precedes *not*). In Modern English, however, a main verb cannot occupy that position; instead *do* is inserted as a dummy auxiliary just when the main verb becomes separated from that position by some other material, such as a negative (You INFL not eat fish → *You* do *not eat fish*), or an inverted subject (INFL you eat fish? → Do *you eat fish?*). The several contexts of the change show differences in their intrinsic favourability to the innovative variant that are stable across time. Each successive age cohort across the 400 years of the change was less and less likely to permit V-to-I raising, triggering the alternative solution of *do*-periphrasis at progressively higher rates.

The constant rate hypothesis follows directly from the VR model, distinguishing overall rates of use from contextual effects. Indeed, the constancy of the rate of change across different contexts constitutes important evidence in favour of VR for both variation and change. Alternative models of change,

such as the OT treatment involving constraint re-ranking, lack any overall parameter comparable to p_0. This implies that change in an OT model should not be a smooth S-curve, but rather, a step function with inconstant contextual rates: each time a pair of constraints is re-ranked, the contexts they affect should show abrupt changes in the rate of occurrence of the variant realisations, while unaffected contexts would show no change. This is at odds with the empirical evidence.

8.6 Conclusion

Work on language variation has, since its earliest inception, addressed questions of language change. Nearly fifty years of research on these problems has turned up a substantial body of knowledge demonstrating that variation and change are in essence a single phenomenon viewed from different perspectives. This discovery requires linguists to develop new methodologies and theoretical approaches that make possible an integrated understanding of what the orthodoxy of twentieth-century linguistics treated as belonging to opposed and unrelated synchrony and diachrony. This is part of a broader integrative trend in twenty-first-century linguistics, bringing the insights of many disciplines together to tackle big issues that they were unable to resolve separately. At the centre of this integration are questions about stability and dynamism in language: why do languages change, and why do they (sometimes) remain the same? These are the questions that research on variation and change is helping to answer.

8.7 Where next?

Readers interested in following up this topic with further study would do well to examine the three volumes of Labov's *Principles of Linguistic Change* (1994, 2001, 2010), which provides an extended treatment of many of these issues. Condensed discussions of the social distribution of changes in progress may be found in Labov 1980 and Guy *et al*. 1986. The sociolinguistic typology of change is treated at length in Thomason and Kaufman 1988 and Van Coetsem 1988, and more succinctly in Guy 1990. Eckert 2000 is a classic source for the relationship of variation and change to social identity. A thorough discussion of the question of regularity in sound change is found in Labov 1981. Sankoff and Blondeau 2007 provide an excellent discussion of the relationship of real and apparent time evidence.

9 Variation and forensic linguistics

Frances Rock

> When a fact has to be impressed on a jury, the most effective method is to go on repeating it over and over, varying the wording, if possible, but relying on reiteration to do its work … The greatest advocates have not disdained to use this simple method to the full and have distinguished themselves from their less accomplished brethren by the variety which they could give to the performance.
>
> Philbrick 1949: 5–6

9.1 Introduction

Variation is crucial to our understanding of the language of individuals and groups. Through its study, we can find out about aspects of who a speaker is, where they are from, what they do, with whom they spend time and even with whom or what they affiliate. Many of these issues are of interest to the law. The police or courts might have a voice recording and a desire to know something about the identity of the speaker, for example. They might have a written police statement and a desire to know something about its situation of production. The police or courts might want information about the influence of the varieties which they themselves use for investigative work, such as interviews, or the varieties which they use for procedural work, such as explaining jury duties. Applying research about varieties to legal settings is exciting work. Such work with, on or for lay people or legal specialists provides opportunities to contribute directly to the operation of justice (Rampton 1992). Of course, this also means that this form of applied sociolinguistics is extremely responsible work.

This chapter explores ways in which the study of variation has been used in legal settings, either to provide evidence or to comment on or influence legal systems and their texts.

I am very grateful to members of the Centre for Language and Communication Research at Cardiff University whose collegiality and support has made it possible for me to write this chapter. I would also like to thank the editors of this volume, who have also provided helpful and constructive feedback on previous drafts of the text. I am delighted to have benefited from their experience and insight. Finally, thanks to the Arts and Humanities Research Council for a period of funded research leave during which this piece was written (AHRC Reference: AH/G007926/1).

9.1.1 Variation and forensic linguistics

Over the past forty years, linguists' knowledge of variation in relation to a range of linguistic and social variables has come to be applied to legal problems in three main spheres.

Firstly, variation is relevant to the functioning of legal systems because the law operates through language – laws are encoded through written documents and enacted through writing and speech. Particular varieties predominate and, as in other social arenas, take precedence in terms of the frequency with which they are heard and the status, roles and power of the speakers and writers who use them. The text below, from the Special Educational Needs and Disability Act, exemplifies:

> Nothing in this Act shall impose any charge on the people or on public funds, or vary the amount or incidence of or otherwise alter any such charge in any manner, or affect the assessment, levying, administration or application of any money raised by any such charge. (HMSO 2001) (Section 43, paragraph 13)

Subordination, co-ordination and the resulting long sentence as well as lexical choices like *incidence*, *manner* and *levying* are just a few of the features which makes this instantly recognisable as a legal text. This variety creates particular relationships with readers.

Secondly, the sociolinguist has become involved with providing expert reports to police and courts when matters of variation become relevant to criminal cases. For example, Svartvik (1968), in perhaps the first critical investigation of a police statement, asked whether a single individual could possibly have produced excerpts like the two below, during a single interview:

> She never said no more about it
> She was incurring one debt after another

He contrasts the multiple negation in the first excerpt with the use of *incurring* in the second to suggest multiple authorship (1968: 22–4).

Finally, linguists have examined variation in texts which are shaped by the law but operate outside legal arenas. These texts remind us that we are never far from the law in any social setting – even activities in own homes are subject to law – contracts and other documents which establish relationships, rights and responsibilities reach out from legal sites. For example, the train ticket which permitted my travel to work today states:

> Travel is subject to the National Rail Conditions of Carriage (NRCoC) and to the conditions of carriage of other operators on whose service this ticket is valid.

Even this short text features intertextuality, the use of an acronym and formal lexis and syntax (e.g. *subject to*, *on whose service*) which carry a legal variety into a service relationship.

Some would say that it is sociolinguists' social responsibility to investigate these contacts with law and its activities by asking:

- how the status and operation of particular varieties influences individuals' progress through criminal justice systems and, ultimately, their access to justice;
- how information about varieties can help legal fact-finders when they confront language issues and language evidence;
- how variation can influence the success with which information is communicated beyond legal systems, and that information's force.

Many people who study language as evidence or language and legal systems would not express their interests in terms of a focus on variation because, in this applied area, the effects of variation are more often the expressed focus of research than variation itself. Yet, as this chapter will show, a great deal of forensic work is underpinned by scholarship in variation.

9.2 Varieties and identification: linguistic evidence

Would you be able to recognise a close friend on the basis of only their speech? Would you recognise them from just their writing? What about someone you knew less well? Or someone you had only met once? These questions concern linguistic identity. They begin to indicate influences on speakers' and writers' distinctiveness and influences on listeners' and readers' abilities to discriminate between individuals. Perhaps the most obvious way in which variation is relevant to forensic settings is when language becomes evidence and variation potentially becomes a way of distinguishing between individuals. Forensic linguists have worked with differences in accent and dialect as well as considering the influence of variation according to occupation, gender, age, and ethnicity.

The happy fact for forensic linguistic practitioners is that there are differences between different speakers and different writers. However, this fact is accompanied by a number of unhappy facts. For example, we do not know with any certainty how constant an individual's speech and writing are and we cannot measure intra-speaker variation sufficiently robustly. Likewise, degrees of difference between individuals are not measurable with fine gradation. There is disagreement about which aspects of speech or writing are likely to be the most reliable, valid markers of identity and which are likely to be uninteresting or, worse, misleading. Furthermore, texts for forensic examination tend to be very short, providing scant data for analysis. These facts, in combination, mean that any examination of language conducted for police, courts or others must be undertaken both carefully and with adequate caveats. In this section I describe and exemplify some of this work and, to conclude, consider why linguists are willing to provide forensic evidence despite these unhappy facts.

This will help to illuminate the variationist-forensic relationship. This relationship exists because matters of identification, the bread-and-butter of investigative forensic linguistics, are also at the sharp end of theoretical debates within sociolinguistics about style, styling, inter- and intra-personal variation, diachronic language change, the influence of context and co-text, genre, topic, setting and interactional goals, as well as speaker characteristics (such as social class). The study of variation in forensic settings, as elsewhere, rests on the assumption that variation is not arbitrary. The extent to which one subscribes to this view ultimately dictates one's theoretical view of variation. This, in turn, determines one's views on the potential of language evidence for identification, the degree of accuracy of identification and the kinds of variation which can reveal identity (see Nolan 1999; McMenamin 2002b for further discussion).

9.2.1 Forensic phonetics

The development of forensic phonetics is charted and evidenced by book-length studies devoted to it (Nolan 1983; Hollien 1990; 2002; Rose 2002). Forensic phoneticians are, very generally, concerned with occasions when the sounds of language can become relevant to a criminal or civil investigation. As French (1994) explains, their activities can be organised under five headings. A brief review of these will highlight some key directions and developments in each in relation to variation:

Speaker comparison (also known as speaker identification) – This involves investigating who might be speaking on a suspect recording, such as a threat, from a restricted set of possible speakers. This kind of examination initially evolved differently in the USA, where technology reigned in the form of the spectrogram (acoustic analysis), and the UK, where analysis by the ear of trained phoneticians predominated (auditory analysis). These traditions led to the emergence of a joint auditory-acoustic approach which used both instrumentation and close listening. This is currently the most active area of forensic phonetics yet, as Jessen has pointed out, must be approached with care, an array of published research, personal research and, if necessary, advice because 'it is near to impossible to be a specialist in all the dialects of the target language, know all about its sociolinguistics, about speech pathological conditions, second language phonetics, and so forth' (2008: 674).

Determination of unclear or contested utterances – Here the phonetician opines on what is said in an unclear recording. Again, both instruments and careful listening feature and methods are frequently developed and appraised (e.g. Howard *et al*. 1995). There are many reasons that recorded utterances may be difficult to hear. The recording may have been made via a particular device such as a landline (Künzel 2001; Lawrence, Nolan and McDougall 2008) or mobile telephone (Byrne and Foulkes 2004; Guillemin and Watson 2008) or

in particular conditions such as out of doors. Additionally, speaker behaviour may have contributed. The speaker may have been shouting (e.g. Blatchford and Foulkes 2006) or deliberately trying to disguise their voice (Zetterholm 2003; Neuhauser 2008), for example. Of particular relevance to variation are cases in which a recorded utterance is unclear because the speaker's variety is unfamiliar to naïve listeners and the linguist is able not only to provide insight into what is said but also to explain how, through knowledge of accent and dialect, they have reached that conclusion.

Authenticity examinations of audio recordings – This concerns whether an audio- or even video-recording has been tampered with through deletion or addition of material, for example. The recent, high-profile, US case in which O. J. Simpson faced charges including kidnapping shows that both analogue and digital recordings can be at issue. The trial in that case heard evidence about two audio recordings. The first, a digital recording, was made covertly during a meeting between Simpson and two sports memorabilia dealers. In the recording a voice could be heard making violent threats but the analyst was unable to confirm that the recording was authentic. The second was an analogue recording in which it was alleged that the attack was planned. The analyst was able to validate that this micro-cassette was genuine, but of course further work would be needed to consider the identities of the speakers on the tapes (only news reports at present give details, e.g. Arseniuk 2008; Elsworth 2008; Powers and Ryan 2008).

Speaker profiling – This typically contributes to the investigative phase of a case rather than trial. It involves analysis of a speech sample in order to provide information about the sample and, by implication, the speaker. Probably the most cited case of this type is that reported in Ellis (1994; see also: Windsor Lewis 1994; French, Harrison and Windsor Lewis 2006). A series of violent murders during the late 1970s in northern England led to an unprecedented police investigation and the use of a number of rather novel forensic techniques including linguistic analysis. Ellis's involvement centred on several audio-recordings, apparently sent by the killer, which taunted police and, through the voice, offered potential clues to the speaker's identity. This case is of particular and obvious interest in the context of variation: indeed, Ellis's experience on the Survey of English Dialects was a major influence on his work (e.g. 1956). His investigation involved scrutiny of the accent and dialect on the tape, compilation and analysis of a collection of non-standard speech samples from adults and children local to the area which Ellis had identified as the speaker's likely home, and interviews with speakers from that area. His analysis proved extremely successful in that the speaker was from the area that Ellis had identified. However, the speaker on the tape and the killer turned out to be different people – the speaker was a hoaxer, clearly something beyond the linguist's control.

Naïve speaker recognition or earwitness evidence – So far, we have observed how expert phoneticians examine speech. Naïve speaker recognition, as the name suggests, involves instead untrained observers. The alternative title 'earwitness evidence' highlights that, like eyewitnesses, 'earwitnesses' may encounter an obvious offence in progress (e.g. an armed robbery during which robbers shout instructions) or something which appears mundane and only later arouses suspicion (e.g. an overheard plot, disguised in innocuous code). As the term 'earwitness line-up' suggests, this form of identification parade is not unlike an eyewitness line-up – several speakers are presented via recordings, and the earwitness is asked to try to identify the voice they heard during the crime. Yet differences between auditory and visual processing, for example, render deriving methods for earwitness line-ups from the eyewitness procedure problematic (Hollien 2002: 94). Like eyewitnesses, earwitnesses are extremely believable in court yet can be easily swayed by suspect presentation or simply mistaken (Hollien 2002: 92–3). Phoneticians have made recommendations about the conduct of voice line-ups (e.g. Hollien 1996; Yarmey 2001), yet there is more to do here in terms of investigating earwitness characteristics and abilities, exploring elicitation of earwitness evidence and educating investigators about earwitnesses' strengths and limitations.

9.2.2 Authorship

Just as speech-sounds evidence is used in legal investigations, a growing literature examines the extent to which written language can reveal identity in texts which are plagiarised (*taken* from an author without consent) (e.g. Johnson 1997; Turrell 2004) or disputed (typically, *assigned* to an author controversially) (e.g. Coulthard 1994b, 2002). Two broad approaches have developed. The first is essentially qualitative and involves closely scrutinising the disputed or 'suspect' text and comparing it to 'known' writings from either the claimed or disputed author in order to identify points of difference which are sufficiently noteworthy, in one way or another, to indicate authorship (e.g. McMenamin 2002b: 109–22). The second is essentially quantitative and involves examining large corpora in order to devise and test 'markers of authorship', in other words to seek to establish whether, for example, frequent use of a particular lexical item will identify an author and whether it will do so more successfully than frequent use of a particular grammatical feature (e.g. Chaski 2001; Grant and Baker 2001). Quantitative and qualitative approaches have never been truly separate in that qualitative work of this type ultimately depends at least on a sense of the predictable in texts, whilst quantitative work is often initiated by questions arising from close textual scrutiny.

Qualitative–quantitative interaction is exemplified by Coulthard's collaborations with other scholars. For example, Coulthard has provided expert

evidence in cases involving disputed police statements – written texts which police claim accurately record suspects' words and which suspects claim have been fabricated by police. Coulthard's work hinges on variation in that, where possible, he identifies features of disputed statements which appear typical or atypical of the suspect's variety, thus suggesting authorship. He has discussed variation in characteristics of lexical items, adjectival choices and arrangements and grammatical structures. He ultimately strives to work at the level of idiolectal variation (e.g. Coulthard 1992, 1995, 2002, 2004; Coulthard and Johnson 2007: 161–73). One statement, which was supposedly written down verbatim – in the suspect's own words – contained the grammatical characteristic of repeatedly placing temporal *then* after the subject rather than before, for example, *I then ran out after them* rather than *Then I ran out after them* (Coulthard 1993,1994a). For Coulthard, this seemed like the language of the police, rather than the suspect. This was supported by analysis of a corpus of police language which showed that this post-positioned *then*, along with other features such as very precise times, use of passive voice and formal vocabulary, all characterised police language (Fox 1993). As work on identifying and using quantitative authorship markers continues (McMenamin 2002a; Grant 2007) alongside closely related work on variation (e.g. Johnstone 2000, 2007) the further developments likely in this area have the potential to be firmly grounded.

9.2.3 Asylum and language analysis

What happens when activities which might have been best undertaken by, or with, researchers with a sound understanding of variation and its implications become the territory of people who appear to lack that understanding? Unfortunately the answer to that question is provided by some studies of language analysis to determine nationality.

If an individual faces persecution in their country of nationality they may travel to escape that persecution. On arrival in a new country they will need to demonstrate that they are fleeing genuine persecution. They will, of course, do this through their narratives' content but in many countries (such as the Netherlands, Australia and Great Britain) they will also need to prove their claimed origin and personal history through their narratives' form – the variety they use during asylum interviews (Eades and Arends 2004: 179–80). Those who examine asylum seekers' speech, using 'language analysis' are clearly working with variation. Their basic assumption, that variation can indicate identity, looks very familiar to sociolinguists. However, the details of language analysis have been called into question as 'linguists are increasingly raising concerns about over-generalized and erroneous assumptions' from those who perform the analyses. Particularly worrying is a dominant assumption that asylum seekers will use only one variety without any influence from other varieties and

that each variety is intimately, and straightforwardly, tied to particular places or communities. Thus, some language analysts are prone to ignore scholarship on such topics as bilingualism, use of a lingua franca, speaking with an interpreter, accommodation and code-switching (Eades and Arends 2004: 180–1). Books and articles which describe and problematise current procedures have raised consciousness (e.g. Blommaert 2001) and *The International Journal of Speech, Language and the Law* has asked how linguists can contribute to determining refugee status and improving language analysis through a special issue (11, 2). Some linguists have been sufficiently concerned that they have produced a set of guidelines, intended to provide for knowledge about variation to enter legal procedures around asylum seekers (Language and National Origin Group 2004). Even low awareness of the diversity of means of expression in different varieties causes practical problems, as this exchange, from Maryns (2006: 229) between an interviewer, asylum seeker and translator, illustrates:

INTERVIEWER:	how many miles (..) do you know (..)
ASYLUM SEEKER:	I don't know miles
TRANSLATOR:	how much how much kilometre (.) mile (.) how much mile from Kabala to urm this urm usay you work
ASYLUM SEEKER:	40 minutes

As Maryns notes, the asylum seeker is not familiar with expressing distance in the measurement units offered. Without the resources which both the interviewer and translator expect him to use, he appeals to a more familiar means, time.

9.2.4 Presenting linguistic evidence

Literature surrounding linguistic evidence is of two sorts. On one hand, authors produce case reports, articles and books based on their direct, data-driven case experience (e.g. Shuy 2005). Thus they present methods for academic scrutiny and make their ideas available to future investigations. The other kind of publication is based in something closer to 'blue-skies' research – research which is not tied to particular cases but explores topics which might assist future casework by increasing its rigour and breadth (e.g. McDougall 2004). Identifying topics for blue-skies research is perhaps easier in areas like forensic phonetics, where a *relatively* limited set of questions are asked *relatively* frequently, than in fields like discourse analysis, where different analytic concepts and methods might be needed for each case. We could see all research on language, including language variation, as blue-skies research because it becomes part of the body of work which equips linguists, of whatever specialism, every time they venture into the police station or courtroom. The courts, in deciding whether to attend to any expert, including linguists, require expert evidence, rooted in

excellent scholarship in recognisable subject disciplines. Increasingly, judicial systems seek to improve their use of experts by asking experts to be clear about what they know and how, and about whether the methods they use are robust (reliability) and the ways in which they use them are relevant to specific questions and cases (validity). In turn, linguists note the need to situate and explain the study of language and to show that their discipline has rigour (Cambier-Langeveld 2007; Eriksson and Lacerda 2007).

In my introduction to this section I noted that one's view of variation and of the difficulties in pinning down an individual's variety, the lack of constancy of any individual's speech and writing and the potential for different speakers and writers to speak or write 'the same way' might be seen as reasons for linguists to stay away from legal settings. However, Nolan's warning against this in relation to forensic phonetics applies to other forensic linguistic tasks too:

> The alternative to phoneticians and speech scientists taking part in the forensic process is not … that evidence on speaker identity and other 'forensic phonetic' aspects would play no part in court cases; rather, if phonetically competent scientists do not offer the help sought by courts it will be provided by others who have much less understanding of the complexity of spoken communication. (Nolan 1999: 747)

Howald similarly problematises methods used by non-linguists for analyses of authorship (2008). From this standpoint, attention to legal systems, their use of language evidence and their understanding of language issues is a responsibility of all linguists, not just those who work in this applied area. To this end, linguists have considered appropriate ways to present their evidence (summarised, for example, in Coulthard and Johnson 2007: 200–13) in the context of wider debates about the place of experts in legal systems and have produced guidance documentation (French and Harrison 2007).

9.3 Varieties and their influence: language in and of legal systems

9.3.1 Code choice by lay participants in legal systems

In the courtroom, spoken evidence has great significance. A witness's words can put a defendant at a crime scene, provide a compelling alibi and even corroborate or contradict others' words. Witnesses have little choice about whether to talk in court. Those who refuse behave in a marked way which is so institutionally unacceptable that it can attract serious sanctions (e.g. BBC 2002, 2008). The fact that many legal systems require victims and witnesses to testify illustrates a performative aspect to testimony. However, it is not only the words said but also the way in which they are said that can persuade those who try cases. One important and apparently influential aspect of how witnesses deliver their testimony is the variety they use.

The Duke Language and Law Project systematically examined variation in speech style in court and its effects on trial decision-makers and, ultimately, outcomes. The work of the project, based at Duke University, has been reported extensively (e.g. Conley, O'Barr and Lind 1978). The project team examined the influence of four dimensions of variation: powerless versus powerful language in the sense proposed by Lakoff (1975); hypercorrection, raised as an influential stylistic dimension by Labov (e.g. 1972b); turn-taking, particularly simultaneous speech, drawing on Sacks, Schegloff and Jefferson (1974) and narrative versus fragmented testimony, which they identified as important using lawyers' comments (presented in O'Barr 1982: 31–8). The project was embedded in a rich framework of data including extensive courtroom observation, audio-recording, note-taking, lawyer interviews and scrutiny of legal texts, particularly lawyers' training manuals. Lawyers' comments and their educational texts illustrated assumptions about varieties in the courtroom; for example, that '"English-speaking" jurors understand "English-speaking" witnesses [and lawyers], regardless of cultural background and differences in dialect'. This aspect of the research also showed that lawyers' training encouraged them to speak, and have witnesses speak, in particular ways, yet the law itself would not recognise style as having any influence on its processes or outcomes (O'Barr 1982: 31–49). Contradictions like this provided questions to frame and stimulate the analysis and ways to interpret courtroom data.

O'Barr and colleagues' examination of powerless speaking style, perhaps the part of the project which is most cited, shows the usefulness of variation to their study. Lakoff's model of women's language (WL) asserts that particular features occur with higher frequency among women than men, for example, hedges (*I'm **kind of** glad to be reading this book*) and tag questions (*this chapter is in this book, **isn't it**?*). The Duke Project examined courtroom transcripts, finding that WL features were 'neither characteristic of all women nor limited only to women' (O'Barr 1982: 69). Indeed, WL features correlated more directly with speech from socially powerless individuals than women (although they noted women's tendency towards such powerlessness; O'Barr 1982: 70–1). Having established the presence of both powerless and powerful speech styles, O'Barr and colleagues' interest in the influence of speech styles in court led them to devise a series of psycholinguistic experiments. These investigated whether speakers who exhibited a high incidence of powerless features (men and women) were perceived differently from powerful speakers despite a universal (mock) courtroom setting and evidence-giving purpose. Their experiment used audio-recordings derived from naturally occurring testimony and delivered one of four circumstances: a female witness speaking in a powerless style; the same witness presenting otherwise identical testimony in a powerful style; and both the powerful and powerless styles delivered by a male witness. Ninety-six experimental subjects, assuming something akin to the

role of juror, were asked to rate the witnesses for convincingness, truthfulness, competence, intelligence and trustworthiness – criteria which are very relevant to credibility of testimony. The experimental results remain compelling. O'Barr and colleagues found that the female, powerful-style witness was rated more highly than the powerless-speaking female witness for all criteria, with varying degrees of statistical significance. Likewise, the powerful male speaker was more positively evaluated than the powerless male (O'Barr 1982: 71–5). The team similarly investigated courtroom variation in relation to the other stylistic dimensions. Although those findings are less striking for various reasons they nonetheless indicated further aspects of style which influence courtroom decisions.

O'Barr found 'the degree to which legal decision-makers altered their opinions about the relative credibility of witnesses on the basis of variation in their presentational "styles"' 'disturbing' (1993: 325). Many sociolinguists have taken up this concern, studying variation to reveal arbitrariness in justice and to educate about its avoidance. If the influence of one's own variety seems worrying, consider how much more potent it is to be assessed in court on the basis of someone else's variety. This is the situation experienced by individuals who must communicate with magistrates, judges and juries through interpreters. Interpreters alter apparently minor aspects of witness' testimony in ways which can have major effects on the impression they create in court (e.g. Berk-Seligson 1990, 2002; Hale 2004). Hale exemplifies variation in register and style caused by grammatical and lexical alteration. In the excerpt below, Spanish testimony is interpreted into English for an Australian court:

WITNESS:	… porque yo le prometí que no la iba a echar
[HALE'S GLOSS:	'… because I promised her that I wouldn't throw her out']
INTERPRETER:	and also I had promised her that I wouldn't evict her

Here, the interpreter has replaced *throw her out* with the more formal *evict*, achieving register shift (Hale 1997: 204–5). The example is not isolated. Hale notes the gap between the discourse of the courtroom (rule-governed, structured, ritualised and formal) and that of the lay witness ('everyday', vivid and detailed using implicature and indirectness). She shows that interpreters tend to 'bridge the discursive gap' by systematically shifting to the court's register when interpreting to them and the witness's when interpreting the other way. This unsettles 'the delicate balance of the adversarial system' (1997: 208) potentially advantaging the speaker on the stand. As Berk-Seligson has pointed out through matched-guise experiments which echo those of the Duke Project but add an interpreter, the interpreter's recasting of the witness's words has the power to influence hearers' perceptions of the witness's convincingness, competence, intelligence and trustworthiness, as the witnesses themselves did in the

Duke studies. This is the case even when the experimental subject-jurors are bilingual and able to understand the witness's original version (2002: 181–2).

Of course interpreters who style-shift (for example introducing hesitations) or register-shift do not necessarily do so in ways that benefit the witness or defendant who happens to be beside them. Berk-Seligson shows that interpreters alter the coerciveness of questions from lawyers, for example, typically reducing questions' pragmatic force. This means that witnesses answer questions which are less leading than those asked and, crucially, different from those which the court believes they were asked (1999). Only half of all leading questions were interpreted accurately in Berk-Seligson's data. The form of lawyers' questions has also been found to be altered by the addition or deletion of discourse markers, apparently systematically, according to whether they are used in direct or cross examination (Hale 1999). Selection of first or third person too, influenced by interpreters' stance towards speakers and by institutional norms, shapes limited English speakers' participation in their own trials (Angermeyer 2009).

Courtroom questions and testimony which are mediated through sign languages are just as open to less than literal interpretation as signing interpreters change the degree of specificity, alter yes–no into either–or questions and even add or remove items (Brennan and Brown 2004: 132–6). When individual courtroom participants begin code-switching things become particularly tricky for interpreters. Legal language in English-speaking jurisdictions, for example, is notorious for its use of Latinate expressions (e.g. Mellinkoff 1963: 71–82) and lawyers could be said to be notorious for the delight with which they use terms like *ex gratia* or *in loco parentis*. Interpreters may resort to omitting such items altogether (Moeketsi 1999: 164).

The role and influence of the interpreter becomes more opaque still in situations where not all courtroom participants agree on whether their presence is even needed. Cooke describes his experiences interpreting in an inquest following a fatal shooting in Australia's Northern Territory (where a relatively large proportion of residents are Aboriginal). He observed that many Aboriginal witnesses needed help to understand and be understood yet this was granted arbitrarily. In some instances an interpreter was assigned at a lawyer's request before the witness took the stand. Other interpreters would serve simply because they happened to have remained in the witness box after previous testimony. Some witnesses struggled without an interpreter (Cooke 1995). Such uncertainty frequently stems from a lack of knowledge on the part of legal personnel about varieties and their influence. It has systematically disadvantaged Aboriginal people. As Mildren explains, most English-speaking Aboriginal people in the Northern Territory speak English as a second language. Their English will fall along a continuum from the acrolectal (closest to the standard variety) to the basilectal (furthest from the standard) and in some situations courts may feel confident about whether an interpreter is needed. Crucially, though, even those who might appear to speak Standard Australian

English (SAE) might, in fact, use and respond to such linguistic features as silence and disagreement, and such paralinguistic features as gesture and gaze, very differently from those who dominate the Australian legal system – first-language SAE speakers (Mildren 1999: 138).

Typical misunderstandings between speakers of SAE and Aboriginal Englishes happen to be particularly detrimental in western legal settings where, for example, questions can function to elicit, seek clarification and even to accuse (Eades 1994: 241). For Aboriginal English speakers, questions take on a very strange significance as they are simply not a recognised interactional device. Information is exchanged in Aboriginal cultures in limited circumstances, some information is imparted according to whether the recipient is male or female and some is not for the uninitiated. The system of entitlements and give-and-take around information is so pervasive and influential that it is usefully characterised as a knowledge economy (Walsh 1994: 225–6). The communicative clash when SAE speakers question Aboriginal English speakers in the legal system can lead at best to the latter appearing unco-operative and at worst to them succumbing to gratuitous congruence – simply answering *yes* to every question – a particularly dangerous strategy in legal settings (Eades 1994: 244–5).

The authors cited above have undertaken descriptive linguistics around problems with the administration of Australian justice. However, their work has an applied dimension, seeking to improve that administration through activism, contact and commentary arising from their research or their work as interpreters or legal practitioners. Eades's most recent book (2008) shows how this effort continues.

Difficulty in delivering justice to speakers whose varieties may be problematised by lay people is not restricted to Australia (Shuy 2003). Creole languages developed in situations of language contact between a superstrate language, whose speakers typically boasted socio-political domination at the time of their formation, and one or more substrate languages, whose speakers were often subjugated, for example in the Caribbean alongside slavery. Creole languages can appear, to speakers of the superstrate language, to be 'imperfect' realisations of that language. Thus, as in Australia, interpreters might seem unnecessary. Brown-Blake and Chambers illustrate that contact between Jamaican Creole and English in the UK legal system can cause 'miscommunication or lack of communication' with serious legal implications (2007: 272). For example, use of a legal term *duress*, which exists in British English but not Jamaican Creole, causes the suspect in the interview below to make an admission about carrying drugs, swallowed in a condom, and even to specify the condom brand:

SOLICITOR: That amounts to duress.
SUSPECT: No, a no durex, a panta dem ina.
[English gloss: No, it was not a Durex, they [the drugs] were in Panther.]

Brown-Blake and Chambers 2007: 280

In Jamaica itself, the importance of the difference between Jamaican Creole and English has been highlighted by proposals to recognise language as a basis for discrimination and to outlaw such discrimination. Yet these have met with resistance because of far-reaching implications for the Jamaican legal system (Brown-Blake 2008).

Difficulties when encountering an unfamiliar variety can be exacerbated when that variety uses a different modality from that familiar to a victim, witness or defendant. This is the situation faced by many deaf people in those roles. The challenges of bimodality can be mystifying for courts which are rooted in the significances of speech and writing as they relate to testimony and statements. An interview conducted through a language like British Sign Language (BSL) which has no written form must be translated, for example into English, before it can be transcribed, introducing much potential for error (Brennan and Brown 2004: 119). Interpreting in court using BSL presents challenges for interpreters concerning register. Some express concern about whether deaf people will understand the legal register, others about whether BSL really has a formal register equivalent to that used in courts (Brennan and Brown 2004: 145). Brennan and Brown conclude that whilst BSL is functionally elaborated, interpreters are wary of using its formal range in court because of a desire to maintain objectivity. This manifests itself as conservative interpreting using a rather limited variety of BSL which is reinforced by strict court processes (2004: 146–8). Whilst the influence of language on the courts is often noted, this influence of the courts on language deserves much more exploration.

9.3.2 Code choice by legal participants in legal systems

This final section considers how variation relates to language produced by the legal system (e.g. legislative language), by those who act for the legal system (e.g. judges), or by those whose language use is influenced by the legal system (e.g. businesses drafting contracts). Here we are predominantly concerned with register variation.

Once upon a time, legal language was unashamedly impenetrable; a rite of passage for legal practitioners and, at best, a puzzle to lay people. Studies during the 1960s described this register and its functions (e.g. Crystal and Davy 1969) but also started to question whether legal language had to remain so challenging to the uninitiated (e.g. Mellinkoff 1963). During the 1970s, public disquiet with communication from a range of public and private institutions grew and was galvanised by organisations like the Plain English Campaign (2008). It was argued that language which is intended to communicate subjects' or citizens' rights and obligations should be 'easy' to understand rather than apparently using a variety designed to keep lawyers employed. Counter-arguments claimed, for example, that legal language was a historical necessity,

provided precision and ensured functionality for legal insiders (discussed in Tiersma 1999).

There have been several consequences of what remains an ongoing debate. One is that some legislation is now enacted in a legal register but also 'translated' in explanatory documents. An example is the Party Wall etc. Act (HMSO 1996) which makes provision for building work on property boundaries. Like other contemporary legislation, the Act itself is available online. However, the government also provides an 'explanatory booklet' also available online and in free hard copy (Department for Communities and Local Government 2004). In addition, a further format – an interactive 'Planning Portal' website – enables users to click on links according to their circumstances and questions (Department for Communities and Local Government 2008). Each source varies in lexis, syntax and organisation. Such simplified texts are apparently influential. For example, when the Animal Welfare Act (HMSO 2006) was enacted in early 2007, it received little press attention. However, when its content was transformed into Codes of Practice intended to help owners of cats, dogs and horses to 'better understand their duties under the Animal Welfare Act' (Department for Environment 2008) there was a vocal if bemused response from the press and subsequently the public.

As well as plain language activists, academics now frequently examine and critique 'difficult' language from a range of agencies. Indeed information design has emerged as an academic field and area of professional practice which unites such endeavours as graphic design, ergonomics and linguistics, seeking to improve understanding and use of texts (e.g. Pettersson 2002; Delin, Searle-Jones and Waller 2006). In addition, linguists have debated the merits of register shift aimed at comprehension and have investigated the consequences, perceptions and implementation of such shifts (e.g. Labov and Harris 1994; Heffer 2005; papers in Wagner and Cacciaguidi-Fahy 2008). One recent study examined how, in practice, register shift is accomplished repeatedly within legal settings, becoming not only the responsibility of legal drafters but also undertaken day-to-day by police officers when mediating information to lay people in police custody (Rock 2007). It becomes an interactional accomplishment, for example, as a police officer transforms part of the formal, formulaic police caution. The caution is issued during police–suspect interviews in order to explain suspects' right to silence. The excerpts below illustrate how several police officers, each in a different police interview, explain the words *Anything you do say can be given in evidence*:

OFFICER 1: these tapes can be used in court

OFFICER 2: anything you do say (.) I can actually tell the court about

OFFICER 3: there's two tapes sat there recording everything we say (.) if at a later date it goes to court (.) they can be used

OFFICER 4: it's all recorded on tape and the court can listen to that if they need to

These officers transform the abstract concept that words can become evidence by drawing on concrete aspects of the process through which words become evidence in this context – the tape recording (examples 1, 3 and 4), the activity of reporting speech (example 2), the court as a location (example 1 and 3) or collective of people (examples 2 and 4) and the activity of 'using' (examples 1 and 3) or listening to (example 4) a recording.

Other studies have turned this mediation around, examining the shifts in the other direction when spoken language from lay people – for example that gathered during statement-taking – is transformed into a written form which will perhaps be more useable by the legal system than the unmediated text (e.g. Gibbons 2001; Rock 2001; Komter 2006). Consider the following excerpts from a telephone call to the police requesting assistance.

CALLER: the owner of [the property] is known to you apparently because um what happened is er it was bought by a husband and wife but they've been divorced for years … since then he moved into it *[CT: uh hm]* and then he disappeared off to Gibraltar for three months *[CT: right]* in the meantime his wife has been trying to get the property back so she then gets a um (0.1) a locksmith to come along … and then he comes back from Gibraltar and he breaks in again *[CT: OK (chuckles) oh right]* so apparently you um she's been ((been been through)) yourselves

The call-taker turns to a colleague to seek guidance and glosses the situation:

CALL TAKER: it's an argument with a domestic next door

Domestic has a particular utility to the call-handler and her colleague, letting her condense a long narrative. However, at several points later in the call when she again needs to summarise, this time for the caller, she does not use *domestic*. Whilst the caller would probably have recognised the word's meaning, it appears problematic as part of legal–lay talk.

There is a need for further work on how texts travel and are transformed through legal processes, and whether this improves comprehension, persuades or serves other purposes. Variation offers a useful and under-exploited starting point for such work.

9.4 Conclusion

By definition, linguistic varieties differ from one another. At some level or levels variant forms of particular variables can be identified and distinguished across different varieties. Some aspects of forensic linguistics spring from an interest in the forms of this difference, others from an interest in its functions. Turning to form first, we have seen that the work of the forensic phonetician and authorship researcher exemplifies how formal aspects of linguistic and social variation, particularly realisations of particular variables, can be used by

legal practitioners to the extent that it provides information about individuals' identities. Other aspects of forensic linguistics are driven by function – both the courtroom studies and the studies of legal language examine ways in which variation can influence how law works or is understood. These issues of language and justice are constantly changing, intimately tied to fairness and politically charged (e.g. de Varennes 2003; Cardi 2007).

Inherent in forensic linguistics is the need to consider the extent to which a focus on the individual versus social groups will figure in particular instances. The traditional sociolinguist, when collecting data and conducting analysis, may focus on interpersonal and interactional levels; *or* large-scale quantitative levels, involving study of class or age, for example; *or* the mezzo-level associated with smaller social groupings such as communities of practice or speech communities. Those working on legal or evidential texts may find imperatives for shunting between these levels during their data collection and analysis. Thus, analysis of legal and evidentiary texts requires use of wide-ranging previous studies although researchers must avoid 'cherry picking' from these. There is a continuing need for focused, 'blue-skies' research provided by traditional and more postmodern variationist scholars to be used by forensic linguists to inform applied work.

If we see forensic linguistics as a microcosm of linguistics or sociolinguistics (e.g. Shuy 2006b: 3–4; 2007: 101) it is possible to recognise that studying variation as it figures in legal or evidential settings is not particularly different from studying variation in other areas of social life. Furthermore, language researchers working on education and workplaces have already applied linguistics research in two senses; firstly by transforming investigative tools to probe those settings meaningfully and secondly by converting research findings into recommendations for practice or at least feedback for practitioners. Forensic linguistics mirrors those endeavours and requires a constant dialogue with less obviously applied scholars of variation in order to understand and use current thinking on the language of individuals and groups. At the same time, for those who do not work on legal or evidential data but on variation, the legal arena can provide many questions which can usefully be investigated outside its confines. These relate to a wide range of issues including some of those touched on here: identity versus identification; constraints on variation; intra- and inter-speaker variation; the relative influence of forms of regional and social variation and language attitudes. Linguistics for testimony or for legal practitioners frequently needs to be rather definitive, and wider sociolinguistic work could help those working in forensic settings to achieve, or make sense of, that requirement. Additionally, as Watt and Smith observe, living language varieties are '"moving targets" often with properties we do not expect either because they have not yet been described in sufficient detail or because

they have changed since last described' (2005: 101). Forensic linguists need help to continue to rise to this challenge.

It can be difficult to obtain naturally occurring forensic data in order to study the operation, distribution and significance of variation in legal settings. The Duke Project and other experimental work on language and law (e.g. Dumas 1990; Foulkes and Barron 2000) illustrate useful, exciting alternatives. Critiques of the Duke Project (Thompson 2002) provide good starting points to think about how a study like this could be replicated in view of more recent linguistic theory. Experimental studies can allow those without access to naturally occurring legal data to get their hands dirty with forensic data by devising projects based on an area of legal procedure which seems questionable or an aspect of expert testimony which might be interestingly replicated.

9.5 Where next?

The presence of a chapter on forensic linguistics in a book on variation indicates that 'language and law' and 'language and evidence' are coming of age. Indeed, many other aspects of forensic linguistics could have been included here. For example, accent can be a factor in trademark cases when the pronunciation of particular product or company names can influence differentiation from competitors (Shuy 2002: 76, 118–21; Gibbons 2003: 286–7). The potential for productive dialogue with colleagues working on other applications of language study, or even in less applied areas, is evidenced by the many collections of papers which, like this one, feature a chapter on forensic linguistics from the perspective of the home volume, such as discourse analysis (Shuy 2001, 2006a) or sociolinguistics (Finegan 1997; Gibbons 2006).

This chapter has given a necessarily brief indication of the diversity of literature which is increasing our knowledge of relationships between language, evidence and law. Over recent years, the rate and diversity of publication here has increased dramatically. Most recently, and in response to this proliferation, a number of book-length introductions to the field have become available. These offer good starting points for newcomers to the field. Additionally, each takes a different focus or analytic perspective so they are complementary. An early, wide-ranging example is Gibbons (2003) which takes an expansive view of legal systems and linguistic analysis although with a focus on Hallidayan theory. Solan and Tiersma (2005) examines linguistic issues in evidence-gathering and courtroom procedures before considering language crimes like bribery, solicitation and conspiracy. Coulthard and Johnson's (2007) two sections cover 'the language of the legal process' and 'language as evidence' from a predominantly discourse-analytic perspective. As forensic linguistic studies continue to diversify, tightly themed introductory texts appear. Current examples include Olsson (2008) which is relatively dominated by authorship and plagiarism and

Tiersma (1999) which concentrates on legal language, its history and social significance. Most prolific in the endeavour of producing introductory texts with specific foci is Shuy, whose writing, typically based on examples from his own casework, illustrates the activities of the forensic linguistic expert witness in cases of language crimes (1993), trademark disputes (2002), in examining police interrogations (1998) and investigations (2005), for example. Recently, Shuy has produced a 'nuts-and-bolts guidebook' for the aspiring forensic linguist (2006b: v) stressing something that is worth reiterating here, that to be a forensic linguist, one first needs to be a linguist, next to be an expert in linguistics and only then to think about applications of linguistics in legal settings (2006b: viii; 3).

Another way to get a big picture of forensic linguistics is through collections of papers such as those edited by Levi and Walker (1990), Gibbons (1994), Kniffka (1996), and Cotterill (2004). As their dates of publication indicate, in combination they offer insights into the trajectory of this area of scholarship. Again specialisation is beginning to become the norm in, for example, collections on law enforcement (Giles 2002) and sexual crime (Cotterill 2007). The perspectives of a variety of authors, on diverse legal or evidentiary topics, are offered by the *International Journal of Speech, Language and the Law* (formerly *Forensic Linguistics*).

Websites

Site	URL (note all sites last accessed 5 August 2010)
International Association of Forensic Linguists	www.iafl.org
International Association for Forensic Phonetics and Acoustics	www.iafpa.net
International Language and Law Association	www.illa.org
National Register of Public Service Interpreters	www.nrpsi.co.uk/about/index.htm
National Association of Judiciary Interpreters and Translators (USA)	www.najit.org
Deaf Lawyers UK	www.deaflawyers.org.uk
International Journal for Speech, Language and the Law	www.equinoxjournals.com/ojs/index.php/IJSLL
Centre for Crime and Justice Studies, London	www.crimeandjustice.org.uk/index.html
The Home Office (UK)	www.homeoffice.gov.uk
The Court Service (UK)	www.courtservice.gov.uk
The Prison Service (UK)	www.hmprisonservice.gov.uk
The Proceedings of the Old Bailey, 1674–1913	www.oldbaileyonline.org

Live webcasts of court proceedings (USA)	www.ncsl.org/programs/lis/webcasts.htm
Peter Tiersma's website	www.languageandlaw.org
Sue Blackwell's website	http://web.bham.ac.uk/forensic/index.html
Helen Fraser's website	www-personal.une.edu.au/~hfraser

10 Variation and identity

Emma Moore

10.1 Introduction

Identity has concerned variationists since the inception of the quantitative paradigm. Labov's discovery that use of centralised diphthongs related to a positive orientation towards the island of Martha's Vineyard demonstrated that 'one cannot understand the development of a language change apart from the social life of the community in which it occurs' (Labov 1963: 275). Understanding the connection between centralised (ay) (in PRICE words) and fishermen, upislanders and island-loyal youngsters required an examination of the local sites, practices and conflicts which constitute social meaning. That is to say, explaining this linguistic pattern was dependent upon an understanding of identity and identity practices and their relationship to the local social order.

Since 1963, identity has continued to be a central concern in variationist research; however, there is ongoing contention in the field about the way in which identity is defined and the function of identity in the explanation of language variation and change. This chapter will chart the role of identity in variationist work and consider the relevance of the concept to contemporary sociolinguistic analyses. I propose that a major part of what we might call variationists' current identity crisis has been caused by the tendency to define identity differently according to where one's work sits in the variationist paradigm. In recent times, as variationists unpack what they mean by identity, there has been increasing interest in the social meaning of variation, and the social and linguistic 'levels' at which meaning and identity are situated. Much of this exciting new work, which draws upon insights from sociology, anthropology and psychology, will be discussed in what follows. This work raises some provocative questions which challenge the view of the sociolinguistic variable provided by traditional language variation and change research. However, whilst it may be possible to interpret these differences of

Several people have helped me to think about the issues I discuss in this paper. In particular, I would like to thank Joan Beal, Mary Bucholtz, Penny Eckert, Paul Foulkes, Kira Hall, Miyako Inoue, Norma Mendoza-Denton and Rob Podesva for sharing their intellect with me. Any errors or shortcomings this work may contain are, of course, my own.

interest (the social meaning of variation versus the progress of language variation and change) as incompatible, following Bucholtz and Hall (2005) and Coupland (2007a), I argue that these approaches simply research different points along a spectrum of meaning. By evoking the notion of the indexical order (Silverstein 1976, 2003; Ochs 1991) and the indexical field (Eckert 2008), I demonstrate that the future of identity in variationist research relies upon our ability to situate our analyses in relevant ideological space and to trace the connections between levels of meaning within that space.

10.2 What is the role of identity in variationist research?

Eckert (2005) has described developments in quantitative sociolinguistics according to three distinct waves, each of which can be seen to conceptualise identity differently. In the first and second waves, sociolinguists have been motivated by a desire to document the spread of language variation and change – that is, to gain a picture of how language features are distributed through communities. In the first wave, the focus is very much upon correlating broad demographic categories such as gender, class, ethnicity and age with language use in geographically delimited speech communities (for instance, Labov 1966 in New York; Wolfram 1969 in Detroit; Trudgill 1974 in Norwich). As Mendoza-Denton (2002: 480) has noted, without these groundbreaking studies, we would not know the status of variables relative to change in progress, nor would we be aware of the social issues pertinent to community language use. For instance, Labov's observation of patterned class and style stratification in New York City enabled us to see the structure of speech communities as norm-based (in the sense that the social stratification of linguistic variables reflects a shared evaluative framework of sociolinguistic meaning within which community members embody social positions relative to one another; cf. Labov 1972b: 120–1).

The second wave of sociolinguistic research attempts to redefine the concept of 'social group' to account for more localised taxonomies. Using observation, researchers look for correlations between language features and participant-defined groups or networks (Labov's 1963 Martha's Vineyard study was perhaps the first to do this, but it was followed by others: for instance, Milroy's 1980 study of neighbourhoods in Belfast; and Cheshire's 1982a study of peer groups in Reading). The aim of the second wave, then, is to give local meaning to the more abstract demographic categories typical of the first wave. These studies are important for what they tell us about local dynamics and for the nuance they add to our account of linguistic patterns. For instance, Cheshire's (1982b) ethnographic study of children in adventure playgrounds suggested that those engaged in the vernacular culture had different language norms to other members of the community (she found that some of the boys in her study

used *more* vernacular variants in the formal context of the school than they did in the informal setting of the adventure playground). This helped us to see that there may be more than one set of norms governing behaviour in a speech community. Woolard (1985) accounts for this difference by positing two competing language markets: the standard and the vernacular.

Despite differences in scale, the first and second waves both imply something similar about the relationship between language and society. Because their analyses are based upon correlations, they suggest that identities are static repositories of the social meaning of language. In this sense, language features are presented as marking social group membership and, as a consequence, 'belonging' to the social groups who use them the most – so postvocalic (r) means 'upper-class New Yorker' because the middle classes of the city use it more; or centralised (ay) means 'Vineyarder' because, statistically, those who are Vineyarders by traditional and historical measures use it more. It may not be that researchers state these meanings in any concrete way (although Labov, at least, explicitly assigns centralised (ay) the meaning 'Vineyarder' in his 1963 study), but the implication of the research is that such correlations equate with identifiable category- or group-level social meaning.

However, the problem with statistical correlations is that they are generalisations. Whilst a variable may be used more by one social group than another, the correlation is never categorical. For instance, it is widely accepted that men will tend to use more non-standard variants than women – so much so that it has been called a 'sociolinguistic universal' (Holmes 1998). However, if we consider a variable like non-standard *were*, studies have found it to correlate with young men in the Outer Banks (Schilling-Estes and Wolfram 1994), young women in York (Tagliamonte 1998), the Lumbee community in North Carolina (Wolfram and Sellers 1999) and adolescent girls in Bolton (Moore 2004). Such diversity indicates that the meaning of non-standard *were* is not directly related to any of these identity types, 'but to something that is related to all of them' (Eckert 2008: 455).

It is for this reason that sociolinguists in the third wave attempt to answer the question of *how* it is that a variable might come to mean 'upper-class New Yorker' or 'rebellious adolescent girl' (and, indeed, how it might come to mean other things besides). This entails analysing meaning at a level which is different from the social groups or categories considered in first- or second-wave research. To do this, sociolinguists have had to look beyond their own discipline. Whilst traditional variationist research has been content to construe identity as the habitual embodiment of a social address, other fields in the humanities have long debated the psychological and material reality of identities. In social constructionist discourse (see, for instance, Bourdieu 1977; de Certeau 1984), identity is not something apart from language, something to be correlated with language; rather language and identity are co-constitutive.

That is to say, language does not just reflect social identities, it also helps to construct them. In this sense, language is a social practice and, by analysing the manner of someone's language use alongside other social practices, we can learn something about who an individual is. The key to understanding what a language feature means, then, is the ability to see how it functions in the production of a complete identity. If a group of underage adolescents engages in drinking practices, it does not mean that 'drinking alcohol' means 'rebellious adolescent' – at least not in a direct way (lots of people of all ages and backgrounds drink alcohol). What allows us to interpret the social meaning of drinking alcohol are the other social practices with which it co-occurs. Likewise, we interpret the meaning of non-standard *were* on the basis of the whole persona of the individual using it. As I show in Moore (2004), it only means 'rebellious adolescent' if it combines with other practices (including drinking) to collectively construct the rebellious adolescent persona of the Townies in my school sample; but it can just as well take on alternative meanings (cool, independent adolescent) when it contributes to the construction of the Popular girl persona in the same school.

By this account, then, the meaning of linguistic variables is not deterministic. Furthermore, given that a change in one aspect of a person's practice can cause the nature of their persona to shift, identities must also be fluid and dynamic. Meanings will shift and adapt dependent upon the wider style in which social and linguistic resources are used, and we will interpret identities based upon our understanding of the whole style. To understand identity, then, we need to understand 'style'; this has been a major goal of the third wave of variationist sociolinguistics.

10.3 How do linguistic variables mean?

'Style' has been used to refer to a range of concepts in variationist research (see Eckert and Rickford 2001; Schilling-Estes 2002; Eckert 2003; Moore 2004 and Coupland 2007b for summaries). The definition I will employ here is that of the Half Moon Bay Style Collective, who define style as 'a socially meaningful clustering of features within and across linguistic levels and modalities' (Campbell-Kibler *et al.* 2006).[1] This definition of style brings the variationist interpretation more in line with that used in other social disciplines (Irvine 2001: 21), where style is considered to be a form of *bricolage* (Hebdige 1979: 102–4). The term bricolage is intended to capture the clustering of resources (practices, ways of being) that occurs in the production of a distinct style. De Certeau (1984: 30) notes that, 'just as in literature one differentiates "styles" or ways of writing, one can distinguish "ways of operating" – ways of walking, reading, producing, speaking, etc'. Styles, then, are 'ways of operating' in the world which embody a range of practices.

Styles are never randomly contrived artefacts. They are determined by the resources available to individuals through their practice. Bourdieu (1977) suggests that individuals' ways of operating are embedded in their *habitus*. The habitus is a system of dispositions (Bourdieu 1977: 72) – a frame of reference which is durable (it accumulates over time via socialisation), transposable (it operates within and beyond the context in which it is acquired) and temporal (it reflects our history and our trajectory). Bourdieu notes that habitus is conditioned by pre-existing structures which are learnt and internalised; consequently, ways of operating, or styles, are regulated by the social and ideological constraints experienced by individuals. In this sense, individuals create styles, but they only do so within the structural constraints of their habitus.

De Certeau (1984) also acknowledges that styles are regulated by social conditioning (which is sustained by institutional power); however, he better accounts for individual agency than Bourdieu by noting that individuals can construct stances that oppose and challenge social order. In this way, local levels of social action can manipulate structural constraints to revise what is stylistically possible. As Rose and Sharma (2002: 4) note, de Certeau treats speakers as being aware of their relative power in a given context, having agency in such contexts and manipulating symbolic means to achieve specific ends. In de Certeau's account, then, styles are constructed in the context of existing social structure, but they may also constitute a challenge towards this structure. In this way, styles comprise an interface between practice (what people do) and structure (where they are positioned in the social order and any constraints this may entail).

Because style is a process informed by the social discourses in which speakers collectively engage, style work requires ethnography: the study of people in their day-to-day existence.[2] Adopting techniques from anthropology, much contemporary variationist work adopts an ethnographic approach (see, for example, Bucholtz's 1999 study of nerd girls in California, Eckert's 2000 study of jocks and burnouts in Detroit, Moore's 2003 study of high-school students in Greater Manchester, Mendoza-Denton's 2008 study of gang girls in California, Zhang's 2005, 2008 study of Chinese professionals in Beijing, and Rose's 2006 study of senior citizens in Wisconsin). Not only has this research provided more nuanced information on the distribution of linguistic variables, it has also done much to contest some of the stereotypes propagated with respect to the connection between linguistic features and 'social address identities'. For instance, Eckert's (2000) famous study of Belten High demonstrates that, whilst social class correlates with sound changes implicated in the Northern Cities chain shift, the distribution of language features is best explained by membership in local communities of practice.

The community of practice (CofP) is a useful concept for the ethnographer as it provides a way to group individuals on the basis of engagement in shared

social practice. First introduced to sociolinguists by Eckert and McConnell-Ginet (1992), but developed by the social learning theorists, Jean Lave and Etienne Wenger (1991), it describes 'an aggregate of people who come together around mutual engagement in an endeavour' (Eckert and McConnell-Ginet 1992: 464). CofPs differ from other types of social aggregate because they situate individuals according to joint engagement in social practice – not just according to membership of decontextualised social structure, or social address. To be a CofP, a social group must be mutually engaged in some joint enterprise, producing a shared repertoire of practices which identify them as a collective (see Wenger 1998: 73–83 for a more comprehensive discussion of these criteria).[3]

Eckert's ethnographic fieldwork, which took place over two years in a high school in Detroit, uncovered two oppositional CofPs: the Jocks and the Burnouts. The fact that these CofPs were oppositional is important as styles are, by definition, distinctive (Irvine 2001). That is to say, social and linguistic practice is as meaningful for what it is not as for what it is. As speakers live alongside other speakers and as disparate practices and forms of participation construct difference between groups, language is implicated in the development of differentiation. As an embodiment of the shared repertoire of a CofP, styles make the identity of a CofP recognisable and distinct from those constructed by other CofPs.

Eckert's work at Belten High demonstrates the importance of differentiation. One of the variables she considered was the raising of (ay). When Eckert looked at the distribution, she found that girls were leading this change, but not all girls. In fact, whilst the Burnout girls led the change, the Jock girls lagged behind everyone else, resulting in the following non-standard–standard continuum: Burnout girls > Burnout boys > Jock boys > Jock girls. Put another way, the extreme behaviour of the female members of the two communities of practice defined the extent of the variation. This was an unusual finding (but one corroborated by Labov 1990), given the commonly reported finding that females lag behind in the use of non-standard language features (see earlier discussion of this 'sociolinguistic universal'). Eckert's (1989, 1998, 2000) work revealed that some females do – but whether they do or not depends upon: (i) how gender is constructed locally, and (ii) how advanced the sound change is (the more stable a change, the more likely we are to see marked differentiation). Her observation of females' situation at the extreme of variation also revealed something else: that females may make more use of symbolic resources in their identity construction than males do (the Jock and Burnout girls use language more extensively to distinguish themselves from one another than do the Jock and Burnout boys). Eckert's analysis, then, not only provides a more subtle analysis of language distribution, it also provides important information on differences in the resources females and males use to construct identities.

Other CofP studies have also contested stereotypes – most recently, Norma Mendoza-Denton's (2008) work on Latina Youth Gangs. Mendoza-Denton's work focuses upon the symbolic construction of distinction between the Norteñas and the Sureñas, two Chicano/Mexican gangs in California, and she considers the raising of (ɪ) in words like 'bit' (2008: ch. 8). In Mendoza-Denton's study, the most significant social factor affecting whether or not a Chicano speaker raises her vowel is her CofP membership. Norteña and Sureña girls both raise (ɪ) more than other kinds of girls (such as Latino Jocks, Disco kids, and wannabe Norteñas and Sureñas). Given that these two CofPs use their other forms of social practice to differentiate themselves, how might this affect the meaning of raised (ɪ)?

Mendoza-Denton argues that both the Norteñas and the Sureñas use raised (ɪ) because it is symbolic of a broader Latina-based identity which has relevance to both gangs; however, each gang reconstitutes the meaning of this variable in the context of their distinct styles. For the Norteñas, who predominantly speak English, raised (ɪ) signifies a Latina identity that is bicultural and modern; whereas for the Sureñas, who predominantly speak Spanish, it signifies a Latina identity that is loyal to Mexico and resistant of American assimilation.

Whilst Eckert's work shows that different kinds of females might use different pronunciations to constitute distinct identities, Mendoza-Denton shows that, even when females use the same language feature, they may use it to different effect. These studies have provided the foundation for a more sophisticated study of social meaning, which is now beginning to consider the social meaning of less conventional linguistic variables. Whereas traditional language variation and change research takes a substitution class approach to variation (where two or more variants compete for the same variable 'slot'), a stylistic approach to variation considers the nature of a variant's occurrence to be as important as its frequency. This raises questions with respect to what we mean by a sociolinguistic variable and, ultimately, what we model as (socio)linguistic knowledge.

10.4 What types of variable can mean?

Researchers often focus their search for meaning on the kind of phonological or morphosyntactic variables typical of traditional language variation and change research – so Eckert considers (ay) raising, Mendoza-Denton considers (ɪ) raising and Moore (2004) considers non-standard *were*.[4] However, summaries of recent sociophonetic work (Foulkes and Docherty 2006; Hay and Drager 2007) demonstrate that researchers are starting to look beyond these traditional variables. In additional to considering vowels and vowel quality, researchers are beginning to study trajectory, duration and voice quality more systematically, in addition to considering the social meaning of consonants and prosody

(Hay and Drager 2007: 92–3). This work reveals the potential for a much wider range of 'variables' to operate as symbols of social meaning.

For instance, in his study of a mid-twenties, white, middle-class gay man, Heath, Podesva (2006, 2007, 2008) shows socially significant variation in segmental phonology (word-final coronal stops), vowel quality (vowels preceding released coronal stops), intonation (declarative contours) and voice quality (falsetto and creaky voice). By carefully analysing the discourse functions of each 'variable', Podesva (2008) is able to define a core function for each which goes across contexts of use which he defines as 'a kernel of similarity'. For instance, he identifies a core 'expressive' meaning for falsetto and a core 'non-threatening' meaning for declaratives with rising intonation. Furthermore, he demonstrates that these meanings can be operationalised in the context of different styles. Heath's identity as 'the caring doctor' in clinic is constructed via a lack of expression (weak falsetto) and a non-threatening stance (frequent rising intonation in declaratives); whereas Heath's identity as the 'diva' at a barbeque is constructed by excessive expressiveness (frequent, extreme falsetto) and animated stance (extreme f0 values in declaratives).[5] Podesva defines these personae using his knowledge of the ethnographic context and evaluating falsetto and intonation alongside an analysis of Heath's segmental phonology and vowel quality to produce prototypical 'style clusters' where these forms are collectively employed.

In addition to demonstrating the range of linguistic resources capable of carrying social meaning and the extent of intra-speaker style shifting, Podesva's work also alludes to another source of social meaning: discourse context. The relevance of this context is supported by evidence that speakers pay more attention to 'pragmatically salient' (Errington 1985) linguistic constructions, such as discourse markers or intensifiers, which encode subjective evaluations (Traugott 2001; Woolard 2008). Such salience may well result in the more explicit monitoring of phonological and syntactic style. Furthermore, Schilling-Estes (2004) has demonstrated that production of phonological or morphosyntactic features can also be affected by topic and alignment. Her analysis demonstrates that ethnically associated markers are more salient when local issues of ethnicity are foregrounded in a conversation between ethnically diverse interlocutors. However, the same speakers use fewer ethnic vernacular forms when relaxed (because they work instead to promote solidarity and limit social difference) – suggesting a complex interplay between linguistic form and positioning in discourse. Similarly, Coupland (2007a), revisiting his (1988) travel agents study, demonstrates that the discursive frame of an interaction interacts with the variables used to construct a range of social meanings. For instance, in a frame where status is called into question (speaking with a non-familiar colleague), Coupland finds that the agent studied, Sue, decreases her use of variants typical of the Cardiff vernacular and increases

her use of technical discourse – behaviour which constitutes a professional identity. However, in a frame where interpersonal relationships are at question, Coupland argues that Sue's increased vernacular symbolises neither 'unprofessional' nor 'lower class' but reflects and constitutes the low personal control she experiences in a conversation about dieting. Here, then, the social meanings and, ultimately, the nature of Sue's identity, depend as much upon the resources used as the social work they are able to do within a given frame.

In my latest work with Rob Podesva (Moore and Podesva 2009), we bring together social and linguistic contexts by examining the *co-occurrence* of stylistic constraints identified in previous research. We examine a variable, the tag question (which in addition to having phonetic, morphosyntactic and discoursal properties, has also long been recognised as socially meaningful – see, for instance, Lakoff 1975; Holmes 1982, 1984, 1995 and Cameron, McAlinden and O'Leary 1989) in the context of the four CofPs observed in my study of Midlan High (Moore 2003). These CofPs comprise the rebellious, anti-school Townies; the cool, independent Populars; the pro-school, knowledgeable Geeks; and the elitist and trendy pro-school Eden Village clique. An analysis of the discourse context of tag usage (placement in turn and agreement patterns) revealed that tag questions seem to have a core meaning, irrespective of which social group uses them. Given that the syntax and semantics of tags encourage an interlocutor to agree with a proposition, we follow Hudson (1975) in identifying this meaning as 'conducive'. However, this underlying function tended to be stylised differently by each social group. For instance, the phonetic design of the tag (phonetic realisation of /t/ and /h/-dropping), the grammatical design (the presence or absence of non-standard grammatical items, such as non-standard *were*), and the content of the tag (who/what was discussed; that is, the wider 'frame' of the tag's occurrence) could all be manipulated in subtly different ways. Consequently, the Townies' largely non-standard, rebellious style meant their tags contributed to the construction of an experienced, authoritative, post-school identity; the Populars' excessively conducive, gossipy, moderately non-standard style meant their tags contributed to the construction of a somewhat bitchy, evaluative identity; the Geeks' impersonal, intellectual style meant their tags contributed to a knowledgeable and authoritative identity; and the Eden Village girls' interactive, facilitative tags contributed to the construction of a collaborative and evaluative identity. Whilst we do not claim that every tag by every group member was constructed in such a way, we use discourse analysis of concrete examples to provide illustrations of the CofPs' prototypical tag questions, which we take to represent iconic performances of group style.

All of the work discussed in this section emphasises the range of resources relevant to meaning-making processes. Social meaning is not just multimodal in as much as it goes across different forms of social practice (ways of dressing, ways of engaging in activities, ways of talking), it is also linguistically

complex (such that syntax, phonology and discourse may work synergistically rather than independently of one another). This latter point has been recognised in conversation analysis (Local and Walker 2005: 122) and is now being acknowledged in variationist accounts of meaning with the recognition that what can vary surpasses the traditional notion of the (socio)linguistic variable (the status of which was debated as early as Lavandera 1978). Until now, the tendency to prioritise the description of a variant's trajectory through a geographical area has marginalised the study of social meaning. Researchers are calling for variationists to reinstate concern with meaning into the variationist paradigm (Silverstein 2003; Eckert 2008; Woolard 2008) – not just for what it would explain about the motivations for language change (see Labov 2001: 325 for an acknowledgement of this necessity), but also because of what it can tell us about human language ability. Engagement with research on the relationship between language, identity and social meaning is helping to refine recent work on the nature of language acquisition (Foulkes and Docherty 2006; Hay and Drager 2007). For instance, exemplar-based models, which propose that speakers store a distribution of socially loaded exemplars (concrete examples of how language has been experienced) as opposed to abstract underlying linguistic forms, have drawn upon findings from the kind of style research examined here. The ongoing development and critique of such cognitive models relies upon our ability to understand social meaning, its manifestation in interaction and its relationship to biological and linguistic constraints.

10.5 Where is meaning situated?

The shift from variable-driven to pattern-driven analyses (Hay and Drager 2007: 90) raises the question of whether third-wave researchers are talking about 'identity' at all. After all, work is focused on examining the processes of social meaning as opposed to correlating identity categories with individual variables. Whether or not one considers third-wave studies to be about identity depends upon how one defines identity. It is clear that different academics use the term to refer to different concepts. For instance, Cameron and Kulick (2003: 104, 2005: 123), referring to Butler's 1990 notion of the epistemological subject, use 'identity' only in the sense of the habitual embodiment or 'claiming' of a social position. 'Identity' by this definition is little more than the 'social address' typical of traditional societal models. It is something fixed, which can be assigned to an individual voluntarily or by someone else (a sociolinguist, perhaps). Given this definition, it is no surprise that Cameron and Kulick have been critical of the hegemonic presence of identity in sociolinguistic research, claiming that such a focus obscures the wide variety of social purposes fulfilled by language. Eckert's (2008) avoidance of the term 'identity' in favour of 'persona' suggests that she too avoids the term because

of its traditional associations. However, other scholars use 'identity' in a much broader sense and, as a consequence, are comfortable with the central position of 'identity' in sociolinguistic research. For these researchers, identity is not just a categorical status but a process which goes across different social levels. For example, Coupland (2007a: 27) refers to macro-, meso- and micro-social identity frames – corroborating Bucholtz and Hall's (2005: 592) argument that '[i]dentities encompass (a) macro-level demographic categories; (b) local, ethnographically specific cultural positions; and (c) temporary and interactionally specific stances and participant roles'. By this definition, all speakers are constantly engaged in identity work (socially positioning themselves or others as 'Geeks' or 'Townies', for instance, or assuming a 'caring' or 'authoritative' stance), even if they are not explicitly engaged in projecting and claiming a particular macro-category 'identity' (such as 'working class' or 'female'). Although it is not entirely a matter of terminology (Cameron and Kulick 2003, 2005 also call for sociolinguists to incorporate different methods of analysis, such as psychoanalysis, into their study of the relationship between language and the social world), the 'levels' of analysis proposed by Bucholtz and Hall (2005) and Coupland (2007a) are not so very different from the distinction Cameron and Kulick (2003: 138–9) make between identity (the conscious embodiment of a social position) and identifications (the conscious or unconscious processes through which individuals assimilate and transform social effects). In both cases, the discussion points to different 'levels' of sociolinguistic work.

Theories of indexicality (Silverstein 1976, 2003; Ochs 1991) have done much to elucidate our understanding of the connections between the identity levels analysed in different types of variationist work. An index simply refers to a meaningful link between a linguistic form and a social meaning. A semiotic link will always be ideologically mediated (Silverstein 2003), in the sense that we use our belief systems to explain and instantiate any connection between language and the social world. This ideological mediation can lead to a series of complex connections, or indexical layers, as we link micro-social meanings (such as 'being tough') to larger socially meaningful units (such as 'being working class'). As Bucholtz and Hall (2005) suggest, basic social meanings or direct indexes (Ochs 1991) (such as 'being tough') are articulated when speakers express stances or orientations in the course of their interactions. These stances, taken at the micro-social level, may be connected to the meso-social level by virtue of their repetition (Du Bois 2002 and Rauniomaa 2003, cited in Bucholtz and Hall 2005, refer to this as a process of *stance accretion*). That is to say, if members of a particular group repeatedly take the same stance in their interactions, we may come to associate that stance with that social group and, in turn, with any behaviours or practices associated with that group. Given that particular social groups come to be associated with

certain category memberships, stances may then track recursively outwards to the macro-social level too. To give an example, imagine a group of high-school kids whose point of commonality is their engagement in a cool, independent style. Imagine that when they discuss their activities, they tend to talk in detail about their friends and associates, using gossip and evaluation to conduce a shared viewpoint about behaviours and group boundaries. To enable this, their use of pragmatically salient variables – such as tag questions – increases. Because their obsession with group boundaries is so prominent and their style so conducive, they come to be viewed as bitchy and divisive, a quality which then becomes ideologically linked to how all of their practices (including speech) are viewed. Now imagine that members of this group also happen to be female. Given that there is an apparent correlation between gender and engagement in this 'bitchy' social group, there is the potential for 'being bitchy' to be tied to the identity 'female' – helped, of course, by dominant ideologies about female style. In this process, tag questions may be semiotically linked to being bitchy, being a member of this local group *and* being female – and one or more of these meanings may be operationalised dependent upon the frame used to interpret the observed behaviour.

In fact, the scenario just described summarises what Rob Podesva and I found in relation to tag-question use in the study I reported in §10.4 (Moore and Podesva 2009), although the range of meanings we found extended well beyond those discussed above. Following Eckert (2008), we were able to construct an *indexical field* for tag questions at Midlan High. This is shown in Figure 10.1.

An indexical field is 'a constellation of ideologically related meanings, any one of which can be activated in the situated use of [a] variable' (Eckert 2008: 464). At the centre of the indexical field lies the *n*-th order (Silverstein 1976), or direct (Ochs 1991), index of the linguistic feature. This meaning may be central if it represents a core association between the feature and its meaning. In our analysis of tag questions, the *n*-th order indexical meaning is 'conducive' (remember, this was the meaning of tag questions which went across our data-set).

Put in Silverstein's (2003) terms, the indexical field represents the ideological associations between the *n*-th order indexical value ('conducive') and the *n*+1st order indexical values ('cool', 'knowledgeable', etc.), the *n*+2nd order indexical values ('popular', 'working class', etc.), and so on. This association of these orders is schematised in Figure 10.2. The creation of any additional indexical order occurs as a consequence of an ideological interpretation of a perceived pattern. Such interpretation simultaneously reconstrues the *n*-th order indexical value and gives shape to (and potentially redefines) the linguistic feature's indexical field (hence the double-headed arrows in Figure 10.2). In this way, the meanings in the indexical field are always available for reconstrual.

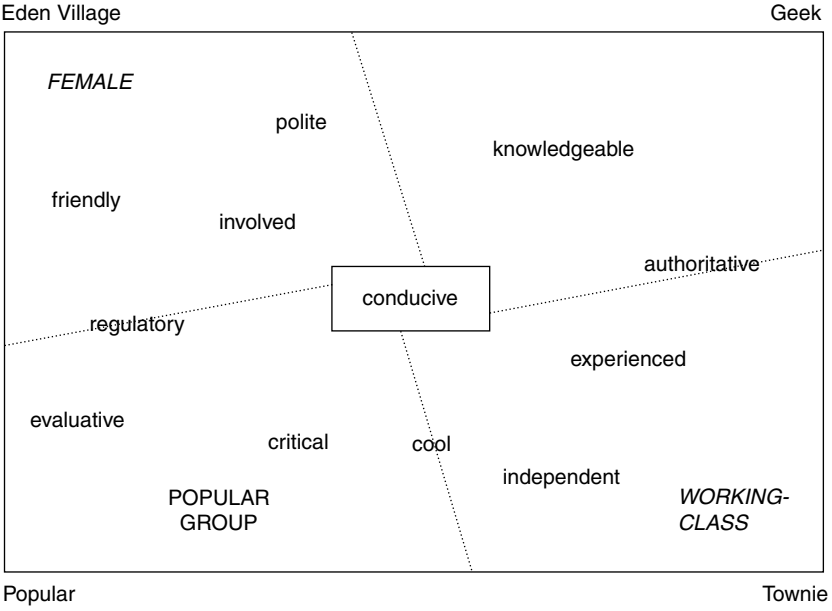


Figure 10.1 Indexical field for tag questions at Midlan High (from Moore and Podesva 2009)

The meanings given in this indexical field represent the findings of our context-specific, bottom-up analysis. We show *n*+1st order indexes in plain text. Thus, there is a range of social meaning at the interactional or micro-social level. For instance, the Populars' use of tag questions to conduce evaluative and critical stances toward girls in the school allows them to portray themselves as 'cool'. In contrast, Townies' use of tag questions to conduce a shared viewpoint around their independence, experience and authoritativeness, indexes their own brand of 'coolness'. That the 'cool' *n*+1st order index is recruited by both Populars and Townies demonstrates that meanings in the indexical field can be repackaged and combined in unique ways to create quite distinct local identities. Similarly, Geeks and Eden Villagers recruit the 'conducive' function of tags to construct their own *n*+1st order indexical meanings – some of which (e.g. 'authoritative', 'regulatory') overlap with other identity-specific meanings.

In addition to the micro-social meanings we observed, we also found a frequency effect in our study, with the Popular CofP using almost twice as many tag questions as the other CofPs that we analysed. This high level of use, which was explicitly acknowledged in metalinguistic comments on tag-question usage, hints at an ideological crystallisation, to the point that the *n*+1st

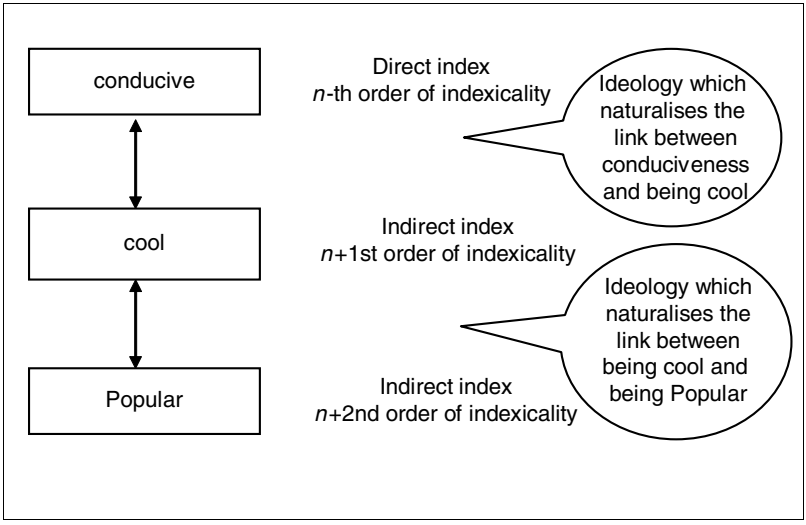


Figure 10.2 An example of the indexical layering of tag-question meaning

order indexes accrete into the Popular identity itself as an *n*+2nd indexical order (represented in capitals in Figure 10.1). This suggests that a particular tag design may become associated with an acknowledged CofP identity at the meso-social level. We did not find similar identity crystallisations for the other CofPs, which is why the CofP labels in Figure 10.1 are positioned outside the borders of the indexical field.

Whilst tag questions only carried meso-social level meaning for the Popular CofP, each of the CofPs we analysed were also constrained by and judged relative to macro-category classifications. Thus, we found some macro-level ideological interpretations at the *n*+2nd order level. For instance, the Townie girls' engagement in working-class culture enabled a 'working-class' categorisation of their tagging practice. Furthermore, the Eden Village girls' *n*+1st order indexes ('polite', 'friendly') sit comfortably alongside dominant discourses of femininity – enabling a 'feminine' interpretation of their tag use. Indexical values at the macro-social level are represented in capitalised italics in Figure 10.1.

Of course, certain practices can only be linked to identities if the interpreter engages with an ideology that facilitates such a link. The role of ideology in determining social meaning cannot be overstressed. In her study of Middlesbrough, Llamas (2007) demonstrates that the meaning of a linguistic variable is very much determined by its historical context and the nature of one's engagement with this context. Her study documents changes in the pronunciations of (p, t, k), demonstrating that older speakers use more of the

released variants associated with Yorkshire, whereas the younger speakers use more of the Tyneside-associated glottalised (p, k) and 'levelled' glottalled (t). However, Llamas argues that, in Middlesbrough, the Tyneside-associated variables do not index Tyneside but contribute to a distinct Middlesbrough identity and thus mean 'Middlesbrough'. Her careful historical analysis (considering changes in administrative boundaries and the subsequent shifting orientations of speakers) gives her access to the local ideologies which imbue the glottalised variants with this meaning. However, Llamas acknowledges that this meaning is not universal – older speakers' production suggests that they do not perceive the variables in the same way. Likewise, one may imagine that those from outside the north-east might well struggle to assign glottalised (p, k) any meaning other than 'Geordie' (given that, in the absence of local knowledge, one tends to assign meaning on the basis of iconic associations).

Similarly, in a fine-grained phonetic analysis of the raising and backing of (ay), Wagner (2007) shows how this 'male-associated' variable signals salient differences in the speech of girls in a south Philadelphia high school. The girls of Irish descent tend to produce backer variants than those of Italian descent and Wagner explains these differences as a consequence of indexical meanings ('tough', 'youthful', 'lack of care') which at once construct and reflect Irish girl style. That some of these meanings also connect to masculinity is a consequence of an ideological association which plays out differently in this female context. What matters here is the contrast between Irish and Italian girls which the variable functions to symbolise.

Both Llamas' (2007) and Wagner's (2007) studies reveal the importance of historical context to our understanding of the social meaning of linguistic variables. The ability for glottalised (p, k) to symbolise 'Middlesbrough' depends upon a history of shifting administrative borders and its effect upon local orientations. Likewise, the meaning of (ay) in Philadelphia depends upon a history of immigration and settlement patterns which impinge upon community relationships. Zhang's (2008) study of rhotacisation in Beijing has demonstrated the benefits of a historically situated analysis more explicitly. By studying a range of historical sources (novels, essays, literary and critical works), Zhang shows that, over time, rhotacisation has come to represent a characteristic (smoothness) and an iconic social character (the Beijing Smooth Operator). She demonstrates that employment of this variable in the locally based Chinese professionals' 'cosmopolitan style' entails engagement with this social history – male professionals use it freely, but female professionals are constrained by its historical and ideologically governed associations.

These studies demonstrate that social meanings rely upon something significant in a community's past. Speakers are unlikely to invest significance in social issues that are trivial to them. Nonetheless, what ends up as significant and symbolic may not be constant within a given community of speakers. Recent work

by Johnstone and Kiesling (2008) shows that in addition to variables having different meanings to different speakers, sometimes – for some people – they may not have any of the meanings we might expect. Using perceptual experiments, they explore the meanings assigned to /aw/-monophthongisation in Pittsburgh and find a complex interaction between production and perception. Whilst those who do not produce /aw/-monophthongisation tend to perceive it as 'local', those who do have it as a feature of their own speech do not. Like Podesva and Chun (2007), Johnstone and Kiesling (2008) stress the multiplicity and indeterminacy of social meaning – such that /aw/-monophthongisation can be interpreted using a range of indexical schemas, some of which imbue it with local social meanings and some of which do not.

Johnstone and Kiesling's (2008) experimental work demonstrates the need to employ a range of methods in our search for the social meaning of linguistic variation. One set of researchers uses fieldwork to obtain knowledge of local ideologies: Moore and Podesva (2009) rely upon a combination of ethnographic knowledge gleaned from participant observation and an analysis of metalinguistic commentary, and Llamas (2007) uses an identity questionnaire and interview to gain knowledge of local practices and orientations (see also Burbano-Elizondo 2006 and Asprey 2008). On the other hand, a number of lab-based studies, like Johnstone and Kiesling (2008), use experiments to access speaker perception. For instance, by manipulating the (ing) variable in spontaneous speech, Campbell-Kibler (2005, 2007) shows that micro-level social meanings (e.g. 'educated', 'urban') depend on the assumptions listeners make about the macro-social characteristics of speakers (e.g. 'southerner', 'gay'). Other work has demonstrated that our perception of social meaning is not only influenced by the social information we have about a speaker but even by our surroundings when we hear a speech sample (Hay, Nolan and Drager 2006; Drager and Hay 2006). This experimental work indicates that both perception and production studies have a role to play in our understanding of social meaning and identity levels.

10.6 Where next?

The work discussed in this chapter suggests that identity research is vital to our understanding of the connection between language and social meaning. If we believe that social meaning has a role to play in explaining language variation and change, it follows that identity research has a vital role to play in variationist study. The research described in this chapter demonstrates what most variationists now readily acknowledge: the study of social meaning is as important as the geographical tracking of sound change in progress; likewise, the study of social constraints is as important as the study of the internal constraints of the linguistic system. The more we learn about social meaning and identity,

the better able we are to explain language variation and change. Likewise, the more we learn about language variation and change, the better able we are to identify potential trajectories of social meaning as they span individuals and communities.

It has only been possible to dismiss the role of social meaning in the past because social processes have been simplified to the point of irrelevance. For instance, in Trudgill's (2001) work on the genesis of New Zealand English (see also Trudgill, Gordon, Lewis and Maclagan 2000) it is clearly demonstrated that New Zealand English is a consequence of majority dialect forms winning out over minority forms. Believing this to be a probabilistic outcome, explainable by sheer numbers alone, leads to the dismissal of the explanatory function of identity and the related issues of stigma and prestige (Trudgill 2001: 44). However, the goal of the New Zealand study was to document the genesis of a new English across geographic and diachronic space. Such a panoramic lens can only provide broad-scale characterisations of social relations – Scottish speakers versus English speakers versus Irish speakers, for instance (and this is, of course, its utility). However, as much of the discussion in this chapter has demonstrated, identity processes inhabit local space, and understanding them requires a more delicate sifting of social context. As Wales (2010) has shown, a little digging reveals a complex of historically grounded, ideologically loaded social forces and conditions which were likely at work in early colonial settings. So, whilst the New Zealand study reveals much about the linguistic genesis of New Zealand English, it only addresses the vaguest 'identity' types in its analysis.

Variationists have now reached a point where the systematic simplification of social processes no longer serves to advance the explanations we can offer in our field. Whether or not we choose to label our search for social meaning 'identity research' is not as relevant as the acknowledgement that there are processes of social meaning which are ideologically mediated and which, ultimately, have the potential to cause and explain language variation and change. The studies discussed in §10.3 suggest that social meaning resides in styles of language use. Consequently, the more we learn about the speakers we analyse, the better equipped we are to understand what their language means. Studies which combine ethnography and quantitative analysis are increasing in frequency. Many have been inspired by Eckert's (2000) work, which provides a powerful illustration of the potential of this type of research. Nonetheless, there is still much to learn about the way speakers operate within communities of practice. Mendoza-Denton's (2008) monograph provides a contemporary account of the extent to which a linguist can document the practices of communities.

Furthermore, given the importance of styles, how we document stylistic variation schematically requires consideration. How do we visually represent

the clustering of variants which accrete in socially meaningful ways? Podesva (2006, 2008) provides some ideas on this, but this is still very much an open question. Coupland's (2007b) monograph, which provides an excellent overview of linguistic style research, will inform any study of style.

We are also still at the early stages of exploring what can meaningfully vary. By examining the latest research on voice quality and discourse context, we have seen that speakers can exploit a wide range of linguistic features in their meaning-making endeavours. But whilst works such as Podesva (2008) and Schilling-Estes (2004) provide striking results, we only know a little about the sociolinguistic patterning of these features of language. Hay and Drager (2007) provide inspiration on a range of phonetic 'variables' and Coupland (2007a) demonstrates the merits of examining discourse contexts for what they contribute to the meaning of linguistic features. These works inform on the types of data one might feasibly consider for analysis.

We also still need to learn about exactly what can stylistically co-vary and what is linguistically constrained from doing so. Are some language features more salient than others? The only way to test this is to combine production and perception analysis. The experimental work examined here (e.g. Johnstone and Kiesling 2008 and Campbell-Kibler 2007) is both exciting and innovative and promises to advance our understanding of the production–perception interface. Those interested in experimental work should draw inspiration from these studies.

Finally, §10.5 suggested that different identity processes can connect to produce a range of social meanings which operate at different interactional levels. Understanding indexical links requires us to learn more about the range of ideologies which affect the speakers we analyse and to ascertain how these ideologies connect the identity levels considered in this chapter. Silverstein's writing on indexicality can be incredibly dense, but Bucholtz and Hall's (2005) discussion of these issues is extremely accessible and will appeal to those interested in the theoretical issues of identity research.

It is clear that exploration of the issues outlined here will require the employment of a range of methodologies – including the experimental and the ethnographic. It will also require us to engage in both perceptual- and production-based analyses and to be fearless enough to wander occasionally outside our own discipline. There is much potential for methodological and analytical innovation. What could be a better goad for research? As variationists grapple with these issues, we may well find ourselves at the vanguard of knowledge on human stylistic capabilities, processes of language acquisition and, of course, language change.

11 Variation and populations

Rob McMahon

11.1 Why populations?

If you have got further than the title, then you are reading a chapter in a book about analysing linguistic variation in English, written by a molecular geneticist, which will be asking whether differences in our genes are relevant to understanding differences in language. Of course, the chapters in this section show that linguistic variation is relevant in all kinds of domains – in legal and educational contexts, for example, and in building speakers' identities – and conversely that those various domains may help us understand more about variation in language, where it has come from, and what it means. Unlikely though it might seem to some, this chapter will explore the possibility that biological or genetic variation falls into the same category of apparently external factors which may cast light on some aspects of linguistic variation.

This idea is not uncontroversial. Some writers on historical linguistics, for instance, regard it as a self-evident truth that genes have nothing whatsoever to do with language; or rather, that while our human genetic make-up might conceivably contribute to our capacity to acquire and use language per se, it certainly has no impact on the specific language or variety we use. Thus, Hale (2007: 226), discussing the generally accepted genetic hypothesis, which proposes that repeated structural similarities in languages today indicate descent from a common ancestor language, suggests that:

> Although it should not be necessary to point this out, the genetic hypothesis is not a theory about gene flow within human population groups. In fact, there is no reason, given what we know of the history of human civilisation, to believe that there is any relationship between the physical transmission of genetic material from one generation to the next, and the transmission of a grammar from one generation to the next. The two are completely independent of one another, and only accidentally coincide in monolingual and monodialectal communities – which probably do not exist (and never have). I point this out only because one continues to see evidence from human genetic lineages cited as support for (or refutation of) theories of human linguistic lineages.

This, however, needs some unpacking. It certainly **is** self-evidently true that our genes do not determine what language we can or do acquire as individuals: the

stock thought-experiment here is that a baby born to Xhosa-speaking biological parents, but with English-speaking adoptive parents who bring the baby up in the UK, will grow up speaking English and not Xhosa. Moreover, the baby in question, if she becomes a linguistics undergraduate, won't necessarily find it any easier to learn to produce click sounds than any other English speaker. However, if we move from the level of the individual to the level of the population, or the speech community, it might well be the case that **on the average**, speakers of Xhosa have non-linguistic factors in common which speakers of English do not, and vice versa. Taking a substantial sample of the relevant populations into account, we might find that traditions of storytelling. architectural practice, preferred food and cooking styles, and aspects of art might fall into two essentially distinct, though marginally overlapping sets. Such constellations of cultural artefacts can and have been used to establish the relationships between populations and their reciprocal influences on each other in times past. Genetic variation patterns in much the same way as these cultural artefacts and has similarly been used as a marker for population contact and interaction (see for example the papers in Bellwood and Renfrew 2002); and if that is the case, there is no reason to assume a priori that the same would not be true of language. Indeed, how could there be a closer connection between two aspects of a culture than between its language and its speakers?

Hale points out that genetic and linguistic variation are independent; but that is exactly what can make studying any potential covariance so interesting and so informative. If one determined the other, then the fact that they patterned in the same way would tell us very little. But if they vary on average in much the same ways (setting aside outliers in populations like our genetically Xhosa but linguistically English linguistics undergraduate above), then that potentially tells us a great deal about some third factor which has had an effect on both. If people typically form groups, and interact more within their group than outside it; and if people both speak languages and carry genes; then it is at least possible (and worth testing) that the histories of those populations, their movements and interactions, are reflected in present-day distributions of variation in both genes and language. And since we find speech communities at the levels of both language (Xhosa versus English) and dialect (Tyneside versus Liverpool, or Somerset versus Fife), we might accordingly expect genetic variation and language variation to follow similar if not identical courses at relatively local as well as national levels.

There are three possible responses to this set of suggestions. First, you might automatically side with Hale: genetic variation has nothing to do with language variation, and any apparent shared similarities in pattern are sheer chance, and are of no consequence for either field. In §11.2.1 below, I will explore some of the reasons why shared similarities might be expected and therefore taken a little

more seriously, however. Second, you might be willing to accept that cases like the English–Xhosa one might have some validity, so that genetic and linguistic variation might pattern similarly if we compare populations divided by vast geographical distance, and languages which have never been thought to form one family even by the most enthusiastic proponents of megafamilies. However, you might well argue that these commonalities of patterning could only be discerned in the most extreme cases of geographical and linguistic separation. Surely, when we turn to related languages, and to populations which (as Hale suggests) have become progressively more intermixed in our current global village, these patterns will also have become irretrievably obscured and confused? And if that is the case for different languages, how much more difficult would it be to see meaningful correlations in the case of dialects of the same language? Add to this the fact that genes are known to evolve and hence to differentiate very slowly, compared to the variation in features of a language which can develop within a single generation, as a word is coined and becomes current, and we surely have a recipe for languages and genes getting seriously out of step. Finally, however, you might be willing to suspend disbelief for the moment and follow the line of reasoning I advocate in §11.2.3 and §11.2.4 below. Here, I will suggest that we don't need to make unrealistic assertions about 'monolingual and monodialectal communities – which probably do not exist (and never have)' (Hale 2007: 226) in order to establish potentially intriguing correlations with genetic variation. We don't have to pretend that genetic intermixture has not happened either. On the contrary, we find that speech communities, especially outside modern cities, often include traditional dialect speakers whose ancestors have lived there for generations – and indeed, this has often been a criterion for selecting speakers to participate in particular linguistic surveys. Even much more intermixed populations will be characterised by their own distinctive frequencies of both genetic and linguistic variants, which modern computational techniques can correlate and contrast with the frequencies found in other populations.

If common patterns of genetic and linguistic variation might reflect common historical events, then we should be able to exploit this to answer questions of interest to both population historians and linguists, whether those linguists are focusing on historical linguistics or dialectology. In §11.3, we will look at the question of the origin of dialect variation in Old English. We will not come to any firm conclusions; the genetic data currently sampled from linguistically relevant groups are woefully inadequate to answer such questions. However, it is possible to see from some recent studies how genetics can assist in understanding the nature of population interactions during (pre)historic periods when we have little or no documentary evidence to help us, and the potential for designing future studies that may therefore cast light on modern linguistic variation.

11.2 Correlations between genetic and linguistic variation

11.2.1 Nature of genetic variation

Before proceeding further we should establish what we mean by genetic variation. There is not sufficient scope in a short chapter to deal with everything that the last few decades have taught us concerning the nature of the inherited material that underpins biological variation and how this variation arises and changes with time, and interested readers should consult a good text book such as Jobling, Hurles and Tyler-Smith (2004), although Jones (1993) also provides a very approachable introduction to genetic concepts. For our purposes it is sufficient to consider the genome as a large library storing the information required to construct a human. The library contains around 30,000 'instruction manuals' (genes), each of which has the information to make at least one 'machine' (protein) required to build a human body, embedded within a much larger mass of material involved in controlling the activity of these genes but whose function is as yet poorly understood. So an individual's genome can be compared to an adult's grammar: it contains an archive of stored information and mechanisms required to convert the stored information into visible (or auditory) output. Physically the genome is inherited as a group of long polymer molecules of the genetic material, DNA. This you can think of as long strings of 'text' made up of four alternating letters, with the sequence of letters containing the 'information'. Our understanding of the genome has advanced tremendously quickly over the last few years as a result of the Human Genome Project (HGP), a worldwide collaboration of geneticists started in 1990 that led to the publication of a near complete ordered sequence of the 3,200 million molecular letters that make up the text of a human genome in 2003 (www.genome.gov/11006943, IHGSC 2004). It should be remembered that this represents a single copy of the genome sequence; in actual fact we each receive two genome copies, one from our father and one from our mother. As an indication of just how rapid the advances in technology have been in genetics, the HGP project involved hundreds of scientists from several countries working on a collection of DNA samples from different individuals and took over thirteen years at an estimated cost of over $2,700,000,000 to obtain the first nearly complete sequence of a 'Mr Average'. In 2007, barely four years later, James Watson (one of the scientists responsible for elucidating the double helix structure of DNA in the 1950s) and Craig Venter (the founder of Celera, a commercial company that prepared an initial draft sequence in parallel with the publicly funded HGP) became the first individuals to have their whole genome sequence published, at an estimated cost of $1,000,000 each, and in January 2008 the 1000 Genomes Project announced its intention to sequence a representative sample of total worldwide genome variation by sequencing the

genomes of one to two thousand individuals chosen from around the world by 2011, at an estimated total cost of $30–50 million (www.1000genomes.org/page.php). Obviously we are about to enter an era with unprecedented access to genetic information.

Genetic variation is any difference that exists in the sequence of DNA between two individuals, which can be detected either directly by molecular biological techniques (the genotype) or indirectly by its effect on what an individual looks like or how they behave (the phenotype). Phenotypes are the consequence of the genotype, but are influenced to varying extents by environmental factors during growth and development. As we have pointed out above, we should not suggest a direct causative relationship between the genes carried by an individual and the language(s) she speaks, but there are rare instances where genetic variants underlie an inability to correctly form any normal working grammar (Fisher *et al*. 2003), and as more is learned about individual genetic variation it is quite likely that the genetics underlying biological differences in linguistic competence will be discovered. However, that is not the focus of this chapter. Rather, we shall focus on the mechanisms by which genetic variation is created, spread and maintained, and to what extent this might mirror linguistic variation.

For biologists, the level of genetic variation detectible in a species is thought of as arising from a balance between three interacting processes: mutation, selection and drift. Mutation introduces novelty in the first place, and results in changes in the DNA transmitted from one generation to the next, either from errors during copying or as a result of repair of damage from environmental agents such as radiation. Of the approximately 6,400 million genetic 'letters' each of us inherits from our parents, around 500 will be new mutations. If a new variant alters a protein product then it is possible that the protein will perform its job better in that individual and her offspring, leading to an increased representation of the new form of the gene in future generations. This process, known as positive selection, can rapidly increase the frequency of a new variant in a population over a few tens of generations. On the other hand, if the environment changes, the protein made by a variant already present in a population might become detrimental, resulting in that variant being less represented in future generations, a process of negative selection. Geneticists see selection as acting counter to mutation, with both positive and negative selection tending to result in one variant form of a gene being lost from a population, either because a new form spreads throughout the whole population, or because it is selected out. Selection is also the mechanism by which groups of individuals become adapted to their environment and explains the functional differences between groups and eventually species. From the point of view of linguistic correlations, selected variants are unlikely to be useful markers of population history or group similarity since the factors creating the selective advantage for most

genetic variants are unlikely to have anything to do with language (although a causative relationship has been proposed in one recent case between two selected genetic variants and the world distribution of languages that use tone contrastively (Dediu and Ladd 2007).

On the other hand, such selected variants represent only a tiny minority of new mutations. Most mutations do not have any effect on protein function and are considered 'neutral' with respect to selection. The frequency of these neutral mutations will be determined by random sampling from one generation to the next. That is, whether a variant is passed on to a surviving child or is lost will simply be a matter of chance (and note that the vast majority of new mutations are lost by chance within a few generations of creation with only a tiny minority being 'lucky' enough to reach detectible frequencies). The frequency of such neutral variants will therefore fluctuate within a population from one generation to the next, and this process, known as genetic drift, is now believed to be the main cause of the variation found within species (Kimura and Crow 1964). Drift is not a rapid process and it has been shown theoretically to take on average four times the population size in generations for one form of a gene to replace another (Kimura 1983: 49). For our species, where the long-term effective population size is estimated at 10,000 individuals (Relethford 2001: ch. 7), that means around 40,000 generations or 800,000–1,200,000 years if you assume around 20–30 years as the average generation interval (the time between the birth of a child and the birth of that child's parents). So, many of the genetic variants that are present at variable frequencies in most human populations today not only pre-date the individual populations but also the origin of our species.

Any globally distributed species like our own can be thought of as being subdivided into more or less isolated groups of interbreeding individuals. A thought-experiment starting from an initially homogeneous population suggests that, if the landscape were completely smooth, the geographic range of the species large, and individual movement restricted, then the chance of two individuals mating would be inversely related to the distance between them; so-called isolation by distance (Wright 1943; Falconer and Mackay 1996, and for a discussion of related models of variant spread in linguistics see Wolfram and Schilling-Estes 2005). Over time, as many new variants arise and those already present drift in frequency, the individuals in the population will start to diverge across its range. Considering a slice from one edge to the other through the middle of the population, the frequency of many variants will tend to show smooth gradients called clines. Two individuals who are physically near each other will have a higher probability of having the same variants than two chosen from the extremes of the range. If we extend this to sampling two groups of twenty individuals from different parts of the overall range, the combined frequency of many variants within each of these samples compared to the difference between the samples will indicate how far from each other the sampled groups were taken. Doing the same experiment for an initially homogeneous

language spoken by the same individuals, we might find a very similar distribution, local variation drifting in frequency resulting in more shared idiolectic features between individuals in close proximity than those from the distant parts of the range (Chambers and Trudgill 1998). So, under these perfect conditions of isolation by distance alone, one might expect a one-to-one correlation between genetic and linguistic variation as a consequence of the autocorrelation of both with geographic distance. That is to say, individuals who share many linguistic features will be likely to have similar genetic variants because they were born close together, while those with very different genetic make-up will also be expected to have less in common linguistically because they were born far apart.

At this point, a note of caution is necessary; these perfect conditions are unlikely ever to have applied to the human species. Geography is not homogeneous and the global range of our species is littered with barriers to interaction between individuals. These barriers, such as continental edges, oceans and mountain ranges, can be argued to increase the local isolation of groups on opposite sides of the barrier, preventing both genetic and linguistic contacts (at least prior to the creation of long-distance travel and telecommunication), hence enhancing the correlation of genetic and language variation. Similarly, local features such as roads and rivers may facilitate contact between individuals resulting in genetic and linguistic features spreading preferentially along these routes of population contact. Once local differences, either in physical form or in linguistic variants, start to emerge they may become identifiable markers of group identity (see also Montgomery and Beal, and Moore, this volume) to be used preferentially by individuals in mate selection, further isolating these local groups from more distant ones. In contrast, long-distance mass movements of people during historic and pre-historic times will have brought groups with both distinct genetic and linguistic markers into contact, and it is possible that **both** significant language change (ranging from lexical borrowing through to creole formation) and genetic admixture will result from such intrusions. On the other hand, there are clearly attested cases of language change without significant genetic change and vice versa. By unpicking the details of linguistic and genetic interaction consequent on these movements, we can hope to reach an understanding of the distribution of, and some of the reasons for, variation in languages and genes today. Before going on to specific cases, however, we will briefly turn our attention to some of the problems involved in sampling to determine genetic and linguistic variation.

11.2.2 How do you pick your population?

Above, it was suggested that samples from twenty individuals might be used to characterise the population at particular geographic regions across the species range. If, without any other knowledge of the range of variation, we had taken

two samples from physically close regions and a third from further away, we might have been tempted to suggest that we had sampled the first two from one population and the third from a distinct population. Thus we must at this point ask the question of what is meant by 'a population', and how, given limited time and resources, can we obtain samples that represent accurately the variation we are interested in? Similar issues arise when sociolinguists try to define what they mean by a linguistic speech community and how to quantify variation within and between such communities (Patrick 2001; Milroy and Gordon 2003; Eckert 2005; Moore this volume).

The arguments in section 11.2.1 assumed a single undivided and idealised population in which the offspring from a mating tended to stay relatively close to where they were born, effectively resulting in pools of genes and language features that remain close to their sites of origin, and cross the species only slowly as a result of diffusion, rather akin to ripples on a pond when stones are dropped in. That similar processes affect mutually intelligible languages has been recognised since the 1870s in the form of the Wave Model proposed by Johannes Schmidt to explain the distribution of certain features across the Germanic languages (see Fox 1995: 6.3) and many linguistic examples can be found in Labov (2001). So although population boundaries are permeable, distance is effectively creating locally closed gene pools within which variants can alter in frequency without affecting other more distant gene pools. An ideal population in this sense is a group of individuals who mainly interact with each other and only relatively infrequently with those outside the population, and whose membership remains relatively stable over a sufficiently long time scale for unique variants or frequencies of variants to accumulate as markers for population membership – a situation effectively identical to that hypothesised for rural villages underlying dialect continua (Chambers and Trudgill 1998: ch. 1). In the case of continua like these, a regular grid drawn across the area to be sampled and with regular numbers of individuals recorded from within each square is as good a way of determining the local distribution of genetic and linguistic factors as any. Correlations in the more complex real-life situation would also be well served by random sampling, but, with a few exceptions, this has not been how data have been obtained.

Most people would have no difficulty in subjectively dividing the overall human species into 'identifiable' groups which we might wish to refer to as populations for sampling (see Corrigan and Buchstaller, this volume, or Britain 2002, on linguistic sampling). However, these groups would be determined by physical or cultural similarity and the boundaries of groups visible today would almost certainly not match those recognised by people living even two centuries ago. Similarly, the recognised boundaries would be very unlikely to mark the extent of any meaningfully closed or restricted gene pools. Many of the 'populations' recognised today, particularly in Europe, are socio-political

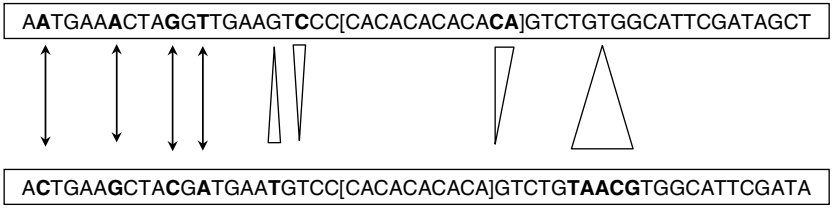
constructs that have been imposed on older, more local patterns of human variation over the last few decades. Geneticists examining big evolutionary questions, such as the date and site of origin of our species, have tended to assume that it is these older patterns that are of interest and that the global mobility of the last few decades will have acted to mask those 'original' patterns (see for example Cavalli-Sforza *et al*. 1991 and Roberts 1992 for a discussion of sampling issues involved in the Human Genome Diversity Project). Genetic sampling for a region has often therefore been by selecting individuals from small isolated groups on the basis of tribal affinity and more often the language spoken. In other words, there has been a simplistic assumption that in recent prehistory the world was a simpler place where most individuals mated within their natal tribe, relatively stable tribal groupings were the norm, and 'except in the case of large modern nations in which the identity of original tribes is usually – though not entirely – lost, languages offer a powerful ethnic guidebook, which is essentially complete' (Cavalli-Sforza *et al*. 1994: 23). Initial work on linguistic dialectology in English applied a similar line of reasoning to focus on sedentary rural communities and more particularly on NORMs (non-mobile older rural males) to elicit variants, mainly lexical items, to build maps of dialect isoglosses used in the construction of dialect atlases, such as the Survey of English Dialects (Orton 1962; Upton and Widdowson 2006; and see Chambers and Trudgill 1998 for a discussion and criticisms of this approach in linguistics and MacEachern 2000 for genetics).

11.2.3 Not all genetic markers are equal: allele frequencies

So far, although we have stated that current variation is a balance between the initial creation of variants and their spread/maintenance, we have effectively ignored the fact that there are different types of genetic change, each with its own individual mutation rate. We can exploit the differences between rapidly and slowly mutating systems to explore different aspects of population history, but first we need to look at the nature of the genetic variants and the relationship between individuals and populations.

The commonest class of genetic variants are the SNPs (single nucleotide polymorphisms) that represent sites in the DNA where a single letter in the code has mutated (see Figure 11.1 for some examples). These 'spelling errors' can take the form of a single replacement of a letter with another or the loss/gain of a letter. Such mutations occur so rarely they are generally considered to be unique events in the history of a species. In other words these are slowly mutating systems where the origin of a particular variant may often pre-date not only the origin of a particular 'population' but also the origin of our whole species.

SNPs are usually found in two forms (called alleles), either the ancestral state, or the derived state resulting from a mutational change. As mentioned



Base substitutions Single Nucleotide Polymorphisms (SNPs)	single base insertion deletions	Microsatellite mutations a $(CA)_6$ on the top line $(CA)_5$ on the bottom	Small deletions or insertions of a few bases. Can also be a few tens of bases or many thousands of bases. Large deletions are often associated with observable phenotype due to loss of genes.

Figure 11.1 A few types of variable genetic markers discussed in this section and in 11.2.4 below

The DNA strands are shown as a line of letters corresponding to the four 'bases'. Two alternative strands are shown with possible changes in sequence between them indicated in bold. Variational changes can occur in either direction so the 'ancestral' allele could be the top or bottom in each case – see text for discussion. These are only a sample of the types of variants found in the human genome for illustrative purposes, for more details of these types of variation and how they arise refer to any genetic textbook, particularly Strachan and Read (2004: ch. 11).

earlier, each of us has two genome copies, so an individual's genotype can then be said to be either homozygous (carrying two identical copies of either the ancestral or derived allele) or heterozygous (one copy of the ancestral and one of the derived alleles). Populations can be characterised and compared either by the number of individuals of each genotype or by the frequency of each of the two alleles. It is not even necessary to understand the underlying genetic mechanism in order to use resulting, observable phenotypes to describe populations in this way, but you do have to determine the population to sample.

Developments in protein chemistry in the 1950s and 1960s led to the realisation that variation was common in many different blood protein systems and a large body of population data rapidly accumulated in blood banks (Mourant, Tills and Domaniewska-Sobczak 1976). These data clearly indicated that there were many different patterns of allele frequency and that most of the systems showed independence; in effect, there was no 'single history' relating the populations sampled. Indeed, apportioning the total world genetic variation into components found within and between 'races' (or continental groups) and populations, Lewontin (1972) demonstrated that more than 80 per cent of the total variation for most genetic systems lies within any single local population and that less than 10 per cent represents the differences between the major

'recognisable races' (see Barbujani 2005 for a review of more recent research and a discussion of the biological invalidity of the race concept applied to human diversity).

The conclusion is that human populations are made up from inter-fertile individuals and people can, and do, pick their mates from outside their natal group. Present-day populations are therefore composed of individuals whose genes derive from diverse ancestral populations. Since the genes in each individual are themselves a random mixture of their parents' genes, every individual can be considered as a 'population' of variant forms, each of which may have a different history. If we were to walk along each human chromosome as if it were a street, and each house (gene) that we pass represented a different architectural style characteristic of the time period when that genetic stretch entered the British Isles, what would we see? Perhaps a shelter from the Palaeolithic hunter-gatherers representing genes that entered Britain when it was a peninsula of Europe, 10–12 kyrs before the present (BP), next to a Roman villa representing genes from a retired Roman legionary, and a Viking longhouse, from genes left by a raiding Dane, who himself returned to his homeland after one summer here. So every individual can be considered as a 'population' of variant forms. Modern molecular approaches that characterise thousands of genetic systems in an individual allow researchers to detect many of these and thus allocate individuals more or less accurately to a geographical area of origin on statistical grounds (Rosenberg *et al*. 2005). Similarly, the variety of English an individual speaks as an adult may well be composed of different components each with its own history, including those features influenced by the learner's parents and those influenced by the speech communities the individual interacts with. Linguistic incomers to a population, whether by physical or social mobility, may 'acquire' linguistic features from that resident population while retaining most of their native variety, or alternatively may initiate novelty that spreads into the local variety without completely replacing it. Hence although it is often convenient to think of a language or dialect of an area, or even of an individual, as a single entity, it must be borne in mind that in some cases a 'speech community' will involve speakers of different varieties or, in the case of *sprachbund* areas such as the Balkans, even languages from different families (see Trask 1996: 11.2, or Heine and Kuteva 2005: ch. 5). So here again we have an expectation of similar processes acting on genetic and linguistic variants.

The increasing availability of genetic data has been accompanied by the development of methods for summarising it so that the 'most interesting' patterns can be easily detected and interpreted. There are many different procedures for converting observed population frequency into measures of 'distance' between populations, and varied approaches to combining and summarising these distances using either numerical or visual approaches (see Barbujani

2000). However, there is no 'one best method' and the selection of techniques is more of an art than a science, with the choice often being determined by the type of story the researcher wishes to tell.

The presence of a pattern of allele frequencies seen today may have very little to do with currently recognised populations and it is impossible to take such patterns as **strong** evidence for any **particular** historical event. The writing of a recent history on the landscape, like a medieval palimpsest, may not have completely removed the patterns of the past. In some areas change will overlay and reinforce prior signals, while in others old and new will become confused, masking both, and it is often impossible to distinguish events that occurred 100 years ago from those of 25,000 years ago (see McMahon 2004 and references therein). These problems are amongst those that have led to scepticism in applying genetics to illuminate linguistic problems.

In the next section, we will look at molecular genetic methods that focus on individuals and offer the possibility of unpicking this confusion by dating and characterising the different genetic contacts between populations.

11.2.4 The Y-chromosome and the mitochondria – haplogroups, founders and dating in genetics

The mutations giving rise to SNPs are so rare that we can consider each as a unique event, and any DNA molecules with a particular mutation today must share a single common ancestor at some time in the past. A mutation occurring in the recent past will therefore be present on a molecule carrying older changes. We can exploit the physical linkage of these changes by aligning molecules into nested groups with shared variants to give the temporal sequence in which the mutations occurred. Trees based on these 'shared derived characters' then link modern molecules to their common ancestors (often called the MRCAs for most recent common ancestors, or 'coalescents', as the mutational lines of extant groups coalesce at those molecules (Rosenberg and Nordborg 2002). In these trees each ancestral mutational event is shown as a branch point, or node, marked by a 'reconstructed' sequence of variants. This process, known as phylogenetic reconstruction (see Page and Holmes 1998 or Felsenstein 2004), bears many similarities to the comparative method in linguistics (Durie and Ross 1996). Each node, and leaf, represents the sequences present on a single DNA molecule at a list of potentially variable sites and is known as a **haplotype** – you could think of these like feature bundles of the lects spoken by individuals. As we noted above, most of our DNA occurs as two copies, so we have two haplotypes for most regions of our genome. Unfortunately for phylogenetic reconstruction, you do not receive a complete copy of one or other grand-paternal haplotype present in your father, but rather a random mixture of both grand-paternal haplotypes recombined

together. This makes phylogenetic reconstructions from much of the genome a challenging process.

However, cells do contain two unpaired DNA molecules: mitochondrial DNA, and the Y-chromosome. The mitochondria are cellular structures responsible for energy processing and contain small closed-circle DNA molecules (mtDNA). Mitochondria are present in most cells (with up to 1,000 copies per cell), but are passed on to children only by mothers. In contrast, the Y-chromosome is one of the sex-determining chromosomes, and if present, results in the bearer developing into a male, so the Y-chromosome is passed from father to son. Here we have two complementary molecules, one passing through the female line and the other the male line, that are free to accumulate mutations without recombination.

Figure 11.2 illustrates the passage of such a system from the past through to the present and a reconstruction of the relationship between the current molecules. By counting the average number of mutations between the 'root' and the extant molecules (the so-called *rho* (ρ) statistic of Forster *et al*. 1996), we can estimate the time elapsed since this common ancestor. In this example, concentrating only on the left-hand population of molecules in 11.2a and considering the 'root' molecule marked by the arrow in 11.2b, there are 9 molecules sampled (1–6, and A–C) with a total of 18 mutational steps from the root; so $\rho = 18/9 = 2$. We next assume that mutations occur at a regular rate like a 'molecular clock' (see Bromham and Penny 2003 for a review and history of the molecular clock hypothesis in biology). So simply multiplying 2 by the mutation rate, say 1 in 10,0000 per year for a stretch of DNA consisting of 1,000 letters, gives an estimate for the time since the observed molecules shared a common ancestor of 20,000 years BP in our example.

We can take this a step further by subtracting the observed haplotypes in one population from the other to give us the 'grey' circles in Figure 11.2b. Since molecule 3/8 is found in both populations we can assume that this was a 'founder' haplotype carried into the second population from the first. Then assuming the more diverse population (the left-hand one) is the 'source' population we can ask how much variation is restricted to the right-hand population since the founder entered. In this case, three related molecules are present, one each with zero, one and two mutations from the assumed founder type. This gives a *rho* of one mutational unit, or an age of 10,000 years since the contact. Obviously this is a crude estimate based on many assumptions. Although **relatively** constant, mutation rates have high stochastic variability; the human generation interval may have varied over time (and indeed, where it has been measured, is different for mtDNA and the Y-chromosome due to different average age at child bearing, e.g. thirty-five for men versus twenty-nine for women in Canada (Tremblay and Vezina 2000)); and what is the correct 'mutation rate' to use when we know that different sites in the genome and different genetic systems

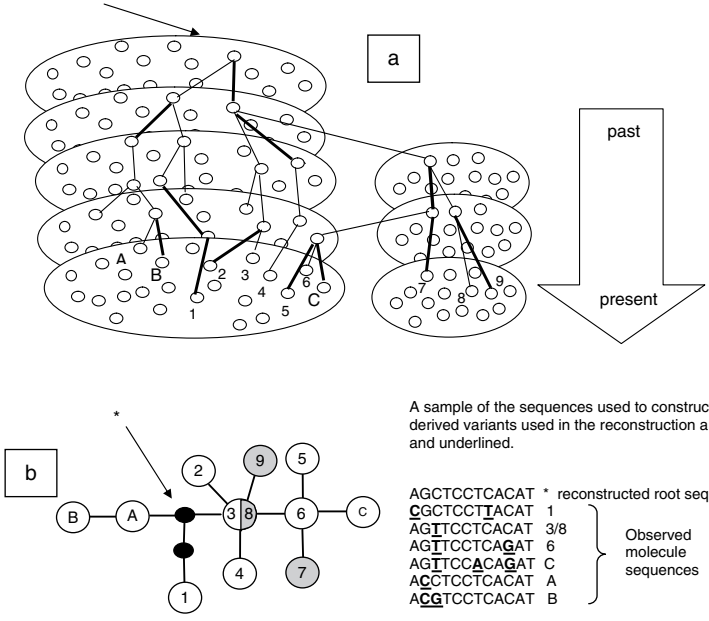


Figure 11.2 An example of phylogenetic reconstruction in molecular genetics

(a) shows the genealogy of twelve groups of molecules sampled from two present-day 'populations'. Each larger oval represents a population at a successive time point. Lines indicate inheritance between successive time points tracing the current molecules back to their common ancestors. Circles represent groups of individuals in each generation. Thick lines correspond to the inheritance associated with a 'new' mutational change. A and B therefore differ by one sequence change (the C to G at the third base of the sequence in (b)), while A and C differ by four.

Using this variation, an unrooted phylogeny or 'tree' of molecular diversity can be constructed by nesting groups of shared derived characters – see text for details.

(b) gives an unrooted phylogeny of the twelve extant molecules A, B, C and 1–9, based on sequence data like the examples on the right. Shading of circles distinguishes sequences drawn from the two populations above. Small black circles represent 'reconstructed nodes' not detected in the samples, but required to connect the observed sequences. Each node in this phylogeny represents a haplotype and groups of haplotypes sharing common ancestral state mutations are termed haplogroups. So C is a haplotype within the haplogroup (5,6,7,C) which we could call haplogroup 1 or define by the presence of the derived state (G) of the C to G mutation between 3/8 and 6.

have their own considerably different rates? At this point you might be beginning to wonder whether or not there is anything to gain in looking at these molecules at all! What I want to do here is to emphasise that the genetic processes are subject to as many uncertainties as the linguistic processes in terms of the assumptions of the underlying mechanisms responsible for detectible variation. In biology this is often apparently ignored, but is implicitly included in the calculation of confidence intervals associated with estimated values, so, as an example, the age of the mutation responsible for converting the A allele to the B allele in the ABO blood system is estimated using haplotype analysis to be 3.5 million years old, but with a 95 per cent confidence interval of 2.64–4.36 million years (Calafell *et al*. 2008).

Haplotypes can combine the unique, but rare, mutation processes responsible for SNPs, to mark deeper time events, with more rapidly mutating systems such as microsatellites (see Figure 11.1) to mark recent events. Microsatellites on their own generate mutations at a rate of 1/1000 or even 1/100 transmissions, bringing the power of molecular dating into the correct time scales for looking at linguistically relevant population interactions, but the rapid mutation process (a sort of molecular stutter) can also result in molecules that have the same number of repeats without shared ancestry. By examining regions of DNA that contain both slow and fast elements, these recurrent mutations often occur on different SNP backgrounds, and so do not become confused. The MtDNA and Y-chromosome represent two very powerful genetic systems that, if fully exploited, could yield a vast amount of information. The frequencies of different haplogroups can be compared between populations as if they were single SNPs, and subsequent founder analysis within haplogroups used to date population interactions. However, it must be emphasised that each behaves as a single super-locus and, while around 200 years ago an estimated 2^{10} (1,024) ancestors contributed to any living child's genes, only one female gave him her mtDNA, and one male his Y-chromosome. So these are powerful but highly selective story tellers, whose history may, as a result of selection, population demographics or mere chance, be very different from the majority of genes present in a given population. In the future whole-genome sequencing will allow founder analysis to be extended to include the majority of the DNA, but today we are limited to these two.

Even for the mtDNA and Y-chromosome, it is only in the last decade that haplogroup phylogenies have become sufficiently detailed to answer questions about local European population histories (Jobling and Tyler-Smith 2003; Pereira *et al*. 2005; Karafet *et al*. 2008). So it may be unsurprising that, as we shall see in §11.3, published genetic analyses of the UK are far from conclusive with regards to questions concerning linguistic history. However, the potential of molecular data to confirm written records is illustrated by the recent observation that 8 per cent of Asian men tested (and by extrapolation

16 million others not tested) share a Y-chromosome that derives from a single male ancestor living 600–1,300 years ago (Zerjal *et al*. 2003). These chromosomes are present in at least sixteen different populations, and while it is possible for genetic drift to result in a single population with a high frequency of a recent Y-chromosome, repeated random sampling is unlikely to have had the same effect in so many. The original bearer of this chromosome is assumed to have been Genghis Khan (circa 1162–1227) and his immediate male children, and the mechanism driving the spread was the differential reproductive success of this prestigious elite within the Mongol empire. Thus genetic data appear to corroborate the apparently outrageous historical reports that Genghis Khan had fathered a dynasty of more than 20,000 descendants less than 100 years after his birth (Juvaini 1260). In the absence of a DNA sample from him or his children we cannot unambiguously prove this hypothesis, nor can we rule out another selective force, such as disease resistance, creating the observed distribution. However, the isolated presence of this marker in the Hazara population of Pakistan, whose oral history claims direct descent from the Mongols, argues in favour of this explanation.

In this example a large number of individual DNA samples had been taken across a geographical range of sufficient size for the exceptional pattern of variation to be visible. Indeed it was only when the very recent nature of the spread was determined that the pattern became exceptional in relation to models of expected population variation and change. So dissecting variation into different sub-systems (in this case haplogroups and haplotypes) and looking at each independently and in relationship to each other provides insight into some of the factors influencing present-day variation that would not otherwise be apparent.

11.3 Genetic clues to linguistic history: analysing variation in English

11.3.1 The origins of English

To understand genetic variation, then, we need to recognise that there are different types; we also require models of the processes that affect the creation and spread of each type and at least a partial understanding of the history of population movements and interactions that affect how individuals have got to where they are today. The situation is the same for varieties of English: current variation is a product of linguistic processes acting in the short and longer terms to create and distribute linguistic features within speech communities, combined with the history of those speech communities and their interactions with others. Understanding the movements and compositions of ‘groups of people’ can thus contribute to understanding both genetic and linguistic variations and it is certainly possible that clues from one domain will help to understand variation in the other.

There is archaeological evidence of humans in the British Islands for hundreds of thousands of years, but because of periodic fluctuations in weather this has been a far from continuous occupation (for details of the climatic and archaeological picture see Stringer 2006). We can say little about the language(s) spoken by these groups, since we have no written records concerning them until the time of the Romans, and even then the picture is at best fragmentary. What is generally believed, is that around 2,000 years BP the inhabitants of Britain were tribal groups speaking some variety of Celtic related to modern Welsh (Brittonic), Irish (Goidelic) or Pictish (unknown, guessed to be Brittonic) (Cunliffe 2003). In the nineteenth and early twentieth century archaeologists and historians saw the prehistoric period in Britain as a succession of population movements and replacements from the continent (see, for example, Stenton 1947), but in the 1980s a more procedural line of thinking, and detailed archaeology, led to the belief that language and cultures may have spread many times across Europe without population replacement (Renfrew 1987; Higham 1992). Such differences in viewpoint will inevitably affect how similarities in genetic and language variation are interpreted and we will return to these issues below.

The cultural exchange associated with the birth of English in particular was historically seen as a period of expulsion, enslavement and extermination of the resident populations. Commencing in the year AD 449 (according to the Anglo-Saxon Chronicles) the fifth and sixth centuries saw groups of Germanic-speaking peoples from Northern Europe crossing the channel and setting up home. What is not clear, because we have no written records until much later, is what dialects or languages were spoken by the incomers, and exactly where and in what numbers they landed. Bede's (731) *Historia ecclesiastica gentis Anglorum* identifies three separate groups of invaders, the Jutes from Denmark who settled the Isle of Wight and Kent, the Saxons from north-west Germany who settled much of the south of England, and the Angles from the south of the Danish Peninsula who settled in East Anglia, the Midlands and the area that became Northumbria. These general classifications may hide a more diverse continental origin, although all can be assumed to have spoken some variety of West Germanic.

Social interaction between populations, be it for trade, co-occupancy of a region or colonisation, will result in genetic exchange to a degree dependent on the intensity of contact. Different types of population interaction will also affect language, with contact-induced changes varying from simple lexical borrowing to complex interference and language shift or replacement (Thomason and Kaufman 1988; McMahon 1994). The influence of the prior residents of these Isles in the formation of Old English is still open to dispute. The traditional view holds that 'outside of place-names the influence of Celtic upon the English language is almost negligible' (Baugh and Cable 2002: 69) and this

lack of apparent influence has been taken as support for the completeness of the destruction of the Romano-Celtic tribes assumed to be a relatively homogeneous presence in these islands before the invasion. However, borrowing of lexical items is often unidirectional and related to the relative status of the languages, and the borrowings into Old English, being limited to a handful of place names and related cultural items (Scheler 1977; Coates and Breeze 2000), might also be expected in a highly differentiated society. The Laws of the Wessex king Ine, from around 694, make special provision for the (lesser) legal rights of the Britons living in his kingdom, suggesting that, at least in some areas, extermination was incomplete and socially stratified 'mixed' populations existed for some time (Grimmer 2007). This can be seen as the foundation for the so-called 'Celtic Hypothesis' (see papers in Filippula and Klemola 2009). This alternative to the traditional view, which has gained popularity in the last few years, holds that several features of English phonology and (morpho)-syntax can be best explained as resulting from contact between West Germanic and the Celtic of the resident population(s). Partial population replacement is unlikely to have resulted in homogeneous proportions of incoming to resident individuals across the country, possibly reflected in regional variation in the proportion of Brittonic features in place names (see, for example, Coates 2007). Similarly, we might also expect some consequent local differences in Celtic-influenced features in the resulting English dialects (e.g. Klemola 2009). If any modern dialect variants reflect underlying differences in population admixture at the time of the invasion they might also be reflected in differential genetic signals of admixture. From evidence in texts from the seventh to tenth centuries, Old English was already split into Northumbrian, Mercian, Kentish and West Saxon by that time. Does this reflect variation generated within Britain either with or without influence from Brittonic speakers, or is it a result of linguistic variation already present between the incoming groups and their subsequent interactions?

From the discussion above we might be able to ask three relevant questions that we could hope to answer partially. Firstly, can genetics establish the extent of population replacement/admixture? Secondly, are there detectible genetic signals specific to the geographical areas or tribes that acted as the source populations for the invaders? Lastly, do any of the genetic patterns match the approximate distribution of dialect boundaries of modern English? While the data currently available can give us pointers as to how these could be addressed, they are at best a low-resolution first step towards providing answers.

11.3.2 Did fifth-century England witness a case of complete genocide?

Although population frequency data could be used to address this question, we will concentrate on three studies that arguably provide the best current

molecular evidence, while at the same time illustrating some of the limitations of genetic data.

The east and southern parts of Britain are undoubtedly genetically closer to the populations of mainland Europe than are those of Wales, Scotland and Ireland. Weale and co-workers (2002) used Y-chromosome haplogroups to investigate this pattern by analysing males from six rural market towns arranged in an east–west transect from East Anglia (North Walsham) to North Wales (Llangefni). They argued that these established market towns mentioned in the Domesday book were likely to contain a high frequency of non-mobile-rural-males with local farming ancestors, so providing a genetic picture relatively unaffected by recent population movements. They characterised the frequencies of different Y-haplogroups in each population, and analysed their results as if each sample was generated from an admixture of allele frequencies found in Wales (Llangefni, representing the pre-Anglo-Saxon Britons), and those from modern Friesland (representing the source population for the 'invasion'). They identified a near hundred per cent population replacement of males in eastern England, with a strong genetic discontinuity between the Welsh and English towns. Of course, one could argue that a handful of samples taken in one small strip of England is hardly representative, and Capelli *et al*. (2003) have extended the Weale analysis to cover twenty-five sample points spread in a grid pattern across the British Isles. Even though they only have twenty-five sample points, they detect a much more complex picture than Weale *et al*. Anglo-Saxon male influence appears to have been highly variable, with Wales, the south of England and Lowland Scotland having relatively little replacement of 'indigenous' chromosomes (around 30 per cent or less) compared to the central and eastern parts of England (approximately 60 per cent replacement for York and Norfolk). A genetic summary of these results placed alongside similar samples drawn from other studies is shown in Figure 11.3, redrawn from Oppenheimer (2006: 369). Figure 11.3 is a two-dimensional summary of the genetic 'distance' between sampled populations. The towns in Weale *et al*. (2002) are joined by a line in the figure and, as expected, they represent only a small proportion of the total variation. Note that, although the combined North Welsh towns (18) **are** distinct from their English neighbours (23, 24), the central Welsh town of Llanidloes (26) is not. Genetic distances between groups match, to some extent, the geographical relationships between the samples with a few interesting exceptions, such as York clustering with Norfolk, and Belgium lying within central England between East Anglia and Uttoxeter!

Oppenheimer (2006) has further analysed these results by dating the molecular variation within the British Isles using founder analysis (see section 11.2.4) and suggests that many of the haplogroups shared between eastern England and the continent actually entered the British Isles well before the fifth century. In fact he distinguishes several different periods of contact between the Continent

and the British Isles, with a significant proportion dating to the period when Britain was a peninsula of Europe rather than an island (earlier than 10,000 years BP, Behre 2007). He therefore recalculates a fifth-century continental input of around 9–15 per cent in Norfolk and the Fens, dropping to around 5 per cent for the rest of England including the south coast, and suggests that an ancestor of English may well have been spoken in the east of England at the time of the Roman occupation.

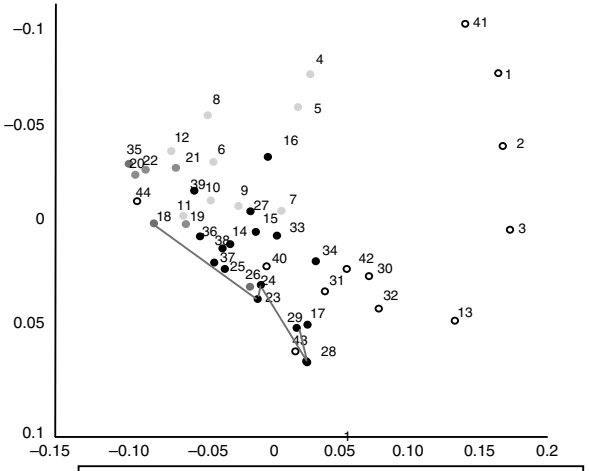
On the basis of these papers, genetic evidence suggests the degree of replacement is anything from 100 per cent to 5 per cent of male lines, illustrating several of the difficulties inherent in interpreting current variation. In analysing the results, the first two groups of researchers have assumed that two current populations can act as proxies for those of the fifth century, with Castlerea/Llangefni representing the entire pre-Anglo-Saxon population of Britain and the current Northern Germanic/Frisian population, the invaders. They then assume that shared allele frequencies between British and continental European populations are the result of population movement only in their period of interest, overestimating the contact by ignoring the palimpsest of older and more recent contact events, which may be partially responsible for the position of the French and Belgian population samples 'within' the English cluster in Figure 11.3.

Oppenheimer, for his part, may have overestimated the proportion of older contacts by assuming that when the Anglo-Saxons invaded only a small fraction of any population actually migrated, so that variation present in Britain has arisen only in Britain, and that no variation has been lost completely from the continental source. Studies of a handful of skeletal remains from Iron Age (2,000–1,800 years BP) villages in Denmark demonstrate clear variation in mtDNA between sites of low status and sites of high status at that time, with only the latter close to present-day Danish patterns (Melchior *et al*. 2008) – there is insufficient evidence at the moment to know whether the population(s) that moved to Britain reflect this variation, or whether they were a biased subsample. The fifth century was a period of sea-level change when whole communities in coastal northern Germany were abandoned (Myres 1989). Some of these communities may have migrated to Britain en masse bringing both dialectal and genetic variation to their new home. This variation would then have become part of that specific to the founder population, even though it pre-dates the movement and, as Barbujani, Bertorelle and Chikhi (1998: 489) note when comparing the histories of molecules and the populations they find themselves in, 'suppose that some Europeans colonize Mars next year: if they successfully establish a population, the common mitochondrial ancestor of their descendants will be Paleolithic. But it would not be wise for a population geneticist of the future to infer from that a Paleolithic colonization of Mars'.

Further genetic data from extant and archaeological populations of Europe and Britain may allow us to quantify the extent of interchange prior to the fifth century, but there are good reasons to think that Oppenheimer is correct in positing cultural and genetic exchange with the Continent prior to the Roman invasion. However, his suggestion that a Germanic language may have been spoken extensively in eastern and southern Britain at this time, by groups related to the Belgae, must remain a conjecture (Oppenheimer 2006: 267–92).

Even though the data are currently insufficiently detailed to be conclusive, the genetic evidence indicates that there was a significant movement of continental populations (males at least) into the east and south of Britain around the fifth century, but for most of the country the incomers represented a relative minority. We must therefore reject widespread genocide as an explanation for the lack of Celtic influence on early English, and look rather for social interpretations.

In Figure 11.3 the influence of a later period of 'Germanic' invasion associated with the Vikings can also be seen. However, the Danish Vikings and the Angles/Jutes may have originated from similar geographical regions and hence have quite similar genetic signatures and related dialects. Also the areas of highest recorded Danish Viking activity in England, along the east coast particularly around York and the Wash, overlap extensively with the area of earlier putative Anglian invasion, making it difficult to distinguish these separate events and their contribution to the similarity of Norfolk and York in Figure 11.3, for example. On the other hand, Norwegian populations have distinctive markers in both mtDNA and Y-chromosomes, making it possible to say that around 30 per cent of Orcadians and 40 per cent of Shetlanders have Norse ancestors. In the Western Isles, the Isle of Man and parts of coastal mainland Scotland this falls to 15–20 per cent male and 10–15 per cent female input. The linguistic consequences for these populations were extensive, with Shetland retaining its own distinctive variety of Norse (Norn) into the eighteenth century, while the Scots and English dialects of both Shetland and Orkney remain heavily influenced by Norn, at the lexical and phonological level (Barnes 1984). Mainland Scots has also been influenced by Norse characteristics, although whether Scots originates from the fifth/sixth-century Anglian of old Northumbria or tenth-century Anglo-Norse originating from Yorkshire is fairly unclear (Macafee, 2004). Within England, Norse influence on place names is rare south of the Danelaw, a hypothetical line drawn roughly from London to Chester (approximately the line of the A5 road today) set up to separate the area subject to Danish Law (to the north of this line) from Saxon Law to the south. Although the exact position of this line is unclear, there is plenty of place-name evidence suggesting substantial Norse presence within the Danelaw, and northern varieties of Middle English show extensive lexical and grammatical characteristics that may have come from Norse interactions. As Thomason and



Graph shows genetic distance between 'populations'. Note that these data come from different source papers and populations were not sampled in identical ways. Redrawn from Oppenheimer (2006: 369). Closed circles represent populations from the British Isles, open circles other groups.

Sample sites arranged approximately North to South

1. Trondheim (Norway)
2. Bergen (Norway)
3. Oslo (Norway)
4. Shetland
5. Orkney
6. Durness
7. Western Isles
8. Eastern Scottish Mainland
9. Western Scottish Mainland
10. Stonehaven
11. Pitlochry
12. Oban
13. Denmark
14. Morpeth
15. Penrith
16. Isle of Man
17. York
18. North Wales
19. Irish (2 samples)
20. Castlerea central Ireland
21. Rush (nr Dublin)
22. Irish West coast
23. Ashbourne
24. Southwell
25. Uttoxeter
26. Llanidloes
27. Sheringham (coastal Norfolk)
28. Fakenham (central Norfolk)
29. North Walsham (E. Norfolk)
30. North West Germany
31. Frisia
32. Dutch
33. Chippenham
34. East Anglia
35. Haverfordwest
36. Faversham
37. Midhurst
38. Dorchester
39. Cornwall
40. Belgium
41. Central Germany
42. Bavaria
43. France
44. Basque region

Figure 11.3 Genetic distance map of selected Western European sites

Kaufman (1988: 264) observe, 'the Norse invasion and possibly heavy settlement in certain areas of Britain during the ninth and tenth centuries is a fact with linguistic consequences'. Viking genetic influence seems relatively small and patchy outside the areas of York and the Wash, and it remains to be seen if more detailed sampling in the future will detect any relationship between the distribution of Norse dialect features and genetic variants in English varieties.

In conclusion, while there are no **definitive** answers to the question of whether genetic variation can tell us anything useful about linguistically relevant population interactions, there is clear indication from the studies above that variation does exist between relevant populations at a level that can be exploited in the future. There are also strong indications that similar patterns do exist at some levels between genetic and linguistic variation within the British Isles. What is not clear is how best to **interpret** these similarities in

terms of particular past events, and the influence of those events on modern variation. However, there is hope; we can already use genetic data to argue against the once widely held view that the 'birth' of English was accompanied by near complete genocide, and in the last section we shall look at some possible questions for the future.

11.4 Where next?

Varieties of English worldwide owe much of their differences in character to exchanges between interacting populations both during the establishment of each regional or social variety and its subsequent development. Thomason (2005: 687) suggests that 'most of what historical linguists study under the designation 'language change' is due to contact … The changes we investigate therefore tend to be those that have spread throughout a speech-(sub)community, and the process of spread is a function of contact between speakers.' In this chapter, we have established that such population interactions may in some cases also leave a genetic signature that can illuminate the nature of the exchange. In §11.2 we observed that different genetic systems have different characteristics and that there is no one system applicable to the study of all population interactions or time scales. Only a small proportion of total genetic variation can be related to any particular event or time slice and only a proportion of interactions leave both a linguistic and a genetic signature. So the degree of illumination generated by genetics is rather akin to a candle flame in a cathedral, and the power of genetic variation to illuminate population movement and cultural history may have been somewhat overstated in the past. We should not, however, reject such analysis out of hand, since a candle in the right place can illuminate a historical text well enough to read its message, and the prospects for more detailed and informed investigations are becoming a reality.

In §11.3, we looked at how recent papers have attempted to apply such illumination to the population events associated with the origin of English, with success mainly limited by the availability of suitable samples. Technology is now developing to permit testing of vast numbers of genetic variants in individuals, which promises to allow the combination of the benefits of haplotype analysis with detailed population methods. However, informed sampling is a key part of investigation, and this is an area where linguistics input would be required in future experimental design. We have focused on a limited time period in §11.3, basically from the pre-Old English period up to Middle English, and have ignored any questions concerning the more recent history of Modern English variation. Lack of sufficiently detailed genetic data has prevented us from examining more recent population movements and their influences, either from outside the Islands, in the form of the Norman conquest or more recent

immigrations and their effect on urban varieties, nor on movements within Britain associated with social and technological change that have led to varieties such as Milton Keynes, or the Midlands influence on Modern Standard English (see, for example, Fennell 2001 and Kerswill 2006).

Another area where there is currently no data to speak of, but where genetics could be very illuminating, is in the influence of population admixture in the founding of World Englishes. The last few centuries have seen a dramatic expansion in the range of English usage, initially associated with population and cultural emigration from Britain, and subsequently promoted by globalisation. While genetics cannot hope to cast any light on the latter process, interactions during the colonial period may have left genetic signatures of population admixture underlying some modern dialect variation. Nichols (1997: 372) recognises three types of language spread: language shift, demographic expansion, and migration, and we might want to extend this to include the spread of varieties. Each of these spreads is likely to leave different substratal signatures in the derived linguistic varieties and in the genetics of the populations speaking those varieties (although it must be remembered that individual features and genetic markers **can** spread across 'population boundaries' independently). English has probably been influenced by or initiated all three forms of interaction during different times of its history, and genetic studies, if performed with care and sufficient resolution, may help to determine where, when and the extent of such population interactions, and thereby help to understand the distribution of linguistic variation today.

For details of how to date genetic molecules, particularly with reference to Europe you should read Richards *et al*. (2000), and good reviews of the Y-chromosome and mtDNA can be found in Jobling and Tyler-Smith (2003) and Torroni *et al*. (2006) respectively.

Two books that you might like to read with particularly positive attitudes to the possibility of combining the study of cultural (including linguistic) and genetic variation, and the methods involved in doing so are Stone, Lurquin and Cavalli-Sforza (2007) and Cavalli-Sforza (2000); but you should contrast these with the arguments in McMahon and McMahon (1995) and Sims-Williams (1998).

An overview of World English varieties and history can be found in Crystal (1995), British English in Trudgill (1999) and American English in Labov, Ash and Boberg (2006) (or the website of the American Linguistic Atlas project (http://us.english.uga.edu/). Historical aspects of Early English can be found in Baugh and Cable (2002) and more modern varieties of World Englishes in Cheshire (1991) or Singh (2005).

12 Variation and education

Graeme Trousdale

12.1 Introduction

Significant numbers of research projects on linguistic variation and change in English have focused on the language of adolescents, who are often seen as the driving force behind the propagations of linguistic innovations (see, for instance, Cheshire 1982a; Eckert 2000; and Moore 2003). Young people, therefore, bring into the classroom a wide range of linguistic forms, some of which may be established and stigmatised, while others may be emergent and not yet subject to overt evaluation. This fact has long been recognised by educators, yet there is often a lack of consensus as to how best to treat dialect variation in the classroom. Furthermore, the relationship between variation in English and educational policy and practice must always be seen in a political context. This manifests itself in a number of different ways: for example, in the debate on African American English in schools in the United States, or in the treatment of non-standard accents and dialects in the development of a national curriculum in England and Wales. In what follows, I address some of the ways in which research into variation in English has helped to inform aspects of educational policy in different parts of the world, as well as some of the ways in which a lack of understanding about the nature of variation (as a consequence of the marginalisation of the study of linguistics in the classroom) has caused significant problems for government, teacher, parent and student alike. I also highlight some of the ways in which linguists working with educators (and with policy makers) can help to clarify the relationship between standard English and other varieties of the language. Conversely, I show how working with teachers and students can also help researchers collect useful data which furthers our understanding of the nature of and constraints on variation in English.

Since I am most familiar with the educational systems in the United Kingdom, much of the evidence is drawn from projects and policies in the countries which make up the UK. Some of this is generalisable to other communities, however,

I am grateful to Dick Hudson, and to the editors of this volume, Warren Maguire and April McMahon, for very helpful comments on an earlier version of this chapter.

and this is pointed out where relevant; evidence from other countries is also discussed in various places below. My focus is predominantly on variation in form, and specifically on the relationship between standard and non-standard English in the classroom (e.g. the difference between standard *I was visiting my sister* and non-standard *I were visiting my sister*, or between *I don't like those pictures* (standard) and *I don't like them pictures* (non-standard). Focusing on these issues means I will have little to say about other ways in which linguistic variation and education interact (e.g. in terms of gendered language, or the discourse structure of oral and written narrative), though again, such more general issues will be addressed where relevant. Specifically, I have tried to write this chapter so that it will be useful to linguists who are concerned to know how dialect variation is dealt with in the English classroom, and to teachers and educationalists looking for information about varieties of English and their place in the classroom. As a result, some of the observations made below may be well known to one group, but perhaps less well known to the other.

The chapter is structured largely around the themes of the present volume. An example of educational policy regarding variation and English is discussed by way of contextualisation (§12.2), §12.3 is concerned with what we know about variation and education, §12.4 with what we don't know, and how we might find the answers, and §12.5 with why it matters. The final section (§12.6) outlines some key references.

12.2 Variation and education in a political context: an example from England and Wales

The National Curriculum (NC) in England and Wales, introduced in 1988, brought about significant changes for the teaching of English in those countries.[1] It should be noted that the NC provides guidance to teachers about expected levels of attainment and programmes of study in all subjects (i.e. not just English) in schools in England and Wales. Hudson and Walmsley (2005) suggest that the NC has brought about improvements in the teaching of grammar; this may in part be due to the fact that its approach to grammar teaching is essentially non-prescriptivist,[2] and that diversity is not condemned as wrong, but examined for its own sake. Standard English still has an important role to play, and teachers are still under the obligation of ensuring that their students are familiar enough with the standard variety that they can use it effectively when the situation arises. But the consequences of this approach to grammar for variation and education (in England and Wales, at least) are far reaching. In their discussion of the teaching of non-standard grammar, Hudson and Walmsley observe:

> Some children will be able to induce the standard rules for themselves, but others will not; those who cannot do this for themselves may benefit from explicit instruction. This logic leads to an even more radical innovation in grammar-teaching: that a teacher

might start by considering the non-standard grammar system as a basis for comparison with the standard one. (Hudson and Walmsley 2005: 614)

The critical issue here is that non-standard varieties – the language that most children bring into the school classroom – may be seen as a resource, available for use in developing literacy skills, acquisition of the standard variety, and knowledge about language more generally.

The NC for English (DfEE 1999) makes particular mention of language variation at all levels, from Key Stages (KS) 1 and 2 (the levels associated with primary education, from age five to eleven) up to KS4 (when children sit national examinations known as GCSEs (General Certificate of Secondary Education) at age sixteen). Although the importance of the standard language is recognised in all subjects in the NC, the focus is understandably greater in the curriculum for English. Yet non-standard varieties are also subject to specific comment in the English NC. The cultural development of the child, for instance, can be furthered through learning, among other things, 'about language variation in English, and how language relates to national, regional and cultural identities' (DfEE 1999). The approach to standard and non-standard English in the NC is one which recognises the importance and value of diversity, as well as the need for children to have an appropriate command of standard English. At KS1, for instance, children are 'introduced to some of the main features of spoken standard English and … taught to use them' (DfEE 1999: 18), but also taught how context and audience correlate with variation in English. This dual approach develops through the key stages (including explicit teaching of the notions of standard and dialect at KS2) until KS4, where the list of expected levels of knowledge about variation in English is impressive:

- why standard English is important, and its place in a national and international context;
- the range of influences on the development of written and spoken language;
- attitudes to usage;
- the main linguistic differences between speech and writing;
- some central lexical and morphosyntactic differences between standard and non-standard dialects;
- some key issues in historical linguistics (lexical innovations, etymology), including the use of English in an electronic age.

The NC in England and Wales provides an example of a rational and informed policy regarding the importance of both standard and non-standard varieties. Despite this, there still remains some widespread suspicion and ignorance regarding the importance of non-standard English in the classroom. The next section is concerned with what we know about variation in English as it relates to education.

12.3 What we know

In order to explore issues relating to standard and non-standard English, it is important to understand what we know about 'Standard English' itself; and while linguists know a great deal about the process of standardisation (i.e. how a particular variety comes to function as the standard language for a community), an exact description of standard English is in fact a rather complicated matter.

12.3.1 Standard English

First, we can differentiate between formal and informal standard English. There are informal standard varieties (which typically are spoken, and are associated with regional norms) and there are formal standard varieties (which typically are written, and are associated with more global norms). For instance, it is common in spoken standard English in the northern part of England to say *I've not* rather than *I haven't* (while the reverse is true in southern England). The norms governing the distribution of variant linguistic forms in these varieties are subjective and arbitrary; the more highly codified the form, the more resistant it will be to change, and the more likely it will be to appear as a feature of the more formal standard variety.

The classic example here is the absence of multiple negation, or as Barber (1993) more appropriately describes it, cumulative negation, in formal standard English. Cumulative negation is the use of many negative markers in one clause, for emphasis (as in *He ain't never done nothing*). It is well attested in the history of the language, but fell into disuse in the standard variety following the codification (the establishment of particular, arbitrary rules) of the standard language in the late Modern period (from the eighteenth century on). As is well known, the practice survives in most non-standard varieties. There is no inherent reason as to why variation between *I've not* and *I haven't* is tolerated as different regional standards, while variation between *He hasn't done anything* and *He ain't never done nothing* is a matter of standard vs. non-standard. The difference is one of convention.

Even allowing for a stylistic range within the standard, it is nonetheless the case that informal standard English is not the norm for the vast majority of English-speaking children; yet as Adger, Wolfram and Christian (2007: 14) observe, it is the speakers of informal standard English, not speakers of highly local vernaculars, whose language determines the shape of the formal standard variety.[3] Some children speak informal standard English more frequently, in a greater range of contexts, with a greater range of interlocutors than others, but for many children, the standard English of the school has some differences of grammar or vocabulary in comparison with the language they use outwith

the classroom. All children have to learn formal standard English for writing, but some are at an advantage because it resembles their spoken language more closely; by contrast, not all children have to learn the informal standard English of the classroom, because some already speak it at home.

This discrepancy has led to discussion about the best way to deal with non-standard varieties in the classroom. An understandable concern – on the part of educators, parents, and employers – is that children leave school without a sufficient command of standard English, such that in formal writing and speaking, they continue to use the non-standard forms associated with informal conversation. This concern conflates a number of issues:

- While it is well established that non-standard forms are more frequent in informal styles, it is also the case that, even in informal styles, speakers from higher social classes use non-standard forms less frequently than their counterparts in the lower social classes. Since standard spoken English is largely defined by middle-class norms, this anxiety over the use of non-standard language in formal contexts is predominantly an issue for children from working-class families, as suggested above.
- Research on the extent of morphosyntactic variation in formal contexts suggested a low incidence of use of non-standard forms in the speech of male and female children aged eleven or fifteen from England. Hudson and Holmes (1995) reported on an analysis of recordings (made in 1988) of 350 children from the following four areas: Merseyside (the area around Liverpool in the north-west); Tyneside (the area around Newcastle in the north-east); Devon and Cornwall in the south-west; and London in the south-east. The report notes that close to a third (32 per cent) did not use a single non-standard variant in the time they were recorded (most informants spoke for between five and ten minutes in total). Fifteen-year-olds used a higher proportion of non-standard forms than did the eleven-year-olds. In one sense this is surprising, because we might expect fewer non-standard features from children who had been exposed to the standard variety for longer (because they had been in the educational system for longer). However, quantitative linguistic research on language change has shown that adolescents are often at the forefront of linguistic innovations (Kerswill 1996), so it is important to discover the extent to which the non-standard features reflect ongoing change in the local vernacular.
- Research on the use of non-standard morphosyntax and lexis in formal conversation suggests that the frequency of local and supralocal non-standard forms is such speech styles is rather low. In a small study conducted in the Tyneside region of England, Crinson and Williamson (2004) report that, among fifteen-year-olds of both sexes from both middle- and working-class backgrounds, and of varying levels of ability in English, very few

non-standard forms were elicited in a series of formal interviews lasting between forty-five minutes and an hour. While some local forms (e.g. *divvent* and *deynt* as negative variants of *do*) and supralocal forms (e.g. syncretism of past tense and past participle forms of *come*) did appear in the corpus, there was considerable idiolectal variation, with some children (including those from a more working-class background) using between zero and two non-standard forms in the entire interview. While there was a greater incidence of non-standard lexical items (again, some highly localised, and some more widespread) these too showed significant idiolectal variation, though there was a greater frequency of local dialect words in the speech of the children from the school with a more working-class catchment area.

- Research in communities in which both standard English and vernaculars are used as the medium of education (see Siegel 1999) has demonstrated that acquisition of and competence in standard English is not compromised in multilingual/bidialectal classrooms.
- Some research in the United States into the use of African American English as an educational tool, and as a linguistic system worthy of investigation, has suggested that exploring differences between varieties may even help young people become more proficient readers (Rickford 2002); similar results for writing were found in a study involving the contrastive approach to the linguistic systems of Kriol and standard English in Belize (Decker 2000, cited in Siegel 2007). As Siegel (2007: 74) observes, 'one reason for the success of awareness programmes with a contrastive component is that they help students separate the vernacular from the standard, no matter how similar or different they are'. The issue of comprehensibility is an important one, but Lippi-Green (1997) has shown that standard English is more understandable to non-standard dialect speakers than non-standard English is to speakers of the standard variety. To a certain extent this is unsurprising given the prevalence of standard English in mainstream media, though increasingly, non-standard English can frequently be found in the new media: the use of non-standard English on the Internet, and the rise of text language, in the context of the global spread of English (Crystal 2001, 2003, 2008), is an interesting development in terms of the emergence of a widespread non-standard orthography, running counter to the typical standardisation process.

A final issue here concerns some of the findings from recent sociolinguistic studies, concerning real-time differences in adolescent speech. Moore (2003) reports evidence from research into adolescent language in the north-west of England. One group of girls (known as the 'Townies') displayed different patterns of variation with respect to some grammatical variables than did another group (the 'Populars'). Over the course of the year in which this ethnographic study took place, the Townie girls (who adhered less to the norms of the school

than did the Populars) showed an increase in the use of some non-standard forms, but only with some variables. For instance, with regard to the use of the past tense of *be* (e.g. standard *he was nice* ~ non-standard *he were nice*), the difference in use of the non-standard form was significantly greater in year 10 than it was in year 9. This divergence contrasted with the convergence between the two groups over the course of the year with respect to a variable known as Right Dislocation (e.g. use of constructions like *He's really funny, your brother*). Notice that it was the more salient and stigmatised variable that distinguished the two groups as they aged: use of forms like *he were nice* is more highly localised, and more often subject to overt correction. This is relevant to education because it suggests that some variables are used differently at different times by children to signal aspects of their identity: this may surface in classroom discourse and written work, and teachers need to be aware of this kind of research, which may explain a perhaps unexpected rise in the use of non-standard variants at a late stage in compulsory education. (For a similar study in a primary school setting, see Snell 2008.)

12.3.2 Switching and crossing

Some research into variation in English and education has explored the linguistic behaviour of children when they are in school and when they are elsewhere, to determine aspects of stylistic variation. This provides interesting data on the frequency and nature of shifting between varieties in different discourse contexts, as well as the nature of shifting within a single variation space. Cheshire (1982), in her study of adolescents in Reading, England, showed that some linguistic forms appeared to be invariant with regard to educational context: *ain't* and syncretised past tense/participle forms like *come* and *done* were reported to be invariant for some of the informants, who used such forms all the time, both in the classroom and in the recordings outside the school. However, this pattern was rare: variation was the norm, with children using standard forms in the school context (without explicit instruction so to do, suggesting that the association between standard language and the school is developed subconsciously).

Switching is not exclusive to students in the classroom. Observations by Wolfram and Adger (reported in Adger, Wolfram and Christian 2007) in Baltimore, and my own observations of teachers in Scotland, have suggested that teachers themselves may use varieties other than standard English. In the Scottish cases, switches on the part of teachers typically involved interaction with students either in small-group work (particularly when discussing issues related to Scotland), or in direct response (or accommodation) to a student who was using a high proportion of Scots[4] variants. Accommodation (see further Trudgill 1986: 1–38) is a regularly occurring feature of normal spoken

interaction, whereby speakers alter their speech to sound more like that of their addressee. This kind of attunement is typical when speakers aim to engender favourable social relations with their audience. Specifically, the accommodation was phonological or lexical, rather than morphosyntactic. In a similar way to that reported by Wolfram and Adger, switching on the part of the teacher was not common during periods of explicit instruction, or when an issue relating to literacy or spoken linguistic form was being discussed.

Shifts may take place between dialects or across languages. This pattern of crossing (Rampton 2005, 2006), in which speakers often mimic other dialects or languages that they hear around them in their community, is regularly found in educational contexts (that is, it is a feature of the classroom as much as it is a feature of the playground). Given the normative ethos associated with school, it is perhaps to be expected that one kind of crossing occurs when speakers of languages other than English are required to use (standard) English in the classroom. Yet it is also the case that crossing away from standard English – even in classroom settings – is not uncommon. A motivation for this can be when the target variety carries a covert prestige within the community: although standard English is valued in the wider world, at local levels, it is often non-standard varieties that are held in high esteem, such that white children may adopt features of African American English, or Panjabi, to show allegiance to others with whom they identify. Such instances of crossing tend to be short-lived (i.e. speakers may only use a couple of phrases from the target variety), but may occur frequently. So another thing we know is that accommodation is widespread, frequent and may take as its target both standard and non-standard English, as well as other languages. Using patterns of shifting and crossing as a resource can enable students to come to a better understanding of how language works, and of why accommodation to the standard may be subconscious in particular discourse settings.

12.3.3 Vernaculars in the classroom

Given such patterns of switching and crossing in educational contexts, the question arises as to how a teacher should 'manage' non-standard forms. As noted in §12.1, some national curricula adopt an inclusive approach, by encouraging discussion of the form and function of variation in the classroom (this is the case in Scotland's 5–14 *Guidelines on English Language*, as it is in the English section of the National Curriculum in England and Wales, for instance). This bidialectal perspective (Trudgill 1975) contrasts with two alternatives: eradication of the non-standard variety, such that only the standard variety is tolerated, and vernacular medium education, in which varieties other than standard English are not only tolerated in schools, they are the medium of instruction. In the United States, where there is no national/federal curriculum as such, the

bidialectal approach is most typically the one adopted (see Adger, Wolfram and Christian 2007: 21–2 for further discussion).

Reaction to publications in non-standard varieties (e.g. readers which were written in African American English (AAE) for use in some classrooms in the United States) has been mixed: some have seen the publication of such texts as educationally beneficial (not just in terms of helping the child to read, but also by virtue of legitimising the non-standard variety), while others have considered this counterproductive for a child who is acquiring the standard variety at school (see Rickford and Rickford 1995 for further discussion of dialect readers for speakers of AAE). The issue as to whether or not a variety other than standard English should be the medium of instruction is a highly contentious one. Siegel (2007) observes that some communities have successfully used creoles as the medium of instruction to encourage acquisition of initial literacy, with standard English being introduced at a later date. Others have respected the local variety, and incorporated it into classroom teaching, but without excluding standard English, which remains the medium of instruction. Yet others have simply tried to raise awareness of the non-standard varieties that children are likely to hear in the local community. (This last is more typical in communities where divergence between standard and non-standard is not as great.)

Siegel's research highlights the range of different issues in different classrooms, suggesting that a 'one size fits all' approach will not work: in some ways it is difficult to make generalisations regarding some aspects of variation in English in an educational context, because those educational contexts themselves vary significantly. Siegel (2007: 67), in a discussion of creoles and minority dialects in education observes that obstacles to student progress could be overcome if teachers recognised creoles and minority dialects as legitimate forms of language, if children were allowed to use their own language to express themselves until they had learned the standard, and if they learned to read in a more familiar language or dialect. However, the extent to which non-standard varieties are recognised in local and national curricula clearly varies across the globe. Attempts to generalise across different educational systems certainly do show that any set of policies and procedures would need to be tailored to fit the community in which the education of the child is taking place. This of course maps on to the nature of the linguistic variation in the community in which the child is being educated.

The diversity of languages in the classroom also has an impact on assessment methods, which may rely implicitly on middle-class norms, so that tests are biased in favour of middle-class children. One example of this concerns narrative. The ability to create a well-crafted written narrative is highly valued in terms of assessment criteria – and children are exposed most frequently to this in instances where they hear typical tales for children (e.g. when they are

read a story by their parents or carers). By contrast, children whose exposure to narrative is more verbal – because on-the-fly storytelling is more common as a means of parent–child interaction, or because children's peer-groups value oral narrative skills more highly – do not get the chance of credit for their equally well-developed command of narrative structure, because those kinds of linguistic skills do not typically feature in language assessment.

There are divided opinions regarding the best way forward for dealing with linguistic diversity in the classroom; these opinions tend to become polarised in educational contexts where one group of students (those who use non-standard variants more frequently, and in more diverse contexts) have lower educational attainment levels in subjects like reading and writing either than that of others of a similar age who use standard English more regularly, or than that of an established national average. Siegel (2007) illustrates this using the examples of Hawaiian Creole and standard English in relation to educational attainment in Hawaii; similarly, Green (2002) reports on the reading standards of fourth-grade white Americans and African Americans. One thing remains clear, however: constant 'correction' of non-standard forms is at best redundant and at worst damaging (Green 2002).

One final issue allied to this is the deficit argument (see Edwards 2006 for a discussion of educational failure, and the deficit/difference debate), which states that children who come to school speaking a non-standard variety have some sort of cognitive weakness that must be repaired, ideally by assimilation to the standard norms of the school. Proponents of the deficit position argue that to encourage non-standard varieties in the classroom is a disservice to children who use those varieties. However, there is no evidence to suggest that speakers of a non-standard dialect are cognitively impaired: the fact that such speakers may do less well in some assessment may be a product of an inherent bias in the construction of the assessment, namely that it is oriented towards those who speak informal standard English in a wide range of discourse contexts.

In sum, the things we know about variation and education are extensive, and include the following:

- we know that multivarietal classrooms are the norm, and that this wealth of varieties can function as a useful resource;
- we know that teaching about standard and non-standard varieties (in terms of both form and function) is more likely to promote greater competence in the standard variety than is the case when the non-standard dialect is marginalised in the classroom;
- we know that failure to respect linguistic diversity can limit educational attainment, but also that explicit teaching about (a) language structure and (b) the specific forms of standard English is sometimes necessary;

- we know that even without explicit instruction, children can accommodate to the normative, standardised language of the school, in speech and (particularly) in writing;
- we know that different teachers will have different issues to deal with when it comes to assisting students in their acquisition of spoken and written standard English. This means that some policies and procedures are more likely to be effective in some communities than in others.

12.4 What we don't know (and how we might find out)

Educational policy reform regarding knowledge about language means that some things which were said to be the case a number of years ago need to be reviewed. For instance, the research by Hudson and Holmes (1995) discussed in §12.2 above was carried out before the introduction of the National Curriculum in England and Wales. A legitimate question to ask, then, is the extent to which the reforms to the curriculum for English – particularly, the reduction in prescriptivist views, the encouragement of a recognition of diversity, and the increase in awareness of the importance of knowledge about language – have brought about changes in the frequency of use of non-standard forms in particular discourse context. Similarly, a follow-up study of the *Survey of British Dialect Grammar* (see Cheshire and Edwards 1989) would achieve two distinct but related aims: it would give us a picture of the extent to which non-standard morphosyntactic forms are emerging, surviving and dying out in a range of communities across England; and it would function as a good resource for students and teachers who wanted to explore the nature of linguistic variation in the local area. These and other suggestions for further research to clarify how things may have changed are discussed below. But first, more general issues concerning linguistic variation in the classroom, which still need to be clarified, are introduced.

12.4.1 The role of factors other than local dialect

There is no doubt that a child's idiolect will influence her ability to acquire competence in standard spoken and written English. But it is important to stress that the appearance of non-standard forms in a child's speech or writing may be the result of a number of different factors. As children grow in confidence as writers, they become more intrepid, as they do in other aspects of their lives; but the more complex the grammatical structure, the greater the risk of the appearance of a feature which is not standard, whether this is associated with their local dialect or not. It is important, therefore, to distinguish between a non-standard pattern (e.g. *that was the man as did it*) and an error (e.g. *I have coming to see my sister*). Some evidence suggests that

dialect variation accounts for very few of the instances of non-standard written English in children's work. Williamson (1990), drawing on research carried out in Tyneside, points out that much of the 'non-standard' nature of children's writing is not really to do with (lack of) knowledge of standard morphosyntactic structures – rather, children typically use non-conventionalised spellings and punctuations.[5] Of the non-standard grammatical features which did appear, cases associated with morphosyntactic variation in the local dialect were less frequent: more often they were the result of more general problems in writing. For example, Williamson (1990: 258) notes that one piece of factual writing began 'We have a problem because we can't move the model', where a noun phrase with a definite article is used, despite no prior mention of the item so determined. His later study (Williamson 1995) suggested that as children age, the proportion of non-standard features decreases, as one might expect. Some of Williamson's evidence suggests that part of the difficulty these children had with writing lies not with the mastery of the distinction between standard and non-standard morphosyntax, but with the intricacies of the conventions of written language. Furthermore, many of the features associated with 'bad' writing have no real local character. The infamous dangling participle, as in *walking down the street, a piano fell on me*, which may be a feature that writers are advised to avoid in formal discourse, is not perceived as a dialect form, in the sense of being localised to a particular regional area. So some of the features of writing that might be corrected by a teacher have very little (if anything) to do with the influence of conventions of the local variety.

A similar issue is of relevance in assessing reading. It is sometimes difficult to distinguish a genuine reading error – 'the selection of the wrong word in a printed text, that is, not the word intended by the writer of that text' (Labov and Baker 2003) – from a case of dialect influence. This can be particularly noticeable in cases of morphophonological variation where the accent variation correlates with a particular grammatical marker, as in the presence or absence of the final consonant in past-tense forms like *rued* and *missed* (cf. *rude* and *mist*, where the final consonant does not have any grammatical function). Labov and Baker's research was based on data collected from 579 children who had difficulty reading, and who were educated in inner-city schools in California and Philadelphia. The study focuses on the relationship between the phonological decoding of words and the overall comprehension of a text (particularly, the ways in which phonological variation may be legitimately classified as an error, such that it casts a 'semantic shadow' over remaining parts of the text, leading to miscomprehension). The data suggested that particular variables need explicit discussion in the classroom: 'Better understanding of the possessive, the copula, and irregular past tense are important for all struggling readers, and direct instruction on the decoding of these signals should lead to a significant advance in reading levels' (Labov and Baker 2003); what remains unclear is the nature

of the 'semantic shadow' which Labov and Baker identify, and which seems to affect comprehension. Furthermore, the results of the study suggested that different variables patterned differently with specific ethnic groups, such that for some groups the variation indicated an error, while for others, it indicated dialect influence. For instance, the speech of some of the children varied between presence and absence of a plural marker (i.e. the *–s* inflection on *cups*): for the African American group, this signalled simple dialect influence (because there was no sense in which this significantly affected comprehension), yet for the Latino group, it signalled a genuine error. Comparisons between patterns in ordinary speech and patterns when reading aloud displayed greater complexity. Some children had a high proportion of non-standard verb concord (e.g. *he stay in bed*) in normal speech; for this group, a high incidence of the non-standard form when reading aloud did not correlate with problems of decoding later parts of the text. Similarly, some children had a high proportion of non-standard possessive marking (e.g. *Jim coat*) in normal speech; for this group, the higher the incidence of the non-standard feature when reading aloud, the more likely there were to be further errors. Research of this kind suggests that there is still some way to go in clarifying the precise role of linguistic variation in reading and writing skills.

12.4.2 Conventionalisation and standardisation

In §12.3.1, I discussed some of the ways in which standard English itself may be seen as a collection of varieties; a consequence of this is that what constitutes the standard may be subject to change. Because of the codification of its form, and its particular functional role, standard English is more resistant to change than many other varieties. But change nonetheless does take place. Like many cases of variation, the incoming forms may be subject to overt comment and stigmatisation, so the progress of the change may be halted, or they may gain greater currency, such that the earlier variant becomes marginalised as a feature of an 'archaic' standard. Particularly, it seems that there may be two different sources of non-standard English: the more widely recognised one, in which variants appear which have their roots in local, informal discourse; and the less-recognised one, in which variants appear which have their roots in highly formal discourse (i.e. they appear as a result of hypercorrection). This can be illustrated by variation in reflexive pronouns.

It is well established that some non-standard speakers and writers of English use some pronouns in a way that displays a difference in syntactic distribution of the reflexive and non-reflexive forms, using the non-reflexive in contexts where a reflexive would be expected, as in (2) below:

(1) I'll get myself some food
(2) %I'll get me some food[6]

It is also the case, however, that some speakers have a different distribution, using a reflexive where a non-reflexive would be expected, as in the examples below:

(3) That's entirely up to yourselves
(4) If myself or any other member of the on-board team can be of service, please speak to one of us as we move through the cabin

Such logophors (where the reflexive is not bound by its antecedent within the appropriate (local, clausal) domain) have featured in a number of discussions about specific issues in linguistic theory, because they raise interesting questions about some of the principles of the theory. But they also raise an interesting question about the nature of the standard, and the nature of the arbitrary conventions that determine what is and is not acceptable. The arbitrariness of certain conventions can also be seen in the case marking of pronouns in co-ordinated and non-co-ordinated noun phrases. Many prescriptivists argue that a phrase like *between you and I* is wrong because the case of the second pronoun should be objective, since it is governed by the preposition. As Huddleston and Pullum (2005: 107) observe, there is no reason why conventions applying to pronouns singly must be the same as those applying to conjoined pronouns; and as Denison (1998) observes, the growth in frequency of this and similar forms since the late Modern English period suggests that speakers have conventionalised this as a standard English pattern.

The issue of standardisation as an instance of language change is a critical one. As English has developed into a global language, new standards have emerged in different communities. There is therefore a constant tension between diversification and conformity, the desire to create a separate identity along with the desire to belong to a larger group. As a result, despite the widespread consensus outlined in §12.3, it becomes rather difficult to define precisely what constitutes 'Standard English': as far as the spoken language is concerned, identifying the features of what unites the grammar of the global standard English, while simultaneously distinguishing that variety from all of the non-standard Englishes, is a complex task, so the standard becomes a (slowly) shifting target. The more local the focus, the more specific we can be – we can say some things about features of standard Singaporean English which differentiate that variety from standard Scottish English. And as noted elsewhere in this chapter, we can be more successful in determining the conventions of the written language than the spoken language. But it is clearly worth noting the observation made by Adger, Wolfram and Christian (2007: 15) regarding the multifaceted nature of standard English: it is 'a collection of the socially preferred dialects from various parts of … English-speaking countries'.

12.4.3 Exploring the unknown

Siegel (2007) observes that despite extensive research on linguistic variation, and on linguistic inequality in educational contexts, much of the new research on the nature and systematicity of non-standard dialects and creole languages has not made its way into the classroom, partly because academic linguists have not engaged with the educational system to the extent that they should. While this may be true generally, there are some notable exceptions, which I mention below. Nonetheless, the failure of much academic research to find its way into the English/literacy classroom is worrying; and this state of affairs is particularly troubling given the widespread recognition – among academics, politicians, employers, and the general public – of the importance of developing children's language skills. That children who speak non-standard varieties should continue to view their own language in a negative light is cause for even greater concern (see Siegel 1999). Siegel (2007: 80) suggests a number of ways in which dissemination of such research might be achieved: involvement in workshops for teachers, writing articles that appear in publications read by teachers, and so on. Some instances of this, and suggestions for future work, are provided below.

Writing for teachers and students on the subject of variation in English can take a number of forms. Hudson (1992), for instance, is concerned primarily with more general issues in grammar teaching for the NC in England and Wales; but as part of this topic there is rightly a discussion of non-standard dialects (what they are, why they matter, and how they might be thought of as a resource); Gordon, Hervey, Leitch and Holstein (1996) is another book on knowledge about language for teachers of English, written specifically for those based in New Zealand.

It is common practice among sociolinguists working on less well-documented or endangered languages to 'give something back' to the community from which they collect their data, through the creation of a grammar or some other linguistic resource. This dual effect – the production of a resource for the community, and the collection of a corpus of use to pure academic research – is also of relevance to variation and education. The *Survey of British Dialect Grammar* (Cheshire and Edwards 1989) achieved a number of objectives relating to both (a) our understanding of the distribution of morphosyntactic variants among young children in the United Kingdom and (b) bringing sociolinguistics into the classroom, and encouraging teachers to see knowledge about language – and particularly, knowledge about grammatical variation in English – as a topic of interest in its own right.

Such writing may involve the development of curricular materials on variation in English, which is important for all levels of school education; and once again, we see evidence of good practice in a number of different communities.

As part of the pilot study work for the *Survey of British Dialect Grammar*, some schemes of work on language variation in contemporary Britain were designed by the principal investigators, to complement teachers' existing knowledge of the subject and to prepare them for the specific research project on morpho-syntactic variation in local dialects (Cheshire and Edwards 1989; Cheshire 2005). Similarly, work in the United States has involved the development and implementation of particular dialect awareness curricula, for example in North Carolina and Baltimore (see Adger, Wolfram and Christian 2007: 151–86 for a detailed account). The topics covered issues in language variation and change, using a variety of media, but critically treating the students as junior researchers, with a strong emphasis on encouraging observations and analysis of language in use in the local community. Furthermore, many areas of linguistic enquiry were incorporated into the curricular materials, and these materials could easily be adapted for work in different English-speaking communities.

Academics working directly with teachers in the classroom can also be a useful way of disseminating research and improving skills in knowledge about language. For instance, a project based at the University of Edinburgh has begun to develop links with specific schools in order to work on particular projects. As part of a school project on names and identity, a linguist was invited to teach a class of thirteen- and fourteen-year-olds on the linguistics of names, with a focus on variation. Some of the topics covered were:

1. The sociolinguistics of names: using nicknames to teach about the linguistic variable (since nicknames are alternative ways of referring to the same thing, and the form chosen may correlate with the social context of the speech act, the relationship between speaker and addressee, and so on);
2. The syntax of names: using naming practices in different languages to show how the order of given and family name may vary, but has a conventionalised structure in all languages;
3. The history of names: using different names to explore how original meanings can become lost in the conventionalisation of names;
4. The typology of names: using naming strategies to illustrate similarities and differences between languages (for example, patronymics and occupation-based names in Arabic and Scots).

As well as introducing some unfamiliar data from earlier stages of English, and from other languages, much of this work used media that the students were familiar with (such as films and novels), including that most valuable of resources, their own use of language. The teacher was able to use this material in follow-up classes, to produce a scheme of work that could be replicated elsewhere, and build towards a national assessment unit.[7]

Finally, a further way of increasing awareness and understanding of linguistic variation in education concerns academic involvement in projects outwith

institutes of higher education. A number of television programmes on language (*Do you speak American?* made in the United States, and *The Story of English* made in the United Kingdom, for instance) have involved significant academic input, reflecting public interest in the subject of linguistic variation. Such productions highlight another area of importance for variation and education – the issue of lifelong learning. The focus in this chapter has been very much on school education, but the issue of adult learning is also relevant, and many projects which have had an academic input have been geared towards recognising interests held by the general public on matters of linguistic diversity. Two current examples from the United Kingdom – the BBC *Voices* project and the British Library's *Sounds Familiar* project – illustrate nicely how projects developed for the general public can nonetheless be of particular interest and relevance for schools. The *Sounds Familiar* project (www.bl.uk/learning/langlit/sounds/index.html) provides an excellent set of interactive resources for work on varieties of British English, and again encourages children to become involved in the project by submitting recordings of their own voices, potentially to be analysed and uploaded on the website.[8] All the material is freely available via the British Library website, and while perhaps part of this is designed with schoolchildren in mind, it is clear that this, and similar pages on the British Library website (e.g. *Changing Language*), will be of interest to anyone (irrespective of their age) wishing to learn more about variation in English. Similar projects, involving museum exhibits and community-based presentations in the United States, are discussed by Wolfram (1999). Projects such as these illustrate the relevance of linguistic variation to opportunities for lifelong learning.

12.5 Why it matters

A greater understanding of linguistic diversity in the classroom matters for a number of reasons. It matters because a lack of knowledge – on the part of the teacher and educational policy maker – about the nature and function of linguistic variation can be disruptive and damaging (leading in some cases to significant educational failure on the part of the non-standard dialect speaker). It matters because any sort of marginalisation based on language difference is unnecessary and counterproductive. It matters because a celebration of different Englishes in the classroom gives all students an insight into local culture (whether they be indigenous or not). Above all, it matters because the different varieties of English which children command are not only a part of a community's identity, they are also part of individual identity – and as such, should be celebrated and enjoyed.

This last point is crucial. Linguistic variation in the classroom could be seen as an opportunity, rather than something to be eliminated, and there is evidence

from a range of communities (see Siegel 2007: 69 for further detail) where particular resources have been produced with the specific aim of educating young people about (not in) the vernacular of their local community. It is also important to stress that studying varieties of English is relevant for education not simply in terms of a greater understanding of how language works: such study also relates to matters of citizenship, understanding and overcoming prejudice, and appreciation of cultural diversity.

By focusing on language variation, on both the systemic and the functional differences between different dialects (or between different languages, in some cases), we can better educate not just about the standard variety – including of course, a discussion of why the standard variety matters – but also about the specific linguistic features which differentiate the standard from other varieties spoken in the classroom. This then becomes a larger issue than simply one concerning the development of appropriate resources, or the creation of a stimulating educational environment for the child. It becomes about finding alternative ways to teach the standard variety while simultaneously exploring the diversity of linguistic forms in the local community, something which in itself is surely an important part of the educational process. These are matters for teachers, who are the experts in knowing what will and will not work in a classroom setting. So collaboration between academic linguists and teachers is vital. It is clear that the linguistic diversity of the classroom has the potential to be a tremendous resource for both the linguist interested in patterns of variation and the teacher interested in developing pupils' skills and knowledge about language. Projects which explore this diversity can further our understanding of (a) the general nature of linguistic variation, (b) the specific distribution of variants in dialects of English, and (c) the development of reading, writing, and speaking skills in children.

12.6 Where next?

This chapter has considered some aspects of the relationship between variation and education (more from the perspective of educational linguistics than from the perspective of variationist linguistics). Specifically, I have argued that knowledge about variation is an important part of knowledge about language, and linguists and educationalists need to work together to produce functional resources that meet particular curricular requirements, as well as establishing what those curricular requirements might be. These issues are also of relevance for teacher training, especially in situations where a teacher's knowledge of the nature of linguistic variation might be more limited than it should be.

For recent books specifically on linguistic variation in education, Adger, Wolfram and Christian (2007) is extremely valuable. More general issues concerning linguistics in the classroom are covered in two books edited by Kristin

Denham and Anne Lobeck (Denham and Lobeck 2005, 2009), the second of which takes a global perspective on language in schools. Interesting comparisons between different European countries in terms of policy and practice regarding standard and non-standard languages in the classroom is addressed in Cheshire, Edwards, Münstermann and Weltens (1989).

A handbook on educational linguistics (Spolsky and Hult 2008) includes sections on dialect variation: see particularly Reaser and Adger (2008), and King and Benson (2008). Another excellent handbook chapter on English dialects and education is Cheshire (2005). A helpful discussion of recent changes in practice (and an account of how some things have unfortunately stayed the same) is provided in two articles by Siegel (1999, 2007). Journals which regularly publish on matters of variation and education in English include *Language and Education* and *Linguistics and Education*. The entries in the section on educational linguistics in the *Encyclopedia of Language and Linguistics* (Brown 2006) are also very useful.

Some websites have particularly useful information for specific countries. Richard Hudson's website (www.phon.ucl.ac.uk/home/dick/education.htm) is invaluable for academics and teachers in England and Wales who want to know more about educational linguistics. Resources (developed jointly by linguists and educationists) for work on American English in schools are available at: www.pbs.org/speak/education/; professional development material for teachers associated with such work is available at: www.pbs.org/speak/education/training/.

There are many possible research avenues which could be developed in this area. An updated version of the *Survey of British Dialect Grammar* for the UK would be very welcome, for instance; and similar surveys could be created and carried out in other parts of the world. It would be particularly useful to know the extent to which innovations in local dialects feature heavily in children's writing, or whether the non-standard grammatical patterns are instances of stable variables. (The suggestion from existing research seems to be that it is the latter, but this is not always made explicit.) In addition, further work on the effects of contact – both language contact and dialect contact – on use of non-standard language in educational settings would be welcome, especially in lesser-studied dialect areas.

Notes

1 COLLECTING DATA ON PHONOLOGY

1. Keywords used mostly follow the system developed by Wells (1982).
2. The sound recordings on which Figures 1.2 and 1.3b are based were provided by Warren Maguire.

2 HOW TO MAKE INTUITIONS SUCCEED: TESTING METHODS FOR ANALYSING SYNTACTIC MICROVARIATION

1. A linguistic variable is a unit comprising at least two 'alternate ways of saying the same thing' (Labov 1972b: 118), which are thus conventionally termed 'linguistic variants'. Syntactic variables are here more loosely defined following Spruit (2006: 494) as 'a form or word order in a syntactic context in which two dialects can differ'.
2. We will not discuss the SuRE method, which was developed mainly for the collection of lexical variability but which certainly has the potential to be adapted for collecting data at other levels of linguistic structure (Kerswill, Llamas and Upton 1999).
3. In our pilot we opted for a ratio of 1:1 of filler/test sentences (after Schütze 1996).
4. For rather more detail regarding the data collection methodology than is warranted here, please see Buchstaller *et al*. (forthcoming).
5. Cowart (1997: 52) has shown that the best effects are achieved by using filler sentences ranging from acceptable to completely ungrammatical in approximately equal numbers.
6. Distinctions such as animate/non-animate as well as the grammatical role of the antecedent can be important features of the variable context (see Tagliamonte *et al*. 2005).
7. This same problematic incongruence is also characteristic of our results for the distribution of acceptability ratings with respect to gender differences in northern England, which also seem to be unduly influenced by the type of test being applied. While space issues preclude a more detailed discussion of our findings here, we refer to Buchstaller and Corrigan (2008) and Buchstaller *et al*. (forthcoming), where we provide more detail about the gender differences yielded across tests in our Northern English study.

3 CORPORA: CAPTURING LANGUAGE IN USE

1. I adopt the terminology of Beal *et al*. (2007a) as convenient labels, but the crucial distinction between the two types relates not to notions of convention (both have

longstanding histories) but rather to the different methodological paradigms driving the construction of corpora in empirical variation studies.
2. The goal for each regional corpus within ICE is 1 million words, 600,000 from spoken texts and 400,000 from written texts (see Nelson 1996), though delimitation of speakers and text types is by no means straightforward (e.g. Holmes 1996 on ICE-NZ, Kallen and Kirk 2007 on ICE-Ireland).
3. GB = Great Britain; AUS = Australia; NZ = New Zealand.
4. There are times when it is necessary to rely on written records; namely, when the area of study extends to periods prior to the availability of audio recordings. The ideal documents for historical sociolinguistic research are those that are intended to represent a speech act, real or imagined (e.g. trial records, amanuensis accounts, letters, diaries, fiction). For discussion see Schneider (2002) and references therein.
5. Some corpora can be purchased for a nominal fee for classroom use (e.g. BNC Baby, a four-million-word subset of the BNC); the full BNC can be searched online for no cost using the interface created by Mark Davies, http://corpus.byu.edu/. This site also provides links to COCA (the Corpus of Contemporary American English) and the TIME Magazine corpus, among others.
6. For descriptions of, and accessibility details for, other public specialised corpora, see many of the contributions in Beal *et al*. (2007a).
7. Before that time, researchers can access CONTE-pC by individual request.
8. Summaries of the sixty discourse segments in the SBC can be found at www.linguistics.ucsb.edu/research/sbcorpus_summaries3.html.
9. While it is possible, in principle, for a corpus of spoken data to be available electronically but to not yet be transcribed, working directly from sound files presents an exception rather than the norm in variation research and is not a possibility I will address here.
10. ONZEminer is Open Source software; it can be downloaded and installed free of charge.
11. *Like* has been performing pragmatic functions since at least the nineteenth century. For discussion and apparent time evidence see D'Arcy (2007); Romaine and Lange (1991: 270) also discuss the history of *like* as a vernacular form.

4 HYPOTHESIS GENERATION

1. www.ncl.ac.uk/necte/
2. www.ncl.ac.uk/necte/appendix1.htm

5 QUANTIFYING RELATIONS BETWEEN DIALECTS

1. Minus 1 since each variety does not need to be compared to itself (always 100 per cent similar), and divided by 2 since the similarity of A to B is the same as the similarity of B to A.
2. Note that different versions of Levenshtein DISTANCE assign different costs to these operations; in particular, substitution may be modelled as a deletion followed by an insertion, hence costing 2.
3. Proto-Germanic because the method is designed to compare not only varieties of English but also varieties of other Germanic languages.

6 PERCEPTUAL DIALECTOLOGY

1. In Lambert *et al*.'s study, these 'guises' were different languages, but have subsequently been different accents (e.g. Giles 1977).
2. The composite map is constructed after Preston (1999c: 362), the dialect names are listed in rank order of perception on the left of the map; on the right-hand side, numbers refer to the number of informants indicating areas, and figures in parentheses refer to the overall percentage recognition level of each area. The lines on the map are shaded differently in order to aid differentiation between different dialect areas.
3. The ≥21% agreement level indicates that the least accurate 20% of placements were removed, the n = value indicates how many lines are displayed on the chart, the mean error indicates the overall mean error of placements for 100% of the placements made.
4. There were very few instances of informants simply drawing the 'home' area and no others; it was usually the case that informants who drew detailed maps, including many areas, included the 'home' area.

10 VARIATION AND IDENTITY

1. The Half Moon Bay Style Collective is an international collective of twenty sociolinguists from six universities, organised by Penny Eckert with funding from the Spencer Foundation, and so named after the place of our first meeting. The Collective raised issues and proposed directions for the study of style and encouraged collaborative work between its members.
2. The combination of ethnography and quantitative analysis has only recently re-emerged in variationist research (following the preponderance of demographic surveys in the 1970s and 1980s). However, it should be noted that, whilst ethnography has a disjointed history in quantitative research, it has long had a role in qualitative sociolinguistics. As Bauman and Sherzer (1974: 3) observe, the publication of Hymes' (1962) paper entitled 'the ethnography of speaking' and the subsequent collection of papers edited by Gumperz and Hymes (1964) introduced ethnography to linguistic enquiry.
3. Some people have found it difficult to see the difference between a community of practice and other social aggregates, such as social networks (see, for instance, Davies 2005 and discussions in the same volume). The difference is largely in the methodology. It is not possible to do a community of practice study without ethnography, whereas one could feasibly reconstruct a social network on the basis of a questionnaire or interview about somebody's social ties. An analysis based on the latter can tell us something about social connection, but – on its own – it can't tell us whether the connections are actually meaningful. Ethnography can help to reveal the quality of connections and the social meaning they encode.
4. It should be noted that there has always been much more research on phonological variation than on morphosyntactic variation. However, see Snell (2008) for current and interesting ethnographic research on the latter; see also Buchstaller and Corrigan (this volume).
5. f0 refers to the fundamental frequency of an utterance – a measure which can be used to study the salience of pitch.

12 VARIATION AND EDUCATION

1. The education and public examination system in Scotland is covered by a different set of regulations, while the system in Northern Ireland is also different, though the curriculum is modelled very closely on the National Curriculum, and Northern Irish children take the same kind of public exams as English and Welsh children.
2. Prescriptive approaches to language are usually concerned with guidance as to which linguistic forms are the correct ones to use, and as such contrast with descriptive approaches. Descriptive linguistics provides an account by the author of what people do when they speak and write; prescriptive linguistics provides a set of instructions as to what the author thinks people should do when they speak and write. An example of a prescriptive approach to variation would be something like *When comparing two things, only the comparative form of the adjective should be used, as in 'I have two friends in York, Sam and Ella. Sam is the younger.'* A descriptive approach to the same phenomenon might point out that many speakers also use superlative forms in such cases (e.g. *Sam is the youngest*). Prescriptivist approaches to grammar still loom large in the English classroom in different parts of the world: for example, Horan (2002) reports on a study of grammar teaching in the English classroom in twenty-four schools in Sydney. In answer to the question '*How would you deal with phrases from students such as*: she done it, me and me friends, I don't know nothing*?*', fifty-one of the fifty-two teachers interviewed adopted a prescriptivist approach, viewing such forms as incorrect rather than non-standard.
3. Local norm enforcement, however, may promote the use of non-standard forms, depending on the social fabric of the local community (Milroy 1992). Issues of overt and covert prestige are relevant here too.
4. Scots is an officially recognised minority language of the United Kingdom, a language which developed from Old Northumbrian and which has dialects currently spoken in both Scotland and Ireland.
5. The issues of spelling and punctuation are targets of criticism about 'standards of English'. In his 2008 conference address, David Cameron, the leader of the UK Conservative Party, commenting on debate on spelling reform, said: 'Listen to this. It's the President of the Spelling Society. He said, and I quote, "people should be able to use whichever spelling they prefer." He's the President of the Spelling Society. Well, he's wrong. And that's spelt with a 'w'.' Such orthographic variation is often – and mistakenly – considered as part of 'grammar'; and such comments show how variation and 'standards' in education often become political issues.
6. The % sign here indicates that the sentence is considered grammatical by a limited number of users of a particular language, here English.
7. In this particular case, a teacher from a different local authority used the same basic material that had been prepared but adapted it for a specific scheme of work on local place names.
8. A further series of projects at the British Library, including an exhibition on varieties of English, is planned, and will involve collaboration with academics interested in variation and change in English.

References

Adger, Carolyn Temple, Walt Wolfram and Donna Christian. 2007. *Dialects in Schools and Communities*, 2nd edn. Mahwah, NJ: Lawrence Erlbaum.

Adger, David. 2006. Combinatorial variability. *Journal of Linguistics* 42, 503–30.

2007. Variability and modularity: a response to Hudson. *Journal of Linguistics* 43, 695–700.

Adger, David and Jennifer Smith. 2005. Variation and the Minimalist Programme. In Cornips and Corrigan (eds.), 149–78.

Adger, David and Graeme Trousdale. 2007. Variation in English syntax: theoretical implications. *English Language and Linguistics* 11, 261–78.

Agard, Frederick. 1971. Language and dialect: some tentative postulates. *Linguistics* 65, 5–24.

Agheyisi, Rebecca and Joshua A. Fishman. 1970. Language attitude studies: a brief survey of methodological approaches. *Anthropological Linguistics* 12, 137–57.

Aitken, A. J. 1981. The Scottish vowel-length rule. In Michael Benskin and Michael Samuels (eds.) *So meny People, Longages and Tonges*, 131–57. Edinburgh: The Middle English Dialect Project.

Allen, Harold (ed.). 1976. *The Linguistic Atlas of the Upper Midwest*. Vol. III: *The Pronunciation*. Minneapolis, MN: University of Minnesota Press.

Allen, Will, Joan Beal, Karen Corrigan, Warren Maguire and Hermann Moisl. 2007. A linguistic 'time capsule': The Newcastle Electronic Corpus of Tyneside English. In Beal, Corrigan and Moisl (eds.), Vol. II, 16–48.

Ammon, Ulrich, Norbert Dittmar, Klaus Mattheier and Peter Trudgill (eds.). 2006. *Sociolinguistics: An International Handbook of the Science of Language and Society*, 2nd edn. Berlin: Mouton de Gruyter.

Andersen, Gisle. 1997. 'They gave us these yeah, and they like wanna see like how we talk and all that': the use of *like* and other discourse markers in London teenage speech. In Ulla-Britt Kostinas, Anna-Brita Stenström and Anna-Malin Karlsson (eds.) *Ungdomsspråk i Norden*, MINS 43, 83–95. Stockholm University, Institutionene für nordiska språk.

1998. The pragmatic marker *like* from a Relevance-Theoretic perspective. In Andreas Jucker and Yael Ziv (eds.) *Discourse Markers: Descriptions and Theory*, 147–70. Amsterdam: John Benjamins.

2001. *Pragmatic Markers and Sociolinguistic Variation.* Amsterdam: John Benjamins.

Anderson, Peter M. 1987. *A Structural Atlas of the English Dialects.* London: Croom Helm.

Angermeyer, Philipp S. 2009. Translation style and participant roles in court interpreting. *Journal of Sociolinguistics* 13(1), 3–28.

Anton, Howard. 2005. *Elementary Linear Algebra*, 9th edn. Hoboken, NJ: Wiley International.

Anttila, Arto. 1997. Deriving variation from grammar. In Hinskens, van Hout and Wetzels (eds.), 35–68.

2002a. Morphologically conditioned phonological alternations. *Natural Language and Linguistic Theory* 20, 1–42.

2002b. Variation and phonological theory. In Chambers, Trudgill and Schilling-Estes (eds.), 206–43.

Arseniuk, Melissa. 2008. FBI expert testifies at O. J. Simpson robbery/kidnapping trial. *Las Vegas Sun* 17 September 2008. www.lasvegassun.com/news/2008/sep/17/fbi-expert-testifies-simpson-robberykidnapping-tri/ (last accessed 30 October 2008).

Asprey, Esther C. 2008. *Black Country English and Black Country Identity*. Unpublished Ph.D. dissertation, University of Manchester.

Atal, Bishnu S. and Suzanne L. Hanauer. 1971. Speech analysis and synthesis by linear predictive coding of the speech wave. *Journal of the Acoustical Society of America* 50, 637–55.

Avery, Peter, J. K. Chambers, Alexandra D'Arcy, Elaine Gold and Keren Rice (eds.). 2006. *Canadian English in the Global Context*. Special issue of *Canadian Journal of Linguistics* 51.

Baayen, R. Harald, Richard Piepenbrock and Léon Gulikers. 1995. *The CELEX Lexical Database* (CD-ROM). Philadelphia: Linguistic Data Consortium, University of Pennsylvania.

BBC. 2002. Witness walks out of Damilola trial. Available online at http://news.bbc.co.uk/1/hi/england/1802072.stm, 5th February, 2002 (last accessed 5 August 2010).

BBC. 2008. Rhys witness refuses to testify. Available online at http://news.bbc.co.uk/1/hi/england/merseyside/7695628.stm. Page last updated at 18:07 GMT, Tuesday, 28 October 2008 (last accessed 5 August 2010).

Bailey, Charles-James N. 1973. *Variation and Linguistic Theory*. Arlington, VA: Center for Applied Linguistics.

1980. Conceptualizing dialects as implicational constellations rather than as entities bounded by isoglossic bundles. Chapter 4 of Bailey (1996). Originally in J. Göschel, P. Ivić and K. Kehy (eds.) *Dialekt und Dialektology*. Wiesbaden: Franz Steiner.

1996. *Essays on Time-Based Linguistic Analysis*. Oxford University Press.

Bailey, Guy, Tom Wikle and Lori Sand. 1991. The focus of linguistic innovation in Texas. *English World-Wide* 12, 195–214.

Bailey, Guy, Tom Wikle, Jan Tillery and Lori Sand. 1991. The apparent time construct. *Language Variation and Change* 3, 241–64.

1993. Some patterns of linguistic diffusion. *Language Variation and Change* 5, 359–90.

Barber, Charles. 1993. *The English Language: A Historical Introduction*. Cambridge University Press.

Barbiers, Sjef, Olaf Koeneman, Marika Lekakou and Margreet van der Ham (eds.). 2002. *Syntactic Doubling in European Dialects*. Amsterdam: Elsevier.

Barbujani, Guido. 2000. Geographical patterns: how to identify them, and why. *Human Biology* 72, 133–53.

2005. Human races: classifying people vs understanding diversity. *Current Genomics* 6, 215–26.

Barbujani, Guido, Giorgio Bertorelle and Lounès Chikhi. 1998. Evidence for Paleolithic and Neolithic gene flow in Europe. *American Journal of Human Genetics* 62, 488–91.

Bard, Ellen, Dan Robertson and Antonella Sorace. 1996. Magnitude estimation of linguistic acceptability. *Language* 72, 1–31.

Barnes, Peter. 1984. Orkney and Shetland Norn. In Trudgill (ed.), 352–66.

Baron, Naomi. 2004. See you online: gender issues in college student use of instant messaging. *Journal of Language and Social Psychology* 23, 397–423.

Bauer, Laurie. 1993. *Manual of Information to Accompany the Wellington Corpus of New Zealand English*. Wellington: Department of Linguistics, Victoria University of Wellington.

2002. Inferring variation and change from public corpora. In Chambers, Trudgill and Shilling-Estes (eds.), 97–114.

Baugh, Albert C. and Thomas Cable. 2002. *A History of the English Language*, 5th edn. London: Routledge.

Bauman, Richard and Joel Sherzer (eds.). 1974. *Explorations in the Ethnography of Speaking*. Cambridge University Press.

Bayley, Robert. 1994. Consonant cluster reduction in Tejano English. *Language Variation and Change* 6, 303–27.

Beal, Joan C. 1993. The grammar of Tyneside and Northumbrian English. In James Milroy and Lesley Milroy (eds.) *Real English: Grammar of English Dialects in the British Isles*, 187–213. London: Longman.

2004. The morphology and syntax of English dialects in the north of England. In Kortmann, Schneider, Burridge, Mesthrie and Upton (eds.), 114–41.

Beal, Joan C. and Karen P. Corrigan. 2005. A tale of two dialects: relativization in Newcastle and Sheffield. In Markku Filppula, Juhani Klemola, Palander Marjatta and Esa Penttilä (eds.) *Dialects Across Borders: Selected papers from the 11th International Conference on Methods in Dialectology (Methods XI), Joensuu, August 2002. CILT, 273*, 211–29 Amsterdam: John Benjamins.

Beal, Joan C., Karen P. Corrigan and Hermann L. Moisl (eds.). 2007a. *Creating and Digitizing Language Corpora, Volume I: Synchronic Databases*. Houndmills: Palgrave Macmillan.

Beal, Joan C., Karen P. Corrigan and Hermann L. Moisl (eds.). 2007b. *Creating and Digitizing Language Corpora, Volume II: Diachronic Databases*. Houndmills: Palgrave Macmillan.

Beckman, Mary E. and Julia Hirschberg. 1994. The ToBI annotation conventions. Online typescript. www.ling.ohio-state.edu/~tobi/ame_tobi/annotation_conventions.html (last accessed 1 August 2010).

Bede, the Venerable. 731. *Ecclesiastical History of the English People*, 1990 Penguin edition, translated by Leo Sherley-Price, edited with new introduction and notes by D. H. Farmer. Harmondsworth: Penguin.

Behre, Karl-Ernst. 2007. A new Holocene sea-level curve for the southern North Sea. *Boreas* 36(1), 82–102.

Bell, Allan. 1984. Language style as audience design. *Language in Society* 13, 145–204.

Bellwood, Peter and Colin Renfrew (eds.). 2002. *Examining the Farming/Language Dispersal Hypothesis*. Cambridge: McDonald Institute for Archaeological Research Monograph series.

Bender, Emily. 2007. Socially meaningful syntactic variation in sign-based grammar. *English Language and Linguistics* 11, 347–81.

Benincà, Paola and Cecilia Poletto. 2007. The ASIS enterprise: a view on the construction of a syntactic atlas for the Northern Italian dialects. In Kristine Bentzen and Øystein Vangsnes (eds.) *Scandiavian Dialect Syntax 2005, Nordlyd* 34, 35–52.

Berk-Seligson, Susan. 1990. Bilingual court proceedings: the role of the court interpreter. In Levi and Walker (eds.), 155–201.

1999. The impact of court interpreting on the coerciveness of leading questions. *The International Journal of Speech, Language and the Law* 6(1), 30–56.

2002. *The Bilingual Courtroom: Court Interpreters in the Judicial Process*. University of Chicago Press.

Bernstein, Cynthia. 1993. Measuring social causes of phonological variables. *American Speech* 68, 227–40.

Biber, Douglas. 1988. *Variation across Speech and Writing*. Cambridge University Press.

1993. Representativeness in corpus design. *Literary and Linguistic Computing* 8, 243–57.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad and Edward Finegan. 1999. *Longman Grammar of Spoken and Written English*. London: Longman.

Blake, Renée and Meredith Josey. 2003. The /ay/ diphthong in a Martha's Vineyard community: what can we say 40 years later? *Language in Society* 32(4), 451–85.

Blatchford, Helen and Paul Foulkes. 2006. Identification of voices in shouting. *The International Journal of Speech, Language and the Law* 13(2), 241–54.

Blommaert, Jan. 2001. Investigating narrative inequality: African asylum seekers' stories in Belgium. *Discourse and Society* 12(4), 413–49.

Bloomfield, Leonard. 1944. Secondary and tertiary responses to language. *Language* 20, 44–55.

Blyth, Tom and Edmund Robertson. 2002. *Basic Linear Algebra*, 2nd edn. Heidelberg and New York: Springer.

Boersma, Paul and Bruce Hayes. 2001. Empirical tests of the Gradual Learning Algorithm. *Linguistic Inquiry* 32, 45–86.

Bourdieu, Pierre. 1977. *Outline of a Theory of Practice*, trans. Richard Nice. Cambridge University Press.

Brennan, Mary and Richard Brown. 2004. *Equality Before the Law: Deaf People's Access to Justice*. Durham: Deaf Studies Research Unit, University of Durham.

Bresnan, Joan, Ashwini Deo and Devyani Sharma. 2007. Typology in variation: a probabilistic approach to *be* and *n't* in the Survey of English Dialects. *English Language and Linguistics* 11, 301–46.

Britain, David. 1992. Linguistic change in intonation: the use of high rising terminals in New Zealand English. *Language Variation and Change* 4, 77–104.

1997. Dialect contact and phonological reallocation: 'Canadian raising' in the English Fens. *Language in Society* 26, 15–46.

2002. Space and spatial diffusion. In Chambers, Trudgill and Schilling-Estes (eds.), 603–37.

Bromham, Lindell and David Penny. 2003. The modern molecular clock. *Nature Reviews Genetics* 4, 216–24.

Brown, Keith (ed.). 2006. *Encyclopedia of Language and Linguistics*, 2nd edn. Amsterdam: Elsevier.

Brown-Blake, Celia. 2008. The right to linguistic non-discrimination and Creole language situations: the case of Jamaica. *Journal of Pidgin and Creole Languages* 23(1), 32–73.

Brown-Blake, Celia and Paul Chambers. 2007. The Jamaican creole speaker in the UK Criminal Justice System. *The International Journal of Speech, Language and the Law* 14(2), 269–94.

Bucholtz, Mary. 1999. 'Why be normal?': language and identity practices in a community of nerd girls. *Language in Society* 28, 203–23.

Bucholtz, Mary and Kira Hall. 2005. Identity and interaction: a sociocultural linguistic approach. *Discourse Studies* 7, 585–614.

Buchstaller, Isabelle and Karen P. Corrigan. 2008. Tapping into the intuitions of dialect speakers in Northern England and Scotland: data, methods and their implications. Paper presented at the Edisyn workshop for Syntactic Atlas projects. Venice, September 2008.

Buchstaller, Isabelle, Karen P. Corrigan and Anders Holmberg. Forthcoming. Introspective judgements and sociolinguistic interviews as testing instruments for revealing syntactic microvariation. In Jeff Parrot (ed.) *Theoretical Perspectives on Intra-Individual Variation and its Empirical Study*. Amsterdam/Philadelphia: John Benjamins.

Buchstaller, Isabelle and Alexandra D'Arcy. 2009. Localized globalization: a multilocal, multivariate investigation of quotative *be like*. *Journal of Sociolinguistics* 13, 291–331.

Bunnin, Nicholas and Jiyuan Yu. 2009. *The Blackwell Dictionary of Western Philosophy*. Hoboken, NJ: Wiley Blackwell.

Burbano-Elizondo, Lourdes. 2006. Regional variation and identity in Sunderland. In Tope Omoniyi and Goodith White (eds.) *Sociolinguistics of Identity*, 113–28. London: Continuum.

2008. Language variation and identity in Sunderland. Unpublished Ph.D. thesis, University of Sheffield.

Burridge, Kate and Bernd Kortmann (eds.). 2008. *Varieties of English 3: The Pacific and Australasia*. Berlin/ New York: Mouton de Gruyter.

Butler, Judith. 1990. *Gender Trouble: Feminism and the Subversion of Identity*. New York and London: Routledge.

Butters, Ronald R. 1991. Review of Dennis Preston, *Perceptual Dialectology. Language in Society* 20, 294–9.

Bybee, Joan. 2001. *Phonology and Language Use*. Cambridge University Press.

Byrne, Catherine and Paul Foulkes. 2004. The 'mobile phone effect' on vowel formants. *The International Journal of Speech, Language and the Law* 11(1), 83–102.

Calafell, Francese, Francis Roubinet, Anna Ramirez-Soriano, Naruya Saltou, Jaume Bertranpetit and Antoine Blancher. 2008. Evolutionary dynamics of the human ABO gene. *Human Genetics* 124, 123–35.

Cambier-Langeveld, Tina. 2007. Current methods in forensic speaker identification: results of a collaborative exercise. *The International Journal of Speech, Language and the Law* 14(2), 223–43.

Cameron, Deborah and Don Kulick. 2003. *Language and Sexuality*. Cambridge University Press.

2005. Identity crisis? *Language and Communication* 25, 107–25.

Cameron, Deborah, Fiona McAlinden and Kathy O'Leary. 1989. Lakoff in context: the social and linguistic functions of tag questions. In Jennifer Coates and Deborah Cameron (eds.) *Women in Their Speech Communities: New Perspectives on Language and Sex*, 74–93. London: Longman.

Campbell-Kibler, Kathryn. 2005. Listener perceptions of sociolinguistic variables: the case of (ING). Unpublished Ph.D. dissertation, Stanford University.

2007. Accent, (ING), and the social logic of listener perceptions. *American Speech* 82, 32–64.

Campbell-Kibler, Kathryn, Penelope Eckert, Norma Mendoza-Denton and Emma Moore. 2006. The elements of style. Poster presented at *New Ways of Analyzing Variation* 35, Columbus, Ohio.

Capelli C., N. Redhead, J. K. Abernethy, *et al.* 2003. A Y chromosome census of the British Isles. *Current Biology* 13, 979–84.

Cardi, Valeria. 2007. Regional or minority language use before judicial authorities: provisions and facts. *Journal on Ethnopolitics and Minority Issues in Europe* 6(2), 1–24.

Carr, Philip. 1993. *Phonology*. Basingstoke: Macmillan.

2006. Philosophy of linguistics. In Keith Brown (ed.) *Encyclopedia of Language and Linguistics*, 2nd edn. Amsterdam: Elsevier.

Carter, Phillip M. 2005. Quantifying rhythmic differences between Spanish, English, and Hispanic English. In Randall S. Gess and Edward J. Ruben (eds.) *Theoretical and Experimental Approaches to Romance Linguistics*, 63–75. Amsterdam/ Philadelphia: John Benjamins.

Cavalli-Sforza, Luigi Luca 2000. *Genes, Peoples and Language*, trans. Mark Seielstad. New York: North Point Press.

Cavalli-Sforza, Luigi Luca, Paolo Menozzi and Alberto Piazza. 1994. *The History and Geography of Human Genes*. Princeton University Press.

Cavalli-Sforza, Luigi Luca, A. C. Wilson, C. R. Cantor, R. M. Deegan and M. C. King. 1991. Call for a worldwide survey of human genetic diversity: a vanishing opportunity for the Human Genome Project. *Genomics* 11, 490–1.

Cedergren, Henrietta. 1973. The interplay of social and linguistic factors in Panama. Unpublished dissertation, Cornell University.

Cedergren, Henrietta and David Sankoff. 1974. Variable rules: performance as a statistical reflection of competence. *Language* 50, 333–55.

de Certeau, Michel. 1984. *The Practice of Everyday Life*. Berkeley, CA: University of California Press.

Chalmers, Alan. 1999. *What is this Thing Called Science?* 3rd edn. New York: McGraw-Hill/Open University Press.

Chambers, J. K. 1994. An introduction to dialect topography. *English World-Wide* 15, 35–53.

1998. Social embedding of changes in progress. *Journal of English Linguistics* 26(1), 5–36.

2002. Patterns of variation including change. In Chambers, Trudgill and Schilling-Estes (eds.), 349–72.
2003. *Sociolinguistic Theory: Linguistic Variation and its Social Significance*, 2nd edn. Oxford: Blackwell.
Chambers, J. K. and Peter Trudgill. 1980. *Dialectology*. (2nd edn, 1998). Cambridge University Press.
Chambers, J. K., Peter Trudgill and Natalie Schilling-Estes (eds.). 2002. *The Handbook of Language Variation and Change*. Malden and Oxford: Blackwell.
Chaski, Carole E. 2001. Empirical evaluations of language-based author identification techniques. *Forensic Linguistics* 8(1), 1–65.
Cheshire, Jenny. 1982a. *Variation in an English Dialect: A Sociolinguistic Study*. Cambridge University Press.
1982b. Linguistic variation and social function. In Suzanne Romaine (ed.) *Sociolinguistic Variation in Speech Communities*, 153–75. London: Edward Arnold.
Cheshire, Jenny (ed.). 1991. *English Around the World: Sociolinguistic Perspectives*. Cambridge University Press.
Cheshire, Jenny. 1999. Taming the vernacular: some repercussions for the study of syntactic variation and spoken grammar. *Cuadernos de Filología Inglesa* 8, 59–80.
2005. Sociolinguistics and mother-tongue education. In Ammon, Dittmar, Mattheier and Trudgill (eds.), 2341–50.
Cheshire, Jenny and Viv Edwards. 1989. *The Survey of British Dialect Grammar*. In Cheshire, Edwards, Münstermann and Weltens (eds.), 200–15.
Cheshire, Jenny, Viv Edwards, Henk Münstermann and Bert Weltens (eds.). 1989. *Dialect and Education: Some European Perspectives*. Clevedon: Multilingual Matters.
Chomsky, Noam. 1965. *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.
1980. *Rules and Representations*. Oxford: Blackwell.
1986 *Knowledge of Language: Its Nature, Origin, and Use*. New York: Praeger.
1995. *The Minimalist Program*. Cambridge, MA: MIT Press.
Chomsky, Noam and Morris Halle. 1968. *The Sound Pattern of English*. New York: Harper and Row.
Clarke, Sandra, Ford Elms and Amani Youssef. 1995. The third dialect of English: some Canadian evidence. *Language Variation and Change* 7, 209–28.
Clopper, Cynthia and John Paolillo. 2006. North American English vowels: a factor-analytic perspective. *Literary and Linguistic Computing* 21(4), 445–62.
Clopper, Cynthia and David Pisoni. 2002. Perception of dialect variation: some implications for current research and theory in speech perception. *Research on Spoken Language Processing: Progress Report No. 25*. Indiana University Press.
2005. Perceptual free classification of dialect variation. Paper presented at the Twelfth International Conference on Methods in Dialectology, Université de Moncton.
Coates, Jennifer (ed.). 1998. *Language and Gender: A Reader*. Oxford: Blackwell.
Coates, Richard. 2007. Invisible Britons: the view from linguistics. In Higham (ed.), 172–91.
Coates, Richard and Andrew Breeze, with a contribution by David Horovitz. 2000. *Celtic Voices, English Places*. Stamford: Shaun Tyas.
Coetzee, Andrew. 2004. What it means to be a loser: non-optimal candidates in optimality theory. Ph.D. thesis, University of Massachusetts Amherst.
Coetzee, Andrew and Joe Pater. To appear. The place of variation in phonological theory. In John Goldsmith, Jason Riggle and Alan Yu (eds.) *The Handbook of Phonological Theory*. Oxford: Blackwell.

Collins, Peter C. 2005. The modals and quasi-modals of obligation and necessity in Australian English and other Englishes. *English World-Wide* 26, 249–73.

Conley, John M., William M. O'Barr and E. Allen Lind. 1978. The power of language: presentational style in the courtroom. *Duke Law Journal* 6, 1375–99.

Cooke, Michael. 1995. Aboriginal evidence in the cross-cultural courtroom. In Diana Eades (ed.) *Language in Evidence: Issues Confronting Multi-Cultural Australia*, 55–96. Sydney: University of New South Wales Press.

Cornips, Leonie and Karen P. Corrigan (eds.). 2005. *Syntax and Variation: Reconciling the Biological and the Social*. Amsterdam: John Benjamins.

Cornips, Leonie and Willy Jongenburger. 2001. Elicitation techniques in a Dutch syntactic dialect atlas project. In Hans Broekhuis and Ton van der Wouden (eds.) *Linguistics in The Netherlands 2001* 18, 161–84. Amsterdam/Philadelphia: John Benjamins.

Cornips, Leonie and Cecilia Poletto. 2005. On standardising syntactic elicitation techniques, PART I. *Lingua* 115(7), 939–57.

2008. Field linguistics meets formal research. Unpublished manuscript, Meertens Institute, Amsterdam.

Cotterill, Janet. 2004. *Language in the Legal Process*. Basingstoke: Palgrave Macmillan.

2007. *The Language of Sexual Crime*. Basingstoke: Palgrave Macmillan.

Coulthard, Malcolm. 1992. Forensic discourse analysis. In *Advances in Spoken Discourse Analysis*, 242–58. London: Routledge.

1993. Beginning the study of forensic texts: corpus, concordance, collocation. In Michael Hoey (ed.) *Data, Description, Discourse*, 86–97. London: Harper Collins.

1994a. On the use of corpora in the analysis of forensic texts. *Forensic Linguistics: The International Journal of Speech, Language and the Law* 1(1), 27–43.

1994b. Powerful evidence for the defence: an exercise in forensic discourse analysis. In Gibbons (ed.), 414–27.

1995. *Questioning Statements: Forensic Applications of Linguistics*. University of Birmingham, English Language Research papers.

2002. Whose voice is it? Invented and concealed dialogue in written records of verbal evidence produced by the police. In Janet Cotterill (ed.) *Language in the Legal Process*, 19–34. Basingstoke: Palgrave.

2004. Author identification, idiolect and linguistic uniqueness. *Applied Linguistics* 25(4), 431–47.

Coulthard, Malcolm and Alison Johnson. 2007. *An Introduction to Forensic Linguistics: Language in Evidence*. London: Routledge.

Coupland, Nikolas. 1988. *Dialect in Use: Sociolinguistic Variation in Cardiff English*. Cardiff: University of Wales Press.

2007a. The discursive framing of phonological acts of identity: Welshness through English. In Janina Brutt-Griffler and Catherine Evans Davies (eds.) *English and Ethnicity*, 19–48. London: Palgrave Macmillan.

2007b. *Style: Language Variation and Identity*. Cambridge University Press.

Cowart, Wayne. 1997. *Experimental Syntax: Applying Objective Methods to Sentence Judgments*. Thousand Oaks, CA: Sage Publications.

Crinson, James and John Williamson. 2004. Non-standard dialect in the formal speech of 15-year-olds on Tyneside. *Language and Education* 18, 207–19.

Cruttenden, Alan. 1995. Rises in English. In Windsor Lewis (ed.), 155–73.

Crystal, David. 1995. *The Cambridge Encyclopedia of The English Language*. Cambridge University Press.
2001. *Language and the Internet*. Cambridge University Press.
2003. *English as a Global Language*, 2nd edn. Cambridge University Press.
2008. *Txting: the gr8 db8*. Oxford University Press.
Crystal, David and Derek Davy. 1969. *Investigating English Style*. London: Longman.
Cunliffe, Barry W. 2003. *The Celts: A Very Short Introduction*. Oxford University Press.
Cutler, Cecilia. 2002. Crossing over: white teenagers, hip-hop, and African American English. Unpublished dissertation, New York University.
D'Arcy, Alexandra. 2001. Beyond mastery: a study of dialect acquisition. Unpublished M.A. thesis. St John's: Memorial University of Newfoundland.
2005a. Like: syntax and development. Unpublished doctoral dissertation. University of Toronto.
2005b. The development of linguistic constraints: Phonological innovations in St. John's. *Language Variation and Change* 17, 327–55.
2007. *Like* and language ideology: disentangling fact from fiction. *American Speech* 82, 386–419.
2008. Canadian English as a window to the rise of *like* in discourse. *Anglistik: International Journal of English Studies* 19, 125–40.
Dailey-O'Cain, Jennifer. 1999. The perception of post-unification German regional speech. In Preston (ed.).
Davies, Bethan. 2005. Communities of practice: legitimacy not choice. *Journal of Sociolinguistics* 9, 557–81.
Davies, Mark. 2002–. *Online Corpora*, http://corpus.byu.edu/.
De Decker, Paul. 2002. Beyond the city limits: the Canadian vowel shift in an Ontario small town. Unpublished MA thesis, York University.
Decker, Ken. 2000. The use of Belize Kriol to improve English proficiency. Paper presented at the 5th International Creole Workshop, Florida International University.
Dediu, Dan and D. Robert Ladd. 2007. Linguistic tone is related to the population frequency of the adaptive haplogroups of two brain size genes, *ASPM* and *Microcephalin. Proceedings of the National Academy of Sciences of the USA* 104, 10944–9.
Delaunay, Boris. 1934. Sur la sphère vide. *Bulletin of the Academy of Sciences of the USSR, Classe des Sciences Mathématiques et Naturelles* 7(6), 793–800.
Delin, Judy, Abi Searle-Jones and Rob Waller. 2006. Branding and relationship communications: the evolution of utility bills in the UK. In Saul Carliner, Jan Piet Verckens and Cathy de Waele (eds.) *Information and Document Design*, 27–59. Amsterdam: John Benjamins.
Denbo, Seth, Heather Haskins and David Robey. 2008. *Sustainability of Digital Outputs from AHRC Resource Enhancement Projects*. Report to the Arts and Humanities Research Council. December 2008. Available online at www.ahrcict.rdg.ac.uk/activities/review/sustainability.htm (last accessed 4 August 2010).
Denham, Kristin and Anne Lobeck (eds.). 2005. *Language in the Schools: Integrating Linguistic Knowledge into K-12 Teaching*. Mahwah, NJ: Lawrence Erlbaum.
2009. *Linguistics at School: Language Awareness in Primary and Secondary Education*. Cambridge University Press.
Denison, David. 1998. Syntax. In Suzanne Romaine (ed.) *The Cambridge History of the English Language, Volume IV: 1776–1997*, 92–329. Cambridge University Press.

Department for Communities and Local Government. 2004. *The Party Wall Act etc. 1996: Explanatory Leaflet*. London: HMSO.

2008. Planning Portal: The Party Wall Act etc. (1996). Available online at www.planningportal.gov.uk/england/genpub/en/1115314019877.html (last accessed 5 August 2010).

DfEE. 1999. *The National Curriculum for England: English*. London: Department for Education and Employment and the Qualifications and Curriculum Authority.

Department for Environment, F. a. R. A. 2008. Press release: DEFRA launches consultation on animal welfare codes of practice.

Deterding, David. 2001. The measurement of rhythm: a comparison of Singapore and British English. *Journal of Phonetics* 29, 217–30.

Diercks, Willy. 2002. Mental maps: linguistic geographic concepts. In Long and Preston (eds.), 51–70.

Di Paolo, Marianna and Alice Faber. 1990. Phonetic difference and the phonetic content of the tense-lax contrast in Utah English. *Language Variation and Change* 2, 155–204.

Docherty, Gerard J. and Paul Foulkes. 1999. Derby and Newcastle: instrumental phonetics and variationist studies. In Foulkes and Docherty (eds.), 47–71.

Docherty, Gerard J., Paul Foulkes, James Milroy, Leslie Milroy and David Walshaw. 1997. Descriptive adequacy in phonology: a variationist perspective. *Journal of Linguistics* 33, 275–310.

Dollinger, Stefan. 2008. *New-Dialect Formation in Canada: Evidence from the English Modal Auxiliaries*. Amsterdam: Benjamins.

Dorling, Daniel, J. Ford, A. E. Holmans and S. Regan. 2005. *The Great Divide: An Analysis of Housing Inequality*. London: Shelter.

Douglas-Cowie, Ellen, Roddy Cowie and Joan Rahilly. 1995. The social distribution of intonation patterns in Belfast. In Windsor Lewis (ed.), 180–6.

Drager, Katie and Jennifer Hay. 2006. Can you really believe your ears? The effect of stuffed toys on speech perception. Paper presented at New Zealand Language Society Conference, Christchurch.

Du Bois, John W. 2002. Stance and consequence. Paper presented at the Annual Meeting of the American Anthropological Association, New Orleans.

Dumas, Bethany. 1990. Adequacy of cigarette package warnings: an analysis of the adequacy of federally mandated cigarette package warnings. In Levi and Walker (eds.), 309–57.

Durie, Mark and Malcolm Ross. 1996. *The Comparative Method Reviewed*. Oxford University Press.

Eades, Diana. 1994. A case of communication clash: Aboriginal English and the legal system. In Gibbons (ed.), 234–64.

2008. *Courtroom Talk and Neocolonial Control*. Berlin, New York: Mouton de Gruyter.

Eades, Diana and Jacques Arends. 2004. Using language analysis in the determination of national origin of asylum seekers: an introduction. *The International Journal of Speech, Language and the Law* 11(2), 179–99.

Eckert, Penelope. 1988. Adolescent social structure and the spread of linguistic change. *Language in Society* 17, 183–207.

1989. The whole woman: sex and gender differences in variation. *Language Variation and Change* 1, 245–67.

1998. Gender and sociolinguistic variation. In Coates (ed.), 64–75.

2000. *Linguistic Variation as Social Practice: The Linguistic Construction of Identity at Belten High*. Oxford: Blackwell.
2003. The meaning of style. *Texas Linguistic Forum* 47, 41–53.
2005. Variation, convention, and social meaning. Paper presented at the Annual Meeting of the Linguistic Society of America. Available at www.stanford.edu/~eckert/EckertLSA2005.pdf (last accessed 2 August 2010).
2008. Variation and the indexical field. *Journal of Sociolinguistics* 12, 453–76.
Eckert, Penelope and Sally McConnell-Ginet. 1992. Think practically and look locally: language and gender as community-based practice. *Annual Review of Anthropology* 21, 461–90.
Eckert, Penelope and John R. Rickford (eds.). 2001. *Style and Sociolinguistic Variation*. Cambridge University Press.
Edwards, John. 1982. Language attitudes and their implications. In Ryan and Giles (eds.), 20–33.
2006. Educational failure. In Keith Brown (ed.) *Encyclopedia of Language and Linguistics*, 2nd edn. Amsterdam: Elsevier.
Ellegård, Alvar. 1953. *The Auxiliary Do: The Establishment and Regulation of Its Use in English*. Stockholm: Almqvist and Wiksell.
Ellis, Stanley. 1956. Dialect-hunting by caravan. *University of Leeds Review* 5 (1956–1957), 39–48.
1994. The Yorkshire Ripper enquiry: Part I. *Forensic Linguistics* 1(2), 197–206.
Elsworth, C. 2008. OJ Simpson audio of alleged raid may have been tampered with. *The Daily Telegraph*, 18 September 2008. Available online at www.telegraph.co.uk/news/newstopics/celebritynews/2981171/OJ-Simpson-audio-of-alleged-raid-may-have-been-tampered-with.html (last accessed 5 August 2010).
Embleton, Sheila. 1987. Multidimensional scaling as a dialectometrical technique. In Rose M. Babitch (ed.) *Papers from the Eleventh Annual Meeting of the Atlantic Provinces Linguistic Association*, 33–49. New Brunswick: Centre Universitaire de Shippagan.
2000. Lexicostatistics/glottochronology: from Swadesh to Sankoff to Starostin to future horizons. In Colin Renfrew, April McMahon and Robert L. Trask (eds.) *Time Depth in Historical Linguistics*, 2 vols, 143–66. Cambridge: McDonald Institute for Archaeological Research.
Embleton, Sheila, Dorin Uritescu and Eric Wheeler. 2007. Data capture and presentation in the Romanian Online Dialect Atlas. *Papers from 12th International Conference on Methods in Dialectology* (Linguistica Atlantica 27–8, 2007).
Embleton, Sheila and Eric Wheeler. 1997. Multidimensional scaling and the SED data. In Viereck and Ramisch (eds.), 5–11.
Eriksson, Andrew and Francisco Lacerda. 2007. Charlatanry in forensic speech science: a problem to be taken seriously. *The International Journal of Speech, Language and the Law* 14(2), 169–93.
Errington, Joseph. 1985. On the nature of the sociolinguistic sign: describing the Javanese speech levels. In Elizabeth Mertz and Richard J. Parmentier (eds.) *Semiotic Mediation*, 287–310. Orlando, FL: Academic Press.
Esling, John H. 1978. The identification of features of voice quality in social groups. *Journal of the International Phonetic Association* 7, 18–23.
Everitt, Brian, Sabine Landau and Morven Leese. 2001. *Cluster Analysis*, 4th edn. London: Arnold.

Facchinetti, Roberta, Manfred Krug and Frank Palmer (eds.). 2003. *Modality in Contemporary English*. Berlin and New York: Mouton de Gruyter.
Fairon, Cédrick. 2000. GlossaNet: parsing a web site as a corpus. *Lingvisticae Investigationes* 22, 327–40.
Falconer, Douglas S. and Trudy F. C. Mackay. 1996. *Introduction to Quantitative Genetics*, 4th edn. Harlow: Longmans.
Fasold, Ralph W. 1984. *The Sociolinguistics of Society*. Oxford: Blackwell.
1991. The quiet demise of variable rules. *American Speech* 66, 3–21.
Fasold, Ralph W. and Deborah Schiffrin (eds.). 1989. *Language Change and Variation*. Amsterdam: Benjamins.
Featherston, Sam 2005. Magnitude estimation and what it can do for your syntax: some *wh*-constraints in German. *Lingua* 115, 1525–50.
Felsenstein, Joseph. 2004. *Inferring Phylogenies*. Sunderland, MA: Sinauer Associates.
Fennell, Barbara A. 2001. *A History of English. A Sociolinguistic Approach*. Oxford: Blackwell.
Ferrara, Kathleen, Hans Brunner and Greg Whittemore. 1991. Interactive written discourse as an emergent register. *Written Communication* 8, 8–34.
Filippula, Markku and Juhani Klemola. 2009. Special issue on re-evaluating the Celtic hypothesis. *English Language and Linguistics* 13(2), 155–61.
Finegan, Edward. 1997. Sociolinguistics and the law. In Florian Coulmas (ed.) *The Handbook of Sociolinguistics*, 421–35. Oxford: Blackwell.
Finnegan, Katie. Forthcoming. Dialect levelling in Sheffield English. Ph.D. thesis, University of Sheffield.
Fisher, Simon E., Cecilia S. L. Lai and Anthony P. Monaco. 2003. Deciphering the genetic basis of speech and language disorders. *Annual Review of Neuroscience* 26, 57–80.
Flanigan, Beverly Olson and Franklin Paul Norris. 2000. Cross-dialectal comprehension as evidence for boundary mapping: perceptions of the speech of southeastern Ohio. *Language Variation and Change* 12, 175–201.
Fletcher, Janet, Esther Grabe and Paul Warren. 2005. Intonational variation in four dialects of English: the high rising tone. In Sun-Ah Jun (ed.) *Prosodic Typology: The Phonology of Intonation and Phrasing*, 396–409. Oxford University Press.
Flew, Antony and Stephen Priest. 2002. *A Dictionary of Philosophy*, 3rd edn. London: PanMacmillan.
Forster, Peter, Rosalind Harding, Antonio Torroni and Hans-Jurgen Bandelt. 1996. Origin and evolution of Native American mtDNA variation: a reappraisal. *American Journal of Human Genetics* 59, 935–45.
Foulkes, Paul and Anthony Barron. 2000. Telephone speaker recognition amongst members of a close social network. *Forensic Linguistics* 7(2), 180–98.
Foulkes, Paul and Gerard J. Docherty (eds.). 1999. *Urban Voices: Accent Studies in the British Isles*. London: Arnold.
Foulkes, Paul and Gerard J. Docherty. 2000. Another chapter in the story of /r/: 'labiodental' variants in British English. *Journal of Sociolinguistics* 4, 30–59.
Foulkes, Paul and Gerard J. Docherty. 2006. The social life of phonetics and phonology. *Journal of Phonetics* 34, 409–38.
Fourakis, Marios and Robert Port. 1986. Stop epenthesis in English. *Journal of Phonetics* 14, 197–221.

Fox, Anthony. 1995. *Linguistic Reconstruction: An Introduction to Theory and Method*. Oxford University Press.

Fox, Gwyneth. 1993. A comparison of 'policespeak' and 'normalspeak': a preliminary study. In John M. Sinclair, Michael Hoey and Gwyneth Fox (eds.) *Techniques of Description: A Festschrift for Malcolm Coulthard*, 183–95. London: Routledge.

Francis, W. Nelson and Henry Kučera. 1964. *A Standard Corpus of Present-day Edited American English*. Providence, RI: Brown University.

French, Peter. 1994. An overview of forensic phonetics with particular reference to speaker identification. *Forensic Linguistics* 1(2), 169–81.

French, Peter and Philip Harrison. 2007. Position statement concerning use of impressionistic likelihood terms in forensic speaker comparison cases. *The International Journal of Speech, Language and the Law* 14(1), 139–44.

French, Peter, Philip Harrison and Jack Windsor Lewis. 2006. R v. John Samuel Humble: the Yorkshire Ripper Hoaxer trial. *The International Journal of Speech, Language and the Law* 13(2), 255–73.

Fromont, Robert and Jennifer Hay. 2008. ONZE Miner: the development of a browser-based research tool. *Corpora* 3(2), 173–93.

Garside, Roger. 1987. The CLAWS word-tagging system. In Roger Garside, Geoffrey Leech and Geoffrey Sampson (eds.) *The Computational Analysis of English: A Corpus-based Approach*, 30–41. London: Longman.

Gibbons, John (ed.). 1994. *Language and the Law*. Essex: Longman.

Gibbons, John. 2001. Legal transformations in Spanish: an 'audencia' in Chile. *Forensic Linguistics* 8(2), 24–43.

2006. Forensic sociolinguistics. In Ammon, Dittmar, Mattheier and Trudgill (eds.), 2316–23.

2003. *Forensic Linguistics: An Introduction to Language in the Justice System*. Oxford: Blackwell.

Giegerich, Heinz. 1992. *English Phonology: An Introduction*. Cambridge University Press.

Giles, Howard (ed.). 1977. *Language, Ethnicity and Intergroup Relations*. London: Academic Press.

2002. *Law Enforcement, Communication and Community*. Amsterdam: John Benjamins.

Giles, Howard and Richard Y. Bourhis. 1976. Voice and racial categorisation in Britain. *Communication Monographs* 43, 108–14.

Gilles, Peter and Jörg Peters (eds.). 2004. *Regional Variation in Intonation*. Tübingen: Max Niemeyer Verlag.

Glauser, Beat. 1974. *The Scottish-English Linguistic Border*. Bern: Francke.

Goebl, Hans 1984. *Dialektometrische Studien: Anhand italoromanischer, rätoromanischer und galloromanischer Sprachmaterialien aus AIS und ALF*, 3 Vols. Tübingen: Max Niemeyer.

1993. Dialectometry. A short overview of the principles and practice of quantitative classification of linguistic atlas data. In Reinhard Köhler and Burghard Rieger (eds.) *Contributions to Quantitative Linguistics*, 277–315. Dordrecht, Boston, London: Kluwer.

Goebl, Hans. 1997. Some dendrographic classifications of the data of CLAE 1 and CLAE 2. In Viereck and Ramisch (eds.), 23–32.

2006. Recent advances in Salzburg dialectometry. *Literary and Linguistic Computing* 21, 411–35.
2007. A bunch of dialectometrical flowers: a brief introduction to dialectometry. In Ute Smit, Stefan Dollinger, Julia Hüttner, Gunther Kaltenböck and Ursula Lutzky (eds.) *Tracing English Through Time: Explorations in Language Variation. Austrian Studies in English* 95, 133–71.
Goebl, Hans and Guillaume Schiltz. 1997. A dialectometrical compilation of CLAE 1 and CLAE 2. Isoglosses and dialect integration. In Viereck and Ramisch (eds.), 13–21.
Goodey, B. 1971a. *City Scene: An Exploration into the Image of Central Birmingham as Seen by Area Residents*. Birmingham University Research Memorandum 10.
1971b. *Perception of the Environment: An Introduction to the Literature*. University of Birmingham: Centre for Urban and Regional Studies.
Goodwin, Marjorie Harness, Charles Goodwin and Malcah Yaeger-Dror. 2002. Multimodality in girls' game disputes. *Journal of Pragmatics* 34, 1621–49.
Gooskens, Charlotte and Wilbert Heeringa. 2004. Perceptive evaluation of Levenshtein Dialect Distance measurements using Norwegian dialect data. *Language Variation and Change* 16, 189–207.
Gordon, Elizabeth [with S. Hervey, R. Leitch and E. Holstein]. 1996. *Exploring Language: A Handbook for New Zealand Teachers*. Wellington: Ministry for Education and Learning Media.
Gordon, Elizabeth, Lyle Campbell, Jennifer Hay, Margaret Maclagan, Andrea Sudbury and Peter Trudgill. 2004. *New Zealand English. Its Origins and Evolution*. Cambridge University Press.
Gordon, Elizabeth, Margaret Maclagan and Jennifer Hay. 2007. The ONZE corpus. In Beal, Corrigan and Moisl (eds.), Vol. II, 82–104.
Gould, Peter and Rodney White. 1986. *Mental Maps*, 2nd edn. Boston, MA: Allen and Unwin.
Grabe, Esther. 2004. Intonational variation in urban dialects of English spoken in the British Isles. In Gilles and Peters (eds.), 9–31.
Grabe, Esther, Brechtje Post, Francis Nolan and Kimberley Farrar. 2000. Pitch accent realization in four varieties of British English. *Journal of Phonetics* 28, 161–85.
Grant, Tim. 2007. Quantifying evidence in forensic authorship analysis. *The International Journal of Speech, Language and the Law* 14(1), 1–25.
Grant, Tim and Kevin Baker. 2001. Identifying reliable, valid markers of authorship: a response to Chaski. *Forensic Linguistics* 8(1), 66–79.
Green, Lisa. 2002. *African American English: A Linguistic Introduction*. Cambridge University Press.
Greenbaum, Sidney. 1992. A new corpus of English: ICE. In Svartvik (ed.), 171–83.
Greenbaum, Sidney (ed.). 1996. *Comparing English Worldwide: The International Corpus of English*. Oxford: Clarendon Press.
Grimmer, Martin. 2007. Britons in early Wessex: the evidence of the Law Code of Ine. In Higham (ed.), 102–14.
Guilford, J. Paul. 1954. *Psychometric Methods*, 2nd edn. New York: McGraw Hill.
Guillemin, Bernard and C. Watson. 2008. Impact of the mobile phone network on the speech signal – some preliminary findings. *International Journal of Speech, Language and the Law* 15(2), 193–218.
Gumperz, John and Dell Hymes (eds.). 1964. The ethnography of communication. *American Anthropologist* 66(6), part 2.

Gut, Ulrike. 2002. Prosodic aspects of Standard Nigerian English. In Gut, Ulrike and Dafydd Gibbon (eds.) *Typology of African Prosodic Systems*, 167–78. Bielefeld University.

Guy, Gregory R. 1980. Variation in the group and the individual: the case of final stop deletion. In William Labov (ed.) *Locating Language in Time and Space*, 1–36. New York: Academic Press.

1990. The sociolinguistic types of language change. *Diachronica* 7, 47–67.

2005. Grammar and usage: a variationist response. *Language* 81, 561–3.

Guy, Gregory R. and Charles Boberg. 1997. Inherent variability and the obligatory contour principle. *Language Variation and Change* 9, 149–64.

Guy, Gregory R., Crawford Feagin, Deborah Schiffrin and John Baugh (eds.). 1996. *Towards a Social Science of Language. Vol. I: Variation and Change in Language and Society*. Amsterdam and Philadelphia: John Benjamins.

Guy, Gregory R., Barbara M. Horvath, Julia Vonwiller, Elaine Daisley and Inge Rogers. 1986. An intonational change in progress in Australian English. *Language in Society* 15, 22–52.

Guy, Gregory R. and Julia Vonwiller. 1984. The meaning of an intonation in Australian English. *Australian Journal of Linguistics* 4, 1–17.

Hägerstrand, Torsten. 1952. The *Propagation of Innovation Waves*. Lund Studies in Geography, B: Human Geography, 4. Lund, Sweden: Gleerup.

Haddican, William. 2010. Theme-goal ditransitives and theme passivisation in British English. *Lingual* 20, 2424–43.

Haeri, Niloofar. 1996a. *The Sociolinguistic Market of Cairo: Gender, Class, and Education*. London and New York: Kegan Paul International.

1996b. 'Why do women do this?' Sex and gender differences in speech. In Guy, Feagin, Schiffrin and Baugh (eds.), 101–14.

Hale, Mark. 2007. *Historical Linguistics: Theory and Method*. Blackwell, Oxford.

Hale, Mark and Charles Reiss. 2008. *The Phonological Enterprise*. Oxford University Press.

Hale, Sandra. 1997. Clash of world perspectives: the discursive practices of the law, the witness and the interpreter. *Forensic Linguistics: The International Journal of Speech, Language and the Law* 4(2), 197–209.

1999. Interpreters' treatment of discourse markers in courtroom questions. *Forensic Linguistics: The International Journal of Speech, Language and the Law* 6(1), 57–82.

2004. *The Discourse of Court Interpreting: Discourse Practices of the Law, the Witness and the Interpreter*. Amsterdam: John Benjamins.

Hand, David, Heikki Mannila and Padhraic Smyth. 2001. *Principles of Data Mining*. Cambridge, MA: MIT Press.

Harris, John. 1984a. English in the north of Ireland. In Trudgill (ed.), 115–34.

1984b. Syntactic variation and dialect divergence. *Journal of Linguistics* 20, 303–27.

1985. *Phonological Variation and Change*. Cambridge University Press.

1994. *English Sound Structure*. Oxford: Blackwell.

Haslerud, Vibecke and Anna-Brita Stenström. 1995. The Bergen Corpus of London Teenage Language. In Geoffrey Leech, Greg Myers and Jenny Thomas (eds.) *Spoken English on Computer*, 235–42. London: Longman.

Hay, Jennifer, Aaron Nolan and Katie Drager. 2006. From fush to feesh: exemplar priming in speech perception. *Linguistic Review* 23, 351–79.

Hay, Jennifer and Katie Drager. 2007. Sociophonetics. *Annual Review of Anthropology* 36, 89–103.

Hebdige, Dick. 1979. *Subculture: The Meaning of Style*. London and New York: Routledge.

Heeringa, Wilbert. 2004. Measuring dialect pronunciation differences using Levenshtein Distance. Ph.D. thesis, Rijksuniversiteit Groningen. Available online at www.let.rug.nl/~heeringa/dialectology/thesis/ (last accessed 6 August 2010).

Heeringa, Wilbert, Peter Kleiweg, Charlotte Gooskens and John Nerbonne. 2006. Evaluation of string distance algorithms for dialectology. In John Nerbonne and Erhard Hinrichs (eds.) *Linguistic Distances Workshop at the Joint Conference of International Committee on Computational Linguistics and the Association for Computational Linguistics*, Sydney, July, 2006, 51–62.

Heeringa, Wilbert and John Nerbonne. 2001. Dialect areas and dialect continua. *Language Variation and Change* 13, 375–400.

Heeringa, Wilbert, John Nerbonne and Peter Kleiweg. 2002. Validating dialect comparison methods. In Wolfgang Gaul and Gunter Ritter (eds.) *Classification, Automation, and New Media*. Proceedings of the 24th Annual Conference of the Gesellschaft für Klassifikation e. V., University of Passau, March 15–17, 2000, 445–52. Berlin, Heidelberg and New York: Springer.

Heffer, Chris. 2005. *The Language of Jury Trial: A Corpus-aided Analysis of Legal-Lay Discourse*. Basingstoke: Palgrave Macmillan.

Heggarty, Paul. Forthcoming. *Measured Language*. Publications of the Philological Society, Oxford: Blackwell.

Heggarty, Paul, April McMahon and Robert McMahon. 2005. From phonetic similarity to dialect classification: a principled approach. In Nicole Delbecque, Johan van der Auwera and Dirk Geeraerts (eds.) *Perspectives on Variation*, 43–91. Amsterdam: Mouton de Gruyter.

Heine, Bernd and Tania Kuteva. 2005. *Language Contact and Grammatical Change*. Cambridge University Press.

Henry, Alison. 1995. *Belfast English and Standard English: Dialect Variation and Parameter Setting*. Oxford University Press.

2002. Variation and syntactic theory. In Chambers, Trudgill and Schilling-Estes (eds.), 267–82.

Henton, Caroline and Anthony Bladon. 1985. Breathiness in a normal female speaker: inefficiency versus desirability. *Language and Communication* 5, 221–7.

Herold, Ruth. 1997. Solving the actuation problem: merger and immigration in eastern Pennsylvania. *Language Variation and Change* 9, 165–89.

Heselwood, Barry and Louise McChrystal. 1999. The effect of age-group and place of L1 acquisition on the realisation of Panjabi stop consonants in Bradford: an acoustic sociophonetic study. *Leeds Working Papers in Linguistics* 7, 49–68.

Hibiya, Junko. 1996. Denasalization of the velar nasal in Tokyo Japanese: observations in real time. In Guy, Feagin, Schiffrin and Baugh (eds.), 161–70.

Higham, Nicholas. 1992. *Rome, Britain and the Anglo-Saxons*. The Archaeology of Change series. London: Seaby.

Higham, Nicholas (ed.). 2007. *Britons in Anglo-Saxon England*. Woodbridge: The Boydell Press.

Hillenbrand, James, Ronald A. Cleveland and Robert L. Erickson. 1994. Acoustic correlates of breathy vocal quality. *Journal of Speech and Hearing Research* 37, 769–78.

Hinskens, Franz, Roeland van Hout and Leo Wetzels. 1997. *Variation, Change and Phonological Theory*. Amsterdam: John Benjamins.
HMSO. 1996. *Party Wall etc. Act*. London: HMSO.
2001. *Special Educational Needs and Disability Act*. London: HMSO.
2006. *Animal Welfare Act*. London: HMSO.
Hoenigswald, Henry. 1960. *Language Change and Linguistic Reconstruction*. University of Chicago Press.
1966. A proposal for the study of folk linguistics. In William Bright (ed.) *Sociolinguistics*, 16–22. The Hague: Mouton.
Hogg, Richard. 1992–2001. *The Cambridge History of the English Language* (6 volumes). Cambridge University Press.
Hogg, Richard and David Denison. 2006. *A History of the English Language*. Cambridge University Press.
Hollien, Harry. 1996. Consideration of guidelines for earwitness lineups. *Forensic Linguistics* 3(1), 14–23.
1990. *The Acoustics of Crime: The New Science of Forensic Phonetics*. New York: Plenum Press.
2002. *Forensic Voice Identification*. New York: Academic Press.
Holmes, Janet. 1982. The functions of tag questions. *English Language Research Journal* 3, 40–65.
1984. Hedging your bets and sitting on the fence: some evidence for hedges as support structures. *Te Reo* 27, 47–62.
1995. *Women, Men and Politeness*. London/New York: Longman.
1996. The New Zealand component of ICE: some methodological challenges. In Greenbaum (ed.), 163–81.
1998. Women's talk: the question of sociolinguistic universals. In Coates (ed.), 461–83.
Honeybone, Patrick. 2001. Lenition inhibition in Liverpool English. *English Language and Linguistics* 5, 213–49.
2005. J. R. Firth. In Siobhan Chapman and Christopher Routledge (eds.) *Key Thinkers in Linguistics and the Philosophy of Language*, 80–6. Edinburgh University Press.
2008. Lenition, weakening and consonantal strength: tracing concepts through the history of phonology. In Joaquim Brandão de Carvalho, Tobias Scheer and Philippe Ségéral (eds.) *Lenition and Fortition*, 9–93. Berlin: Mouton de Gruyter.
Hooper, Joan. 1976. *An Introduction to Natural Generative Phonology*. New York: Academic Press.
Hoppenbrouwers, Cor and Geer Hoppenbrouwers. 1988. De featurefrequentiemethode en de classificatie van Nederlandse dialecten. *TABU: Bulletin voor taalwetenschap* 18(2), 51–92.
2001. *De indeling van de Nederlandse streektalen. Dialecten van 156 steden en dorpen geklasseerd volgens de FFM*. Assen: Koninklijke Van Gorcum B.V.
Horan, Anne. 2002. English grammar in schools. Proceedings of the 2002 conference of the Australian Linguistic Society.
Horvath, Barbara M. 1985. *Variation in Australian English: The Sociolects of Sydney*. Cambridge University Press.

Howald, Blake S. 2008. Authorship attribution under the rules of evidence: empirical approaches in a layperson's legal system. *International Journal of Speech, Language and the Law* 15(2), 219–47.

Howard, David, Allen Hirson, T. Brookes and A. M. Tyrrell. 1995. Spectography of disputed speech samples by peripheral human hearing modelling. *Forensic Linguistics* 2(1), 28–38.

Huddleston, Rodney and Geoffrey K. Pullum. 2005. *A Student's Introduction to English Grammar*. Cambridge University Press.

Hudson, Amelia I. and Anthony Holbrook. 1981. A study of reading fundamental frequency of young Black adults. *Journal of Speech and Hearing Research* 24, 197–201.

Hudson, Richard. 1975. The meaning of questions. *Language* 51, 1–31.

1992. *Teaching Grammar: A Guide for the National Curriculum*. Oxford: Blackwell.

1997. Inherent variability and linguistic theory. *Cognitive Linguistics* 8, 73–108.

2007. English dialect syntax in Word Grammar. *English Language and Linguistics* 11, 383–405.

Hudson, Richard and Jasper Holmes. 1995. *Children's Use of Spoken Standard English*. A short report prepared for the Schools Curriculum and Assessment Authority. London: SCAA.

Hudson, Richard and John Walmsley. 2005. The English patient: English grammar and teaching in the twentieth century. *Journal of Linguistics* 41, 593–622.

Hughes, Arthur, Peter Trudgill and Dominic Watt. 2005. *English Accents and Dialects: An Introduction to Social and Regional Varieties of English in the British Isles*, 4th edn. London: Hodder Arnold.

Hundt, Marianne, Nadja Nesselhauf and Carolin Biewer (eds.). 2007. *Corpus Linguistics and the Web*. Amsterdam and New York: Rodopi.

Huson, Daniel and David Bryant. 2006. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution* 23(2), 254–67.

Hymes, Dell. 1962. The ethnography of speaking. In Thomas Gladwin and William Sturtevant (eds.) *Anthropology and Human Behavior*, 13–53. Washington, DC: Washington Anthropological Society.

IHGSC (International Human Genome Sequencing Consortium). 2004. Finishing the euchromatic sequence of the human genome. *Nature* 431, 931–45.

Inoue, Fumio. 1999a. Classification of dialects by image: English and Japanese. In Preston (ed.), 147–60.

1999b. Subjective dialect division in Great Britain. In Preston (ed.), 161–76.

Irvine, Judith T. 2001. 'Style' as distinctiveness: the culture and ideology of linguistic differentiation. In Eckert and Rickford (eds.), 21–43.

Janson, Tore. 1983. Sound change in perception and production. *Language* 59, 93–103.

1986. Sound change in perception: an experiment. In John J. Ohala and Jeri J. Jaeger (eds.) *Experimental Phonology*, 253–60. Orlando, FL: Academic Press.

Janson, Tore and Richard Schulman. 1983. Non-distinctive features and their use. *Journal of Linguistics* 19, 321–36.

Jessen, Michael. 2008. Forensic phonetics. *Language and Linguistics Compass* 2(4), 671–711.

Jobling, Mark A., Matthew Hurles and Chris Tyler-Smith. 2004. *Human Evolutionary Genetics: Origins, Peoples and Disease*. Garland Science: Oxford.

Jobling, Mark and Chris Tyler-Smith. 2003. The human Y chromosome: an evolutionary marker comes of age. *Nature Reviews Genetics* 4, 598–612.

Johansson, Stig, Geoffrey Leech and Helen Goodluck. 1978. *Manual of Information to Accompany the Lancaster-Oslo-Bergen Corpus of British English, for Use with Digital Computers*. Oslo: Department of English, Oslo University.

Johnson, Alison. 1997. Textual kidnapping – a case of plagiarism among three student texts? *Forensic Linguistics* 4(2), 210–25.

Johnstone, Barbara. 2000. *Qualitative Methods in Sociolinguistics*. Oxford University Press.

2007. Linking identity and dialect through stancetaking. In Robert Englebretson (ed.) *Stancetaking in Discourse: Subjectivity, Evaluation, Interaction*, 49–67. Amsterdam: John Benjamins.

Johnstone, Barbara and Scott Kiesling. 2008. Indexicality and experience: exploring the meanings of /aw/-monothongization in Pittsburgh. *Journal of Sociolinguistics* 12, 5–33.

Jones, Charles. 1997. *The Edinburgh History of the Scots Language*. Edinburgh University Press.

Jones, Steve. 1993. *The Language of the Genes*. London: Harper Collins.

Jun, Sun-Ah and Christina Foreman. 1996. Boundary tones and focus realization in African-American intonation. Paper presented at the 3rd joint meeting of the Acoustical Society of America and the Acoustical Society of Japan, Honolulu, HI, 6 December.

Juvaini, Ata-Malik. 1260. *Genghis Khan: The History of the World-Conqueror*, trans. John A. Boyle, Manchester University Press/UNESCO, 1997.

Kallen, Jeffrey and John Kirk. 2007. ICE-Ireland: Local variations on global standards. In Beal, Corrigan and Moisl (eds.), Vol. I, 121–62.

Karafet, Tatania M., Fernando L. Mendez, Monica B. Meilerman, Peter A. Underhill, Stephen L. Zegura and Michael Hammer. 2008. New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. *Genome Research* 18, 830–9.

Kaufman, Leonard and Peter Rousseeuw. 2005. *Finding Groups in Data. An Introduction to Cluster Analysis*, 2nd edn. Hoboken, NJ: Wiley Blackwell.

Kennedy, Graeme. 1998. *An Introduction to Corpus Linguistics*. London and New York: Longman.

Kerswill, Paul. 1996. Children, adolescents and language change. *Language Variation and Change* 8, 177–202.

2003. Dialect levelling and geographical diffusion in British English. In D. Britain and J. Cheshire (eds.) *Social Dialectology: In Honour of Peter Trudgill*, 223–43. Amsterdam: John Benjamins.

2006. Migration and language. In Ammon, Dittmar, Mattheier and Trudgill (eds.), 2271–85.

Kerswill, Paul, Carmen Llamas and Clive Upton. 1999. The first SuRE moves: early steps towards a large dialect database. In Clive Upton and Katie Wales (eds.) *Dialectal Variation in English: Proceedings of the Harold Orton Centenary Conference 1998*. Leeds Studies in English 30.

Kerswill, Paul and Ann Williams. 2000. Creating a New Town koine: children and language change in Milton Keynes. *Language in Society* 29, 65–115.

2002. Dialect recognition and speech community focussing in new and old towns in England: the effects of dialect levelling, demographic and social networks. In Long and Preston (eds.), 173–204.

Kessler, Brett. 1995. Computational dialectology in Irish Gaelic. In *Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics*, 60–7. Dublin: Association for Computational Linguistics.

2005. Phonetic comparison algorithms. *Transactions of the Philological Society* 103(2), 243–60.

Kimura, Motoo. 1983. *The Neutral Theory of Molecular Evolution*. Cambridge University Press.

Kimura, Motoo and James F. Crow. 1964. The number of alleles that can be maintained in a finite population. *Genetics* 49, 725–38.

King, Kendall A. and Carol Benson. 2008. Vernacular and indigenous literacies. In Spolsky and Hult (eds.), 341–54.

King, Robert D. 1969. *Historical Linguistics and Generative Grammar*. Englewood Cliffs, NJ: Prentice-Hall.

Kiparsky, Paul. 1968. *How Abstract is Phonology?* Bloomington, IN: Indiana University Linguistic Club.

To appear. Compensatory lengthening. www.stanford.edu/~kiparsky (last accessed 26 July 2010).

Kirk, John. 1992. The Northern Ireland Transcribed Corpus of Speech. In Gerhard Leitner (ed.) *New Directions in Language Corpora*, 65–74. Berlin and New York: Mouton de Gruyter.

Klemola, Juhani. 2009. Traces of historical infinitive in English dialects and their Celtic connections. *English Language and Linguistics* 13(2), 295–308.

Klima, Edward. 1964. Relatedness between grammatical systems. *Language* 40, 1–20.

Kniffka, Hannes (ed.). 1996. *Recent Developments in Forensic Linguistics*. Frankfurt: Peter Lang.

Knowles, Gerald. 1973. Scouse: the urban dialect of Liverpool. Unpublished Ph.D. thesis, University of Leeds.

Komter, Martha. 2006. From talk to text: the interactional construction of a police record. *Research on Language and Social Interaction* 39(3), 201–28.

Kortmann, Bernd. 2002. New prospects for the study of English dialect syntax: impetus from syntactic theory and language typology. In Barbiers, Koeneman, Lekakou and van der Ham (eds.), 185–213.

Kortmann, Bernd, Edgar Schneider, Kate Burridge, Rajend Mesthrie and Clive Upton (eds.). 2004. A *Handbook of Varieties of English: A Multimedia Reference Tool* (2 volumes). Berlin, New York: Mouton de Gruyter.

Kortmann, Bernd and Clive Upton (eds.). 2008. *Varieties of English 1: The British Isles*. Berlin/New York: Mouton de Gruyter.

Krämer, Jörg. 1995. Delaunay triangulation in two and three dimensions. Ph.D. thesis, University of Tübingen.

Kremer, Ludger. 1999 (1984). The Netherlands-German border as a subjective dialect boundary. In Preston (ed.), 31–6.

Kretzschmar, William A., Jean Anderson, Joan C. Beal, Karen P. Corrigan, Lisa Lena Opas-Hänninen and Bartlomiej Plichta. 2006. Collaboration on corpora for regional and social analysis. *Journal of English Linguistics* 34, 172–205.

Kretzschmar, William A., Jr., Virginia G. McDavid, Theodore K. Lerud and Ellen Johnson (eds.). 1993. *Handbook of the Linguistic Atlas of the Middle and South Atlantic States*. University of Chicago Press.

Kroch, Anthony. 1989a. Function and grammar in the history of English: periphrastic 'do'. In Fasold and Schiffrin (eds.), 133–72.

1989b. Reflexes of grammar in patterns of language change. *Language Variation and Change* 1(3), 199–244.

2000. Syntactic change. In Mark Baltin and Chris Collins (eds.). *The Handbook of Contemporary Syntactic Theory*, 629–739. Malden, MA: Blackwell.

Krug, Manfred. 2000. *Emerging English Modals: A Corpus-based Study of Grammaticalization*. Berlin and New York: Mouton de Gruyter.

Kruskal, Joseph. 1999. An overview of sequence comparison. In Sankoff and Kruskal (eds.), 1–44.

Künzel, Hermann. 2001. Beware of the 'telephone effect': the influence of telephone transmission on the measurement of formant frequencies. *The International Journal of Speech, Language and the Law* 8(1), 80–99.

Kurath, Hans. 1972. *Studies in Area Linguistics*. Bloomington, IN: Indiana University Press.

Kurath, Hans, Miles Hanley, Bernard Bloch and Guy S. Lowman. 1939–43. *The Linguistic Atlas of New England*. 3 volumes in 6 parts. Providence, RI: Brown University Press.

Kurath, Hans and Guy S. Lowman, Jr. 1970. *The Dialectal Structure of Southern England: Phonological Evidence*. Publication of the American Dialect Society 54. Tuscaloosa, AL: University of Alabama Press.

Kurath, Hans and Raven I. McDavid, Jr. 1961. *Pronunciation of English in the Atlantic States*. Ann Arbor, MI: University of Michigan Press.

Kytö, Merja. 1996. *Manual to the Diachronic Part of the Helsinki Corpus of English Texts: Coding Conventions and Lists of Source Texts*, 3rd edn. Department of English, University of Helsinki.

Kytö, Merja, Juhani Rudanko and Erik Smitterberg. 2000. Building a bridge between the present and the past: a corpus of 19th-century English. *ICAME Journal* 24, 85–97. Available online at http://icame.uib.no/ij24/ (last accessed 4 August 2010).

Labov, William. 1963. The social motivation of a sound change. *Word* 19, 273–309.

1966. *The Social Stratification of English in New York City*. Washington, DC: Center for Applied Linguistics [2nd edition, 2006, New York: Cambridge University Press].

1969. Contraction, deletion, and inherent variability of the English copula. *Language* 45, 715–62.

1972a. *Language in the Inner City: Studies in the Black English Vernacular*. Philadelphia, PA: University of Pennsylvania Press.

1972b. *Sociolinguistic Patterns*. Philadelphia, PA: University of Pennsylvania Press.

1972c. *The Design of a Sociolinguistic Research Project*. Report of the Sociolinguistics Workshop, Central Institute of Indian Languages.

1975. Empirical foundations of linguistic theory. In Robert Austerlitz (ed.) *The Scope of American Linguistics. The First Golden Anniversary Symposium of the Linguistic Society of America*, 77–133. Lisse: The Peter de Ridder Press.

1980. The social origins of sound change. In *Locating Language in Time and Space*, 251–65. New York: Academic Press.
1981. Resolving the neogrammarian controversy. *Language* 57(2), 267–308.
1984. Field methods of the project on linguistic change and variation. In John Baugh and Joel Sherzer (eds.) *Language in Use: Readings in Sociolinguistics*, 28–54. Englewood Cliffs, NJ: Prentice-Hall.
1990. The intersection of sex and social class in the course of linguistic change. *Language Variation and Change* 2, 205–54.
1991. The three dialects of English. In Penelope Eckert (ed.) *New Ways of Analyzing Sound Change*, 1–44. New York: Academic.
1994. *Principles of Linguistic Change, Vol. I: Internal Factors*. Language in Society 20. Oxford: Blackwell.
1996. When intuitions fail. *Papers from the 32nd Regional Meeting of the Chicago Linguistics Society* 32, 76–106.
2001. *Principles of Linguistic Change, Vol. II: Social Factors*. Language in Society 20. Oxford: Blackwell.
2010. *Principles of Linguistic Change, Vol. III: Cognitive and Cultural Factors*. Malden, MA: Blackwell.

Labov, William and Sharon Ash. 1997. Understanding Birmingham. In Cynthia Bernstein, Thomas Nunnally and Robin Sabino (eds.) *Language Variety in the South Revisited*, 508–73. Tuscaloosa, AL: University of Alabama Press.

Labov, William, Sharon Ash and Charles Boberg. 2006. *The Atlas of North American English: Phonetics, Phonology and Sound Change. A Multimedia Reference Tool*. Berlin: Mouton de Gruyter.

Labov, William and Bettina Baker. 2003. What is a reading error? www.ling.upenn.edu/~wlabov/Papers/WRE.html] (last accesssed 27 July 2010).

Labov, William and Wendell A. Harris. 1994. Addressing social issues through linguistic evidence. In Gibbons (ed.), 265–305.

Labov, William, Mark Karen and Corey Miller. 1991. Near-mergers and the suspension of phonemic contrast. *Language Variation and Change* 3, 33–74.

Labov, William, Malcah Yaeger and Richard Steiner. 1972. *A Quantitative Study of Sound Change in Progress*. Philadelphia: US Regional Survey.

Ladd, F. C. 1970. Black youths view their environment. *Environment and Behaviour* 2, 74–99.

Laing, Margaret and Roger Lass. 2007. *A Linguistic Atlas of Early Middle English, 1150–1325*. The University of Edinburgh.

Lakoff, Robin. 1975. *Language and Woman's Place*. New York: Harper & Row.

Lambert, Wallace, E. R. Hodgson, R. C. Gardner and S. Fillenbaum. 1960. Evaluation reactions to spoken languages. *Journal of Abnormal and Social Psychology* 60, 44–51.

Language and National Origin Group. 2004. Guidelines for the use of language analysis in relation to questions of national origin in refugee cases. *The International Journal of Speech, Language and the Law* 11(2), 261–6.

Lavandera, Beatriz. 1978. Where does the sociolinguistic variable stop? *Language in Society* 7, 171–82.

Lave, Jean and Etienne Wenger. 1991. *Situated Learning: Legitimate Peripheral Participation*. Cambridge University Press.

Laver, John. 1980. *The Phonetic Description of Voice Quality*. Cambridge University Press.

Lawrence, Sophie, Francis Nolan and Kirsty McDougall. 2008. Acoustic and perceptual effects of telephone transmission on vowel quality. *International Journal of Speech, Language and the Law* 15(2), 161–92.

Lee, David. 2001–. *Bookmarks for Corpus-based Linguists*. Available online at http://personal.cityu.edu.hk/~davidlee/devotedtocorpora/CBLLinks.htm (last accessed 4 August 2010).

Leech, Geoffrey. 1993a. Corpus annotation schemes. *Literary and Linguistic Computing* 8, 275–81.

1993b. 100 million words of English. *English Today* 9, 9–15.

2003. Modality on the move: the English modal auxiliaries 1961–1992. In Facchinetti, Krug and Palmer (eds.), 223–40.

Levenshtein, Vladimir. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Cybernetics and Control Theory* 10(8), 707–10.

Levi, Judith and Anne Walker (eds.). 1990. *Language in the Judicial Process*. New York: Plenum Press.

Lewontin, Richard C. 1972. The apportionment of human diversity. *Evolutionary Biology* 6, 381–98.

Lightfoot, David. 1999. *The Development of Language: Acquisition, Change, and Evolution*. Oxford: Blackwell.

Lindblom, Björn. 1963. Spectrographic study of vowel reduction. *Journal of the Acoustical Society of America* 35, 1773–81.

Ling, Rich and Naomi Baron. 2007. Text messaging and IM: linguistic comparison of American college data. *Journal of Language and Social Psychology* 26, 291–8.

Lippi-Green, Ros. 1997. *English with an Accent: Language, Ideology and Discrimination in the United States*. New York: Routledge.

Llamas, Carmen. 1999. A new methodology: data elicitation for social and regional language variation studies. *Leeds Working Papers in Linguistics and Phonetics* 7, 95–118.

2007. A place between places: language and identities in a border town. *Language in Society* 36(4), 579–604.

Local, John and Gareth Walker. 2005. Methodological imperatives for investigating the phonetic organization and phonological structures of spontaneous speech. *Phonetica* 62, 120–30.

Lodge, Ken. 2009. *Fundamental Concepts in Phonology: Sameness and Difference*. Edinburgh University Press.

Loman, Bengt. 1975. Prosodic patterns in a Negro American dialect. In Håkan Ringbom, Alfhild Ingberg, Ralf Norrman, Kurt Nyholm, Rolf Westman and Kay Wikberg (eds.) *Style and Text: Studies Presented to Nils Erik Enkvist*, 219–42. Stockholm: Språkförlaget Skriptor AB.

Long, Daniel. 1999a. Geographical perception of Japanese dialect regions. In Preston (ed.), 177–98.

1999b. Mapping nonlinguists' evaluations of Japanese language variation. In Preston (ed.), 199–226.

Long, Daniel and Dennis R. Preston (eds.). 2002. *Handbook of Perceptual Dialectology*. Amsterdam: John Benjamins.

Long, Daniel and Young-Cheol Yim. 2002. Regional differences in the perception of Korean dialects. In Long and Preston (eds.), 249–75.

Longobardi, Giuseppe and Cristina Guardiano. 2009. Evidence for syntax as a signal of historical relatedness. *Lingua* 119, 1679–706.

Low, Ee Ling, Esther Grabe and Francis Nolan. 2000. Quantitative characterizations of speech rhythm: syllable-timing in Singapore English. *Language and Speech* 43, 377–401.

Lynch, Kevin. 1960. *The Image of the City*. Cambridge, MA: MIT Press.

Macafee, Caroline. 2004. A History of Scots to 1700. In *The Dictionary of the Scots Language*, www.dsl.ac.uk/ (last accessed 3 August 2010).

Macaulay, Ronald. 1991a. 'Coz it izny spelt when they say it': displaying dialect in writing. *American Speech* 66, 280–91.

1991b. *Locating Dialect in Discourse: The Language of Honest Men and Bonnie Lassies in Ayr*. New York: Oxford University Press.

MacEachern, Scott. 2000. Genes, tribes, and African history. *Current Anthropology*, 41, 357–84.

Macleod, Catherine, Nancy Ide and Ralph Grishman. 2000. The American National Corpus: a standardized resource for American English. *Proceedings of the Second Language Resources and Evaluation Conference*, Athens, Greece, 831–6.

MacWhinney, Brian. 2000. *The CHILDES Project: Tools for Analyzing Talk*, 3rd edn. Mahwah, NJ: Erlbaum.

Maguire, Warren. 2008. Quantifying dialect similarity by comparison of the lexical distribution of phonemes. *International Journal of Humanities and Arts Computing* 2(1–2), 261–77.

Maguire, Warren, April McMahon, Paul Heggarty and Dan Dediu. (2010). The past, present and future of English dialects: quantifying convergence, divergence and dynamic equilibrium. *Language Variation and Change* 22(1), 1–36.

Manning, Christopher, Prabhakar Raghavan and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

Manning, Christopher and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, MA: Cambridge University Press.

Markham, D. 1999. Listeners and disguised voices: the imitation of dialectal accent. *Forensic Linguistics* 6: 289–99.

Martinet, Andre. 1952. Function, structure, and sound change. *Word* 8, 1–32.

Maryns, Katrijn. 2006. *The Asylum Speaker: Language in the Belgian Asylum Procedure*. London: St Jerome Publishing.

Mase, Yoshio. 1999 (1964). Dialect consciousness and dialect divisions. In Preston (ed.), 71–99.

Mather, James Y. and Hans H. Speitel. 1986. *The Linguistic Atlas of Scotland: Scots Section. Volume III: Phonology*. London: Croom Helm.

McDonald, Christine and Joan C. Beal. 1987. Modal verbs in Tyneside English. *Journal of the Atlantic Provinces Linguistic Association* 9, 43–55.

McDougall, Kirsty. 2004. Speaker-specific formant dynamics: an experiment on Australian English /aI/. *The International Journal of Speech, Language and the Law* 11(1), 103–30.

McMahon, April. 1992. Underspecification theory and the analysis of dialect differences in lexical phonology. *Transactions of the Philological Society* 90, 81–119.

1994. *Understanding Language Change*. Cambridge University Press.

2000. *Lexical Phonology and the History of English*. Cambridge University Press.

McMahon, April, Paul Heggarty, Robert McMahon and Warren Maguire. 2007. The sound patterns of Englishes: representing phonetic similarity. *English Language and Linguistics* 11(1), 13–142.

McMahon, April and Robert McMahon. 1995. Linguistics, genetics, and archaeology: internal and external evidence in the Amerind controversy. *Transactions of the Philological Society* 93, 125–225.

2005. *Language Classification by Numbers*. Oxford University Press.

McMahon, Robert. 2004. Genes and languages. *Community Genetics* 7, 2–13.

McMenamin, Gerald. 2002a. Style markers in authorship studies. *The International Journal of Speech, Language and the Law* 8(2), 93–7.

2002b. *Forensic Linguistics: Advances in Forensic Stylistics*. Boca Raton: CRC Press.

Melchior, L., M. T. P. Gilbert, T. Kivisild, N. Lynnerup and J. Dissing. 2008. Rare mtDNA haplogroups and genetic differences in rich and poor Danish Iron-Age villages. *American Journal of Physical Anthropology* 135, 206–15.

Mellinkoff, David. 1963. *The Language of the Law*. Boston: Little, Brown and Company.

Mendoza-Denton, Norma. 2002. Language and identity. In Chambers, Trudgill and Schilling-Estes (eds.), 475–99.

2008. *Homegirls: Language and Cultural Practice Among Latina Youth Gangs*. Malden, MA: Wiley-Blackwell.

Mesthrie, Rajend (ed.). 2008. *Varieties of English 4: Africa, South and Southeast Asia*. Berlin/New York: Mouton de Gruyter.

Meyer, Charles. 2002. *English Corpus Linguistics: An Introduction*. Cambridge University Press.

2004. Can you really study language variation in linguistic corpora? *American Speech* 79, 339–55.

Meyer, Charles and Gerard Nelson. 2006. Data collection. In Bas Aarts and April McMahon (eds.) *The Handbook of English Linguistics*, 93–113. Malden: Blackwell Publishing.

Mildren, Dean. 1999. Redressing the imbalance: Aboriginal people in the criminal justice system. *Forensic Linguistics* 6(1), 83–109.

Miller, Jim and Regina Weinert. 1995. The function of *like* in dialogue. *Journal of Pragmatics* 23, 365–93.

Milroy, James. 1992. *Linguistic Variation and Change*. Oxford: Blackwell.

Milroy, James and Lesley Milroy. 1985. Linguistic change, social network, and speaker innovation. *Journal of Linguistics* 21, 339–84.

Milroy, Lesley. 1980. *Language and Social Networks*. Oxford: Basil Blackwell.

1987. *Observing and Analysing Natural Language*. Oxford: Blackwell.

Milroy, Lesley and Matthew Gordon. 2003. *Sociolinguistics: Method and Interpretation*. Malden, MA/Oxford, UK: Blackwell.

Milroy, Lesley, Jim Milroy and Gerard J. Docherty. 1997. Phonological variation and change in contemporary spoken British English. *SRC Unpublished Final Report*, Dept. of Speech, University of Newcastle upon Tyne, UK.

Mitkov, Ruslan. 2005. *The Oxford Handbook of Computational Linguistics*. Oxford University Press.

Modaressi, Yahya. 1978. A sociolinguistic investigation of modern Persian. Unpublished dissertation, University of Kansas.

Moeketsi, Rosemary. 1999. *Discourse in a Multilingual and Multicultural Courtroom: A Court Interpreter's Guide*. Pretoria: Van Schaik.

Moisl, Hermann. 2007. Data nonlinearity in exploratory multivariate analysis of language corpora. In John Nerbonne, Mark Ellison and Grzegorz Kondrak (eds.) *Computing and Historical Phonology. Proceedings of the Ninth Meeting of the ACL Special Interest Group in Computational Morphology and Phonology, June 28 2007*, 93–100. Association for Computational Linguistics, www.let.rug.nl/alfa/Prague/proceedings.pdf (last accessed 3 August 2010).

2008. Exploratory multivariate analysis. In Anke Lüdeling and Merja Kytö, (eds.) *Corpus Linguistics. An International Handbook*. Berlin: Mouton de Gruyter.

2010. Sura length and lexical probability estimation in cluster analysis of the Qur'an. *ACM Transactions on Asian Language Information Processing* (forthcoming).

Moisl, Hermann and Warren Maguire. 2008. Identifying the main determinants of phonetic variation in the Newcastle Electronic Corpus of Tyneside English. *Journal of Quantitative Linguistics* 15, 46–69.

Moisl, Hermann, Warren Maguire and Will Allen. 2006. Phonetic variation in Tyneside: exploratory multivariate analysis of the Newcastle Electronic Corpus of Tyneside English. In Frans Hinskens (ed.) *Language Variation – European Perspectives: Selected Papers from the Third International Conference on Language Variation in Europe (ICLaVE 3), Amsterdam, June 2005*, 127–41. Amsterdam: John Benjamins.

Montgomery, Chris. 2006. Northern English dialects: a perceptual approach. Unpublished Ph.D. thesis. University of Sheffield.

Moore, Emma. 2003. Learning style and identity: a sociolinguistic analysis of a Bolton high school. Unpublished Ph.D. dissertation, University of Manchester.

2004. Sociolinguistic style: a multidimensional resource for shared identity creation. *Canadian Journal of Linguistics* 49, 375–96.

Moore, Emma and Robert J. Podesva. 2009. Style, indexicality and the social meaning of tag questions. *Language in Society* 38(4), 447–85.

Morley, Barry. 2006. WebCorp: a tool for online linguistic information retrieval and analysis. In Antoinette Renouf and Andrew Kehoe (eds.) *The Changing Face of Corpus Linguistics*, 283–96. Amsterdam: Rodopi.

Moulton, William G. 1962. Dialect geography and the concept of phonological space. *Word* 18, 23–32.

Mourant, A. E., D. Tills and K. Domaniewska-Sobczak. 1976. *The Distribution of the Human Blood Groups and Other Polymorphisms*. London: Oxford University Press.

Mugglestone, Lynda. 2006. *The Oxford History of English*. Oxford University Press.

Murray, James. 1873. *The Dialects of the Southern Counties of Scotland: Its Pronunciation, Grammar, and Historical Relations with an Appendix on the Present Limits of the Gaelic and Lowland Scotch, and the Dialectical Divisions of the Lowland Tongue, and a Linguistical Map of Scotland*. London: Asher and Co.

Myres, John Nowell Linton. 1989. *The English Settlements*. Oxford University Press.

Nelson, Gerald. 1996. Markup systems. In Greenbaum (ed.), 36–53.

Nerbonne, John. 2006. Identifying linguistic structure in aggregate comparison. *Literary and Linguistic Computing* 21(4), 463–76.

Nerbonne, John and Wilbert Heeringa. 1997. Measuring dialect distance phonetically. In John Coleman (ed.) *Workshop on Computational Phonology, Special Interest*

Group of the Association for Computational Linguistics, Madrid, 1997, 11–18. Association for Computational Linguistics.

2001. Computational comparison and classification of dialects. *Dialectologia et Geolinguistica. Journal of the International Society for Dialectology and Geolinguistics* 9, 69–83.

Nerbonne, John, Wilbert Heeringa and Peter Kleiweg. 1999. Edit distance and dialect proximity. In Sankoff and Kruskal (eds.), v–xv.

Nerbonne, John and Erhard Hinrichs. 2006. Linguistic distances. In *Linguistic Distances. Workshop at the Joint Conference of International Committee on Computational Linguistics and the Association for Computational Linguistics, Sydney, July, 2006*, 1–6.

Nerbonne, John and Peter Kleiweg. 2007. Toward a dialectological yardstick. *Journal of Quantitative Linguistics* 14(2), 148–67.

Nerbonne, John and William A. Kretzschmar Jr. 2003. Introducing computational methods in dialectometry. In John Nerbonne and William A. Kretzschmar, Jr (eds.) *Computational Methods in Dialectometry.* Special issue of *Computers and the Humanities* 37(3), 245–55.

Nerbonne, John and William A. Kretschmar Jr. 2006. Progress in dialectometry: toward explanation. *Literary and Linguistic Computing* 21(4), 387–97.

Neuhauser, Sara. 2008. Voice disguise using a foreign accent: phonetic and linguistic variation. *International Journal of Speech, Language and the Law* 15(2), 131–59.

Newmeyer, Frederick. 2003. Grammar is grammar and usage is usage. *Language* 79, 682–707.

2005. A reply to the critiques of 'Grammar is grammar and usage is usage'. *Language* 81, 229–36.

Newton, Brian. 1972. *The Generative Interpretation of Dialect: A Study of Modern Greek Phonology*. Cambridge University Press.

Nichols, Joanna. 1997. Modeling ancient population structures and movements in Linguistics. *Annual Review of Anthropology* 26, 359–84.

Niedzielski, Nancy. 1999. The effect of social information on the perception of sociolinguistic variables. *Journal of Language and Social Psychology* 18, 62–85.

Niedzielski, Nancy and Dennis R. Preston. 2003. *Folk Linguistics*. Berlin: Mouton de Gruyter.

Nolan, Francis. 1983. *The Phonetic Bases of Speaker Recognition*. Cambridge University Press.

1999. Speaker recognition and forensic phonetics. In William J. Hardcastle and John Laver (eds.) *The Handbook of Phonetic Sciences*, 744–67. Oxford: Blackwell.

Nomoto, Kikuo. 1999 (1963). Consciousness of linguistic boundaries and actual linguistic boundaries. In Preston (ed.), 63–9.

O'Barr, William. 1982. *Linguistic Evidence: Language, Power, and Strategy in the Courtroom*. New York: Academic Press.

1993. Professional varieties: the case of language and law. In Dennis R. Preston (ed.) *American Dialect Research*, 319–29. Amsterdam: John Benjamins.

Ochs, Elinor. 1979. Transcription as theory. In Elinor Ochs and Bambi Schieffelin (eds.) *Developmental Pragmatics*, 43–72. New York, Academic.

1991. Indexing gender. In Alessandro Duranti and Charles Goodwin (eds.) *Rethinking Context: Language as an Interactive Phenomenon*, 335–58. Cambridge University Press.

Olsson, John. 2008. *Forensic Linguistics*. London: Continuum.
Oppenheimer, Stephen. 2006. *The Origins of the British: A Genetic Detective Story*. London: Constable.
Orleans, Peter. 1967. Differential cognition of urban residents: effects of social scale on mapping. In John G. Truxal (ed.) *Science, Engineering, and the City*. Washington DC: National Academy of Sciences.
1973. Differential cognition of urban residents: effects of social scale on mapping. In Roger M. Downs and D. Stea (eds.) *Image and Environment: Cognitive Mapping and Spatial Behaviour*. Chicago, IL: Aldine Publishing Company. 115–30.
Orton, Harold. 1962. *Survey of English Dialects: An Introduction*. Leeds: E. J. Arnold and Son.
Orton, Harold and Eugen Dieth. 1962–71. *Survey of English Dialects: Basic Material*. University of Leeds Press.
Orton, Harold, Michael V. Barry, Wilfrid J. Halliday, Philip M. Tilling and Martyn F. Wakelin. 1962–71. *Survey of English Dialects*, 4 volumes. Leeds: E. J. Arnold and Son.
Orton, Harold, Stewart Sanderson and John Widdowson (eds.). 1962–71. *Survey of English Dialects (B): The Basic Material*. Leeds: Arnold and Son.
Orton, Harold, Stewart Sanderson and John Widdowson (eds.). 1978. *The Linguistic Atlas of England*. London: Croom Helm.
Page, Roderic D. M. and Edward C. Holmes. 1998. *Molecular Evolution: A Phylogenetic Approach*. Oxford: Blackwells.
Paltridge, J. and H. Giles. 1984. Attitudes towards speakers of regional accents of French: effects of rationality, age and sex of listeners. *Linguistische Berichte* 90, 71–85.
Påhlsson, Christer. 1972. *The Northumbrian Burr: A Sociolinguistic Study*. Lund: Gleerup.
Pashler, Harold E. (ed.). 2002. *Steven's Handbook of Experimental Psychology, Volumes I–IV*, 3rd edn. New Jersey: John Wiley and Sons.
Patrick, Peter. 2001. The speech community. In Chambers, Trudgill and Schilling-Estes (eds.), 573–99.
Pederson, Lee A., Susan Leas McDaniel, Guy Bailey, Marvin H. Basset, Carol M. Adams, Caisheng Liao and Michael B. Montgomery (eds.). 1986–92. *The Linguistic Atlas of the Gulf States*. 7 vols. Athens: University of Georgia Press.
Peeters, Wilhelmus Johannes Maria. 1991. Diphthong dynamics: a cross-linguistic perceptual analysis of temporal patterns in Dutch, English, and German. Ph.D. dissertation, Reiksuniversiteit te Utrecht.
Pellowe, John and Val M. Jones. 1978. On intonational variability in Tyneside speech. In Peter Trudgill (ed.) *Sociolinguistic Patterns in British English*, 101–21. London: Arnold.
Pellowe, John, Barbara Strang, Graham Nixon and Vince McNeany. 1972. A dynamic modelling of linguistic variation: the urban (Tyneside) linguistic survey. *Lingua* 30, 1–30.
Pereira, Luísa, Martin Richards, Ana Goios, *et al*. 2005. High-resolution mtDNA evidence for the late-glacial resettlement of Europe from an Iberian refugium. *Genome Research* 15, 19–24.
Pesetsky, David. 1997. Some optimality principles of sentence pronunciation. In Pilar Barbosa, Danny Fox, Paul Hagstrom, Martha McGinnis and David Pesetsky (eds.) *Is the Best good Enough? Optimality and Competition in Syntax*, 337–83. Cambridge, MA: MIT Press.

Pettersson, Rune. 2002. *Information Design: An Introduction*. Amsterdam: John Benjamins.

Philbrick, Frederick. 1949. *Language and the Law: The Semantics of Forensic English*. New York: Macmillan.

Phillips, Betty. 2006. *Word Frequency and Lexical Diffusion*. New York: Palgrave Macmillan.

Pierrehumbert, Janet. 2002. Word-specific phonetics. In Carlos Gussenhoven and Natasha Warner (eds.) *Laboratory Phonology* VII, 101–40. Berlin: Mouton de Gruyter.

Pietsch, Lukas. 2005a. *Variable Grammars. Verbal Agreement in Northern Dialects of English*. Tübingen: Max Niemeyer.

2005b. 'Some do and some doesn't': verbal concord variation in the north of the British Isles. In Bernd Kortmann, Tania Hermann, Lukas Pietsch and Susanne Wagner (eds.) *A Comparative Grammar of English Dialects: Agreement, Gender, Relative Clauses*, 125–210. Berlin/New York: Mouton de Gruyter.

Plain English Campaign. 2008. Plain English Campaign: fighting for crystal clear communication since 1979. www.plainenglish.co.uk (last accessed 5 August 2010).

Plichta, Bartlomiej. 2006. Interdisciplinary perspectives on the northern cities chain Shift. Ph.D. dissertation, Michigan State University.

Podesva, Robert J. 2006. Phonetic detail in sociolinguistic variation: its linguistic significance and role in the construction of social meaning. Unpublished Ph.D. dissertation, Stanford University.

2007. Phonation type as a stylistic variable: the use of falsetto in constructing a persona. *Journal of Sociolinguistics* 11, 478–504.

2008. Three sources of stylistic meaning. *Texas Linguistic Forum* (Proceedings of the Symposium About Language and Society – Austin 15) 51, 1–14.

Podesva, Robert J. and Elaine Chun. 2007. On indeterminacy in the social meaning of variation. Paper presented at UK Language Variation and Change 6, Lancaster.

Poole, David. 2005. *Linear Algebra: A Modern Introduction*. Florence KY: Brooks Cole.

Poplack, Shana. 1989. The care and handling of a mega-corpus: the Ottawa-Hull French Project. In Fasold and Schiffrin (eds.), 411–51.

2007. Foreword. In Beal, Corrigan and Moisl (eds.), Vol. I, ix–xiii.

Poplack, Shana and Sali A. Tagliamonte. 2001. *African American English in the Diaspora: Tense and Aspect*. Oxford: Blackwell Publishers.

Poplack, Shana, James Walker and Rebecca Malcolmson. 2006. An English 'like no other'? Language contact and change in Quebec. In Avery, Chambers, D'Arcy, Gold and Rice (eds.), 185–213.

Popper, Karl. 1959. *The Logic of Scientific Discovery*. New York: Basic Books.

1963. *Conjectures and Refutations: The Growth of Scientific Knowledge*. Florence KY: Routledge/Taylor and Francis Group.

Powers, Ashley and Harriet Ryan. 2008. Audio expert testifies in O. J. Simpson robbery-kidnap trial. *Los Angeles Times*, 18 September 2008. Available online at http://articles.latimes.com/2008/sep/18/nation/na-oj18 (last accessed 5 August 2010).

Preston, Dennis R. 1981. Perceptual dialectology: mental maps of United States dialects from a Hawaiian perspective (summary). In H. Warkentyne (ed.) *Methods IV (Papers from the Fourth International Conference on Methods in Dialectology)*, 192–8. British Columbia.

1985. The li'l abner syndrome: written representations of speech. *American Speech* 60, 328–36.
1988. Change in the perception of language varieties. In Jacek Fisiak (ed.) *Historical Dialectology: Regional and Social*. Berlin: Mouton de Gruyter.
1989. *Perceptual Dialectology: Non-Linguists' View of Aerial Linguistics*. Dordrecht: Foris.
Preston, Dennis R. (ed.). 1993. *American Dialect Research*. Amsterdam: John Benjamins.
Preston, Dennis R. (ed.). 1999a. *Handbook of Perceptual Dialectology*. Amsterdam: John Benjamins.
Preston, Dennis R. 1999b. Introduction. In Preston (ed.), xxiii–xxxix.
1999c. A language attitude approach to the perception of regional variety. In Preston (ed.), 359–75.
2000. Mowr and mowr bayud spellin': confessions of a sociolinguist. *Journal of Sociolinguistics* 4, 614–21.
2002a. Language with an attitude. In Chambers, Trudgill and Schilling-Estes (eds.), 40–66.
2002b. Perceptual dialectology: aims, methods, findings. In Jan Berns and Jaap van Marle (eds.) *Present-day Dialectology: Problems and Findings*, 57–104. Berlin: Mouton de Gruyter.
Prince, Alan and Paul Smolensky. 1993/2004. *Optimality Theory: Constraint Interaction in Generative Grammar*. Rutgers University and University of Colorado at Boulder. Published 2004. Oxford: Blackwell.
Prokić, Jelena and John Nerbonne. 2008. Recognizing groups among dialects. *International Journal of Humanities and Arts Computing*, Special Issue on Language Variation, ed. by Nerbonne, John, Charlotte Gooskens, Sebastian Kürschner and Renée van Bezooijen, 153–72.
Purnell, Thomas, Joseph Salmons and Dilara Tepeli. 2005. German substrate effects in Wisconsin English: evidence for final fortition. *American Speech* 80, 135–64.
Purnell, Thomas, Joseph Salmons, Dilara Tepeli and Jennifer Mercer. 2005. Structured heterogeneity and change in laryngeal phonetics. *Journal of English Linguistics* 33, 307–38.
Pyle, Dorian. 1999. *Data Preparation for Data Mining*. San Francisco: Morgan Kaufmann.
Quirk, Randolf. 1968. The survey of English usage. In *Essays on the English Language: Medieval and Modern*, 70–87. London: Longman.
Rae, Megan and Paul Warren. 2002. Goldilocks and the three beers: sound merger and word recognition in NZE. *New Zealand English Journal* 16, 33–41.
Rampton, Ben. 1992. Scope for empowerment in sociolinguistics? In Deborah Cameron (ed.) *Researching Language: Issues of Language, Power and Method*, 29–64. London: Routledge.
2005. *Crossing: Language and Ethnicity among Adolescents*, 2nd edn. Oxford: Blackwell.
2006. *Language in Late Modernity: Interaction in an Urban School*. Cambridge University Press.
Rauniomaa, Mirka. 2003. Stance accretion. Paper presented at the Language, Interaction, and Social Organization Research Focus Group, University of California, Santa Barbara.

Reaser, Jeffrey and Carolyn Temple Adger. 2008. Vernacular language varieties in educational settings: research and development. In Spolsky and Hult (eds.), 161–73.

Relethford, John H. 2001. *Genetics and the Search for Modern Human Origins*. New York: Wiley-Liss.

Renfrew, Colin. 1987. *Archaeology and Language*. London: Jonathan Cape.

Renouf, Antoinette. 1993. A word in time: first findings from the investigation of dynamic text. In Jan Aarts, Pieter de Haan and Nelleke Oostdijk (eds.) *English Language Corpora: Design, Analysis and Exploitation. Papers from the Thirteenth International Conference on English Language Research on Computerized Corpora, Nijmegen 1992*, 279–88. Amsterdam: Rodopi.

2003. WebCorp: providing a renewable data source for corpus linguists. In Sylviane Granger and Stephanie Petch-Tyson (eds.) *Extending the Scope of Corpus-based Research: New Applications, New Challenges*, 39–58. Amsterdam: Rodopi.

Rensink, W. G. 1999 (1955). Informant classification of dialects. In Preston (ed.), 3–7.

Richards, J. C., J. Platt and H. Platt. 1992. *Longman Dictionary of Language Teaching and Applied Linguistics*. Harlow: Longman.

Richards, Martin, Vincent Macaulay, Eileen Hickey, *et al.* 2000. Tracing European founder lineages in the Near Eastern mtDNA pool. *American Journal of Human Genetics* 67, 1251–76.

Rickford, John. 2002. Linguistics, education and the Ebonics firestorm. In James E. Alatis, Heidi E. Hamilton and A.-H. Tan (eds.) *Linguistics, Language and the Professions*, 25–45. Washington, DC: Georgetown University Press.

Rickford, John and Angela Rickford. 1995. Dialect readers revisited. *Linguistics and Education* 7, 107–28.

Roberts, Leslie. 1992. Genome diversity project: anthropologists climb (gingerly) on board. *Science* 258, 1300–1.

Rock, Frances. 2001. The genesis of a witness statement. *Forensic Linguistics* 8(2), 44–72.

2007. *Communicating Rights: The Language of Arrest and Detention*. Basingstoke: Palgrave Macmillan.

Romaine, Suzanne. 2005. Variation in language and gender. In Janet Holmes and Miriam Meyerhoff (eds.) *The Handbook of Language and Gender*, 98–118. Malden, MA, Blackwell.

Romaine, Suzanne and Deborah Lange. 1991. The use of like as a marker of reported speech and thought: a case of grammaticalization in progress. *American Speech* 66, 227–79.

Romesburg, H. Charles. 1984. *Cluster Analysis for Researchers*. Florence, KY: Wadsworth.

Rose, Mary. 2006. Language, place, and identity in later life. Unpublished Ph.D. dissertation, Stanford University.

Rose, Mary and Devyani Sharma. 2002. Introduction: ideology and identity in practice. In Sarah Benor, Mary Rose, Devyani Sharma, Julie Sweetland and Qing Zhang (eds.) *Gendered Practices in Language*, 1–20. Stanford, CA: CSLI Publications.

Rose, Philip. 2002. *Forensic Speaker Identification*. London: Taylor and Francis.

Rosenberg, Noah A., Saurabh Mahajan, Sohini Ramachandran, Chengfeng Zhao, Jonathan K. Pritchard and Marcus W. Feldman. 2005. Clines, clusters, and the effect of study design on the inference of human population structure. *PLoS Genetics* 1, 660–71.

Rosenberg, Noah A. and Magnus Nordborg. 2002. Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nature Reviews Genetics* 3, 380–90.
Russell, Dave. 2004. *Looking North: Northern England and the National Imagination*. Manchester University Press.
Ryan, Ellen and Howard Giles (eds.). 1982. *Attitudes Towards Language Variation*. Edward Arnold: London.
Sacks, Harvey, Emanual Schegloff and Gail Jefferson. 1974. A simplest systematics for the organization of turn-taking in conversation. *Language* 50, 696–735.
Sankoff, David. 1978. Probability and linguistic variation. *Synthese* 37, 217–38.
1988. Sociolinguistics and syntactic variation. In Frederick Newmeyer (ed.) *Linguistics: The Cambridge Survey*, 140–61. Cambridge University Press.
2005. Problems of representativeness. In Ammon, Dittmar, Mattheier and Trudgill (eds.), 998–1002.
Sankoff, David and Joseph Kruskal (eds.). 1999. *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*. Stanford, CA: CSLI.
Sankoff, David and Gillian Sankoff. 1973. Sample survey methods and computer-assisted analysis in the study of grammatical variation. In Regna Darnell (ed.) *Canadian Languages in their Social Context*, 7–63. Edmonton: Linguistic Research Inc.
Sankoff, Gillian. 2004. Adolescents, young adults and the critical period: two case studies from Seven Up. In Carmen Fought (ed.) *Sociolinguistic Variation: Critical Reflections*, 121–39. Oxford University Press.
Sankoff, Gillian and Hélène Blondeau. 2007. Longitudinal change across the life-span: /r/ in Montreal French. *Language* 83(3), 560–88.
Santa Ana, Otto. 1996. Sonority and syllable structure in Chicano English. *Language Variation and Change* 8, 63–89.
Santorini, Beatrice. 1993. The rate of phrase structure change in the history of Yiddish. *Language Variation and Change* 5, 257–83.
Saussure, Ferdinand de. 1916. *Cours de linguistique générale*. Edited by Charles Bally and Albert Sechehaye, with the collaboration of Albert Riedlinger. Paris: Payot.
1986. *Course in General Linguistics*. Translated by Roy Harris. La Salle, IL: Open Court.
Scheler, Manfred. 1977. *Der englische Wortschatz*. Berlin: Erich Schmidt.
Schilling-Estes, Natalie. 2002. Investigating stylistic variation. In Chambers, Trudgill and Schilling-Estes (eds.), 375–401.
2004. Constructing ethnicity in interaction. *Journal of Sociolinguistics* 8, 163–95.
Schilling-Estes, Natalie and Walt Wolfram. 1994. Convergent explanation and alternative regularization patterns: *were/weren't* leveling in a vernacular English variety. *Language Variation and Change* 6, 273–302.
Schneider, Edgar. 2002. Investigating variation and change in written documents. In Chambers, Trudgill and Schilling-Estes (eds.), 67–96.
Schneider, Edgar (ed.). 2008. *Varieties of English 2: The Americas and the Caribbean*. Berlin/New York: Mouton de Gruyter.
Schütze, Carson T. 1996. *The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology*. University of Chicago Press.
1999. English expletive constructions are not infected. *Linguistic Inquiry* 30(3), 467–84.

Séguy, Jean. 1971. La relation entre la distance spatiale et la distance lexicale. *Revue de Linguistique Romane* 35, 335–57.
1973. La dialectométrie dans l'Atlas linguistique de la Gascogne. *Revue de Linguistique Romane* 37, 1–24.
Shackleton, Robert. 2005. English-American speech relationships: a quantitative approach. *Journal of English Linguistics* 33(2), 99–160.
2007. Phonetic variation in the traditional English dialects: a computational analysis. *Journal of English Linguistics* 35(1), 30–102.
Shastri, S. V. 1988. The Kolhapur Corpus of Indian English and work done on its basis so far. *ICAME Journal* 12, 15–26.
Shuy, Roger. 1993. *Language Crimes: The Use and Abuse of Language Evidence in the Courtroom*. Oxford: Blackwell.
1998. *The Language of Confession, Interrogation and Deception*. Thousand Oaks, London: Sage Publications.
2001. Discourse analysis in the legal context. In Deborah Schiffrin, Deborah Tannen and Heidi Hamilton (eds.) *The Handbook of Discourse Analysis*, 437–52. Oxford: Blackwell.
2002. *Linguistic Battles in Trademark Disputes*. Basingstoke: Palgrave Macmillan.
2003. The language problems of minorities in the legal setting. In Ceil Lucas (ed.) *Language and the Law in Deaf Communities*, 1–20. Washington, DC: Gallaudet University Press.
2005. *Creating Language Crimes: How Law Enforcement Uses (and Misuses) Language*. Oxford University Press.
2006a. Discourse analysis and the law. In Ammon, Dittmar, Klaus and Trudgill (eds.), 2323–33.
2006b. *Linguistics in the Courtroom: A Practical Guide*. Oxford University Press.
2007. Language in the American courtroom. *Language and Linguistics Compass* 1(1–2), 100–14.
Sibata, Takesi. 1999 (1959). Consciousness of dialect boundaries. In Preston (ed.), 39–63.
Siegel, Jeff. 1999. Creoles and minority dialects in education: an overview. *Journal of Multilingual and Multicultural Development* 20, 508–31.
2007. Creoles and minority dialects in education: an update. *Language and Education* 21, 66–86.
Siewierska, Anna and Willem Hollmann. 2007. Ditransitive clauses in English with special reference to Lancashire dialect. In Mike Hannay and Gerard Steen (eds.) *Structural-Functional Studies in English Grammar*, 83–102. Amsterdam: John Benjamins.
Silverstein, Michael. 1976. Shifters, linguistic categories, and cultural description. In Keith Basso and Henry Selby (eds.) *Meaning in Anthropology*, 11–55. Albuquerque: University of New Mexico Press.
2003. Indexical order and the dialectics of sociolinguistic life. *Language and Communication* 23, 193–229.
Simpson, Rita C., Sarah L. Briggs, Janine Ovens and John M. Swales. 2002. *The Michigan Corpus of Academic Spoken English*. Ann Arbor, MI: The Regents of the University of Michigan.
Sims-Williams, Patrick. 1998. Genetics, linguistics, and prehistory: thinking big and thinking straight. *Antiquity* 72, 505–27.
Sinclair, John. 1991. *Corpus, Concordance, Collocation*. Oxford University Press.

1992. The automatic analysis of corpora. In Svartvik (ed.), 379–97.
1997. *English Grammar* (Collins COBUILD English Grammar). London: Harper Collins.
Singh, Ishtla. 2005. *The History of English*. London: Hodder Arnold.
Smith, Jeremy. 2000. Synchrony and diachrony in the evolution of English: evidence from Scotland. Unpublished Ph.D. thesis, University of York.
Smith, Nicholas. 2003. Changes in the modals and semi-modals of strong obligation and epistemic necessity in recent British English. In Facchinetti, Krug and Palmer (eds.), 241–66.
Sneath, Peter and Robert Sokal. 1973. *Numerical Taxonomy: The Principles and Practice of Numerical Classification*. San Francisco, CA: W. H. Freeman and Company.
Snell, Julia. 2008. Pronouns, dialect and discourse: a socio-pragmatic account of children's language in Teesside. Unpublished Ph.D. thesis, University of Leeds.
Solan, Lawrence and Peter Tiersma. 2005. *Speaking of Crime: The Language of Criminal Justice*. University of Chicago Press.
Spolsky, Bernard and Francis Hult (eds.). 2008. *The Handbook of Educational Linguistics*. Oxford: Blackwell.
Spruit, Marco René. 2006. Measuring syntactic variation in Dutch dialects. In John Nerbonne and William A. Kretzschmar, Jr (eds.) *Progress in Dialectometry: Toward Explanation. Special issue of Literary and Linguistic Computing* 21(4), 493–506.
Stenton, Frank M. 1947. *Anglo-Saxon England*, Oxford History of England, Vol. II, 2nd edn. Oxford: Clarendon Press.
Stone, Linda, Paul Lurquin and Luigi Luca Cavalli-Sforza. 2007. *Genes, Culture, and Human Evolution: A Synthesis*. Oxford: Blackwell.
Strachan, Tom and Andrew P. Read. 2004. *Human Molecular Genetics*, 3rd edn. London: Garland Science.
Strang, Barbara. 1968. The Tyneside Linguistic Survey. *Zeitschrift für Mundartforschung*, Neue Folge 4, 788–94.
Stringer, Chris. 2006. *Homo britannicus: The Incredible Story of Human Life in Britain*. London: Penguin.
Stuart-Smith, Jane. 1999. Glasgow: accent and voice quality. In Foulkes and Docherty (eds.), 203–22.
Stuart-Smith, Jane, Claire Timmins and Fiona Tweedie. 2007. 'Talkin' Jockney?' Variation and change in Glaswegian accent. *Journal of Sociolinguistics* 11, 221–60.
Stubbs, Michael. 1996. *Text and Corpus Analysis Computer-Assisted Studies of Language and Culture*. Oxford: Blackwell.
Svartvik, Jan. 1968. *The Evans Statements. A Case for Forensic Linguistics*. Göteborg: University of Göteborg.
Svartvik, Jan (ed.). 1992. *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82, Stockholm, 4–8 August 1991*. Berlin: Mouton de Gruyter.
Swadesh, Morris. 1950. Salish internal relationships. *International Journal of American Linguistics* 16, 157–67.
1952. Lexico-statistic dating of prehistoric ethnic contacts. *Proceedings of the American Philosophical Society* 96, 453–63.
1955. Towards a greater accuracy in lexicostatistic dating. *International Journal of American Linguistics* 21, 121–37.

Syrdal, Ann K. 1996. Acoustic variability in spontaneous conversational speech of American English talkers. Paper presented at ICSLP 96 (Fourth International Conference on Spoken Language Processing), Philadelphia, October 3–6. www.asel.udel.edu/icslp/cdrom/vol1/582/a582.pdf (last accessed 1 August 2010).

Szmrecsanyi, Benedikt. 2008. Corpus-based dialectometry: aggregate morphosyntactic variability in British English dialects. *International Journal of Humanities and Arts Computing* 2(1–2), 279–96.

Tagliamonte, Sali A. 1998. *Was/were* variation across the generations: view from the city of York. *Language Variation and Change* 10, 153–91.

2002. Comparative sociolinguistics. In Chambers, Trudgill and Schilling-Estes (eds.), 729–63.

2006a. *Analysing Sociolinguistic Variation*. Cambridge University Press.

2006b. So cool, right?: Canadian English entering the 21st century. In Avery, Chambers, D'Arcy, Gold and Rice (eds.), 309–31.

2007. Representing real language: consistency, trade-offs, and thinking ahead! In Beal, Corrigan and Moisl (eds.), Vol. I, 205–40.

2008. Sometimes there's universals; sometimes there aren't – a comparative sociolinguistic perspective on 'default singulars'. In Markku Filppula, Juhani Klemola and Heli Paulasto (eds.) *Vernacular Universals and Language Contacts: Evidence from Varieties of English and Beyond*, 103–32. London: Taylor and Francis/Routledge.

Tagliamonte, Sali A. and Alexandra D'Arcy. 2009. Peaks beyond phonology: adolescence, incrementation, and language change. *Language* 85, 55–105.

Tagliamonte, Sali A. and Derek Denis. 2008. Linguistic ruin? LOL! Instant messaging and teen language. *American Speech* 83, 3–34.

Tagliamonte, Sali A. and Jennifer Smith. 2006. Layering, competition and a twist of fate: deontic modality in dialects of English. *Diachronica* 23, 341–80.

Tagliamonte, Sali A., Jennifer Smith and Helen Lawrence. 2005. No taming the vernacular! Insights from the relatives in northern Britain. *Language Variation and Change* 17(1), 75–112.

Tarallo, Fernando. 1996. Turning different at the turn of the century: 19th century Brazilian Portuguese. In Guy, Feagin, Schiffrin and Baugh (eds.), 199–220.

Tarone, Elaine E. 1973. Aspects of intonation in Black English. *American Speech* 48, 29–36.

Taylor, Ann. 1994. The change from SOV to SVO in Ancient Greek. *Language Variation and Change* 6, 1–37.

Taylor, James B. and Howard A. Parker. 1964. Graphic ratings and attitude measurement: a comparison of research tactics. *Journal of Applied Psychology* 48(1), 37–42.

Thagard, Paul. 2005. *Mind: Introduction to Cognitive Science*, 2nd edn. Cambridge MA: MIT Press.

Thibault, Pierrette and Diane Vincent. 1990. *Un corpus de français parlé*. Quebec City: Université Laval.

Thomas, Erik R. 2000. Spectral differences in /ai/ offsets conditioned by voicing of the following consonant. *Journal of Phonetics* 28, 1–25.

2001. *An Acoustic Analysis of Vowel Variation in New World English*. Publication of the American Dialect Society 85. Durham, NC: Duke University Press.

2002. Sociophonetic applications of speech perception experiments. *American Speech* 77, 115–47.

2003. Secrets revealed by southern vowel shifting. *American Speech* 78, 150–70.

Thomas, Erik R. and Phillip M. Carter. 2006. Prosodic rhythm and African American English. *English World-Wide* 27, 331–55.
Thomason, Sarah Grey. 2005. Contact as a source of language change. In Brian D. Joseph and Richard D. Janda (eds.) *The Handbook of Historical Linguistics*, 687–712. Oxford: Blackwell.
Thomason, Sarah Grey and Terrence Kaufman. 1988. *Language Contact, Creolization, and Genetic Linguistics*. Berkeley, CA: University of California Press.
Thompson, Joanna K. 2002. 'Powerful/powerless' language in court: a critical re-evaluation of the Duke Language and Law Programme. *Forensic Linguistics* 9(2), 153–67.
Tiersma, Peter M. 1999. *Legal Language*. Chicago, London: University of Chicago Press.
Tôjô, Misao. 1927. *The Great Japanese Map of Dialect Divisions*. Tokyo: Ikvei Shoin.
Toon, Thomas. 1983. *The Politics of Early Old English Sound Change*. New York: Academic Press.
Torroni Antonio, Alessandro Achilli, Vincent Macaulay, Martin Richards and Hans-Jürgen Bandelt. 2006. Harvesting the fruit of the human mtDNA tree. *Trends in Genetics* 22(6), 339–45.
Trager, George and Henry Smith. 1951. *An Outline of English Structure*. Battenburg: Norman.
Trail, Anthony, Martin J. Ball and Nicole Müller. 1995. Perceptual confusion between South African and British English vowels. In Kjell Elenius and Peter Branderud (eds.) *Proceedings of the 13th International Conference of Phonetic Sciences, ICPhS 95*: Stockholm, Sweden, 13–19 August 1995, 620–3. Stockholm: Royal Institute of Technology and Stockholm University.
Trask, Robert L. 1996. *Historical Linguistics*. London: Arnold.
Traugott, Elizabeth Closs. 2001. Zeroing in on multifunctionality and style. In Eckert and Rickford (eds.), 127–36.
Tremblay, Marc and Hélène Vezina. 2000. New estimates of intergenerational time intervals for the calculation of age and origins of mutations. *American Journal of Human Genetics* 66, 651–8.
Trousdale, Graeme and David Adger (eds.). 2007. Special issue of *English Language and Linguistics* 11(2). Cambridge University Press.
Trubetzkoy, Nikolaj. 1931. Phonologie et géographie linguistique. *Travaux du Cercle Linguistique de Prague* 4, 228–34.
Trudgill, Peter. 1974. *The Social Differentiation of English in Norwich*. Cambridge University Press.
1975. *Accent, Dialect and the School*. London: Arnold.
1983. *On Dialect: Social and Geographical Perspectives*. New York University Press.
Trudgill, Peter (ed.). 1984. *Language in the British Isles*. Cambridge University Press.
Trudgill, Peter. 1986. *Dialects in Contact*. Oxford: Blackwell.
1990. *The Dialects of England*. Oxford: Blackwell.
1999. *The Dialects of England*, 2nd edn. Oxford: Blackwell.
2001. On the irrelevance of prestige, stigma and identity in the development of New Zealand English phonology. *New Zealand English Journal* 15, 42–6.
Trudgill, Peter, Elizabeth Gordon, Gillian Lewis and Margaret Maclagan. 2000. Determinism in new-dialect formation and the genesis of New Zealand English. *Journal of Linguistics* 36, 299–318.

Turrell, M. Teresa. 2004. Textual kidnapping revisited: the case of plagiarism in literary translation. *Forensic Linguistics* 11(1), 1–26.
Udofot, Inyang. 2003. Stress and rhythm in the Nigerian accent of English: a preliminary investigation. *English World-Wide* 24, 201–20.
Underhill, Robert 1988. Like is like, focus. *American Speech* 63, 234–46.
Upton, Clive and John Widdowson. 2006. *An Atlas of English Dialects*. London: Routledge.
Van Coetsem, Frans. 1988. *Loan Phonology and the Two Transfer Types in Language Contact*. Dordrecht: Foris.
de Varennes, Fernand. 2003. Language rights and human rights: the international experience. In Ó Dónall Riagáin (ed.) *Language and Law in Northern Ireland*, 5–16. Belfast: Queen's University.
Vaux, Bert. 2008. Why the phonological component must be serial and rule-based. In Bert Vaux and Andrew Nevins (eds.) *Rules, Constraints, and Phonological Phenomena*, 20–61. Oxford University Press.
Viereck, Wolfgang and Heinrich Ramisch. 1997. *The Computer Developed Linguistic Atlas of England 2*. Tübingen: Niemeyer.
Vine, Bernadette, Gary Johnson and Janet Holmes. 1998. *Guide to the Wellington Corpus of Spoken New Zealand English*. Wellington, Victoria: University of Wellington.
Voronoi, Georgy. 1907. Nouvelles applications des paramètres continus à la théorie des formes quadratiques. *Journal für die Reine und Angewandte Mathematik* 133, 97–178.
Wagner, Anne and Sophie Cacciaguidi-Fahy. 2008. *Obscurity and Clarity in the Law: Prospects and Challenges*. Aldershot: Ashgate.
Wagner, Suzanne Evans. 2007. 'We act like girls and we don't act like men': the use of the male-associated variable (ay0) in South Philadelphia. *University of Pennsylvania Working Papers in Linguistics* 13(1), 1–14.
Wakelin, Martyn. 1984a. Rural dialects in England. In Trudgill (ed.), 70–93.
1984b. Cornish English. In Trudgill (ed.), 195–8.
Wales, Katie. 1999. North and south: a linguistic divide? *Reporter* 439.
2006a. Dialects in mental contact: a critique of perceptual dialectology. In Herbert Grabes and Wolfgang Viereck (eds.) *The Wider Scope of English*, 57–66. Frankfurt: Peter Lang.
2006b. *Northern English: A Social and Cultural History*. Cambridge University Press.
2010. Northern English transported: the nineteenth-century goldrushes and the formation of a diaspora? In B. Heselwood and C. Upton (eds.) *Papers from the Thirteenth International Conference on Methods in Dialectology*. Bamburg: Peter Lang.
Walsh, Michael. 1994. Interactional styles in the courtroom: an example from Northern Australia. In Gibbons (ed.), 217–33.
Walton, Julie H. and Robert F. Orlikoff. 1994. Speaker race identification from acoustic cues in the vocal signal. *Journal of Speech and Hearing Research* 37, 738–45.
Wang, William S.-Y. (ed.). 1977. *The Lexicon in Phonological Change*. The Hague: Mouton.
Warren, Paul. 2005. Patterns of late rising in New Zealand English: intonational variation or intonational change? *Language Variation and Change* 17, 209–30.
Watson, Kevin. 2007. The phonetics and phonology of plosive lenition in Liverpool English. Unpublished Ph.D. thesis, Edge Hill College/Lancaster University.
Watt, Dominic and Jennifer Smith. 2005. Language change. In Martin Ball (ed.) *Clinical Sociolinguistics*, 101–19. Oxford: Blackwell.

Weale, Michael E., Deborah A. Weiss, Rolf F. Jager, Neil Bradman and Mark G. Thomas. 2002. Y Chromosome evidence for Anglo-Saxon mass migration. *Molecular Biology and Evolution* 19, 1008–21.

Weijnen, Antonius A. 1946. De grenzen tussen de Oost-Noordbrabantse dialecten onderling. In Antonius A. Weijnen, J. M. Renders and Jac. Von Ginneken (eds.) *Oost-Noordbrabantse dialecteproblemen*, 1–15.

1999. On the value of subjective dialect boundaries. In D. R. Preston (ed.) *Handbook of Perceptual Dialectology*, 131–4. Amsterdam: John Benjamins.

Weinreich, Uriel, William Labov and Marvin Herzog. 1968. Empirical foundations for a theory of language change. In Winfred P. Lehman and Yakov Malkiel (eds.) *Directions for Historical Linguistics: A Symposium*, 95–188. Austin, TX: University of Texas Press.

Wells, John C. 1982. *Accents of English*. 3 vols. Cambridge University Press.

Wenger, Etienne. 1998. *Communities of Practice: Learning, Meaning, and Identity*. Cambridge University Press.

Williams, Angie, Peter Garret and Nikolas Coupland. 1999. Dialect recognition. In Preston (ed.), 345–58.

Williamson, John. 1990. 'Divvent write that, man': the influence of Tyneside dialect forms on children's free writing. *Educational Studies* 16, 251–60.

1995. Canny writers: Tyneside dialect and the writing of secondary school students. *Educational Studies* 21, 3–12.

Williamson, Keith. 2008. *A Linguistic Atlas of Older Scots, Phase 1: 1380–1500*. University of Edinburgh.

Willis, Clodius. 1972. Perception of vowel phonemes in Fort Erie, Ontario, Canada, and Buffalo, New York: an application of synthetic vowel categorization tests to dialectology. *Journal of Speech and Hearing Research* 15, 246–55.

Windsor Lewis, Jack. 1994. The Yorkshire Ripper enquiry: Part II. *Forensic Linguistics* 1(2), 207–16.

Windsor Lewis, Jack (ed.). 1995. *Studies in General and English Phonetics: Essays in Honour of Professor J. D. O'Connor*. London/New York: Routledge.

Wolfram, Walt. 1969. *A Sociolinguistic Description of Detroit Negro Speech*. Washington, DC: Center for Applied Linguistics.

1999. Dialect awareness programs in the school and community. In Rebecca Wheeler (ed.) *Language Alive in the Classroom*, 47–66. Westport, CT: Greenwood Press.

Wolfram, Walt and Natalie Schilling-Estes. 2005. Dialectology and linguistic diffusion. In Brian D. Joseph and Richard D. Janda (eds.) *The Handbook of Historical Linguistics*, 713–35. Oxford: Blackwell.

Wolfram, Walt and Jason Sellers. 1999. Ethnolinguistic marking of past be in Lumbee vernacular English. *Journal of English Linguistics* 27, 94–114.

Wolfram, Walt and Erik R. Thomas. 2002. *The Development of African American English*. Language in Society 31. Oxford, UK/ Malden, MA: Blackwell.

Woolard, Kathryn A. 1985. Language variation and cultural hegemony. *American Ethnologist* 12, 738–48.

2008. Why *dis* now? Linguistic anthropological contributions to the explanation of sociolinguistic icons and change. *Journal of Sociolinguistics* 12, 432–52.

Wright, Sewell. 1943. Isolation by distance. *Genetics* 28, 114–38.

Yang, Charles. 2001. Internal and external forces in language change. *Language Variation and Change* 12, 231–50.

Yarmey, A. Daniel. 2001. Earwitness descriptions and speaker identification. *Forensic Linguistics* 8(1), 113–59.

Zalta, Edward. 2009. *Stanford Encyclopedia of Philosophy*. The Metaphysics Research Lab, Stanford University, http://plato.stanford.edu/.

Zerjal, Tatiana, Yali Xue, Giorgio Bertorelle, *et al.* 2003. The genetic legacy of the Mongols. *American Journal of Human Genetics* 72, 717–21.

Zetterholm, Elisabeth. 2003. *Voice Imitation: A Phonetic Study of Perceptual Illusions and Acoustic Success*. Lund University Press.

Zhang, Qing. 2005. A Chinese yuppie in Beijing: phonological variation and the construction of a new professional identity. *Language in Society* 34, 431–66.

2008. Rhotacization and the 'Beijing Smooth Operator': the social meaning of a linguistic variable. *Journal of Sociolinguistics* 12, 201–22.

Zilles, Ana M. S. 2005. The development of a new pronoun: the linguistic and social embedding of *a gente* in Brazilian Portuguese. *Language Variation and Change* 17, 19–53.

Index