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,

1

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 phono-logical 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 vari-ables need their own kinds of analysis.

The shifting sands of theory and technology create more challenges. Theoretical stances in phonology, such as generativism, autosegmental phon-ology, optimality theory, and exemplar theory, have at times induced variation-ists to adjust aspects of how they study data. However, variationists have often been content to let phonology work out its own issues without adapting phono-logical 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

8   *Erik R. Thomas*

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 inter-pretive 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 phon-ology – 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 some-what tied to structuralist theories of phonology, particularly those of George Trager and Bernard Bloch. For example, they recognised three levels of phon-emic 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 phono-logical 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 poly-phoneme. 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	ɖ	ɑ	ai	ɔi	au	ɒ	ɔ	ʌ	ɵ	o	ʊ	u	
three	ïi																ʊʉ		two
grease	ji																ɵʉ		tooth
six		ɪ															ʊˀ		wood
crib		ɪˑˀ															ʊˑˀ		pull
ear	ɪɚ																	ʉɚ	poor
beard	jɚˑ															ɝˑʊ			ago
eight			eɫ													ɝʊ			coat
April			eɫ													ɝʊ			road
ten		ɪˑˀ														ɝʊ			home
egg				eˀ												ɝʊ			know
head				eˑˀ												oˀɚ			four
Mary			ɛɫ													oˀɚ			door
stairs						æɚˑ										oˀɚ			hoarse
care						æɚ										oˑɚ			mourn
merry				—															
thirty					ˀɚˑ									ʌˀ					sun
sermon					ɚˑ									ʌˑˀ					brush
furrow					ɚˑ														
ashes						æˑ							ɔˑˀ						frost
bag						æˑˀ	ɖᵒ												log
married						æˀ							ɔː						dog
half						æˀ							ɔˑˀ						water
glass						æˀ							ɔˀ						daught
aunt						æˑˀ							ɔˀ						law
father							ɖᵒ												
palm						æˑˀ	ɖˑˀ												
barn							ɖɚ						ɔˀɚ						forty
garden							ɖɚ						ɔˀɚ						morning
crop							ɖˑˀ						ɔˀɚ						corn
John							ɖˑˀ						ɔˀɘ						horse
college							ɖᵒ												
borrow							ɖᵒ												
five									ɑˑɫ		æˑʉ̈								down
twice									ɑˑɫ		æY								out
wire							ɖˑɚ				aʊɚ								flower
										ɔˑɫ									joint
										ɔˀɫ									boil
	i	ɪ	e	ɛ	ɜ	æ	ɖ	ɑ	ai	ɔi	au	ɒ	ɔ	ʌ	ɵ	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 phys-
ical landscape. Results were then shown collectively by region. The physio-
geographic features dictated farming practices and industry in the LAGS terri-
tory, 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