English Language and Linguistics 11.1: 113–142. © Cambridge University Press 2007
doi:10.1017/S1360674306002139 Printed in the United Kingdom

The sound patterns of Englishes: representing phonetic similarity

APRIL MCMAHON, PAUL HEGGARTY, ROBERT MCMAHON, and WARREN MAGUIRE
University of Edinburgh
(Received 15 February 2006; revised 2 October 2006)

Linguists are able to describe, transcribe, and classify the differences and similarities between accents formally and precisely, but there has until very recently been no reliable and objective way of measuring degrees of difference. It is one thing to say how varieties are similar, but quite another to assess how similar they are. On the other hand, there has recently been a strong focus in historical linguistics on the development of quantitative methods for comparing and classifying languages; but these have tended to be applied to problems of language family membership, at rather high levels in the family tree, not down at the level of individual accents. In this article, we outline our attempts to address the question of relative similarity of accents using quantitative methods. We illustrate our method for measuring phonetic similarity in a sample of cognate words for a number of (mainly British) varieties of English, and show how these results can be displayed using newer and more innovative network diagrams, rather than trees. We consider some applications of these methods in tracking ongoing changes in English and beyond, and discuss future prospects.

1 How are accents different; and how different are accents?

In November 2004, the BBC commissioned an online poll on attitudes to accents, as part of the 'Voices' project (which more generally explored 'how we speak in the UK now'). The fact that over 5,000 responses were received shows how interested speakers of English are in each others' accents; and the results and comments make fascinating reading (http://www.bbc.co.uk/voices/yourvoice/poll_results.shtml, accessed 13 September 2006). Part of the fascination, however, lies in the opacity of many of the comments, which are difficult to interpret in phonetic terms, as shown in (1).

(1) (a) I don't really like the Birmingham accent that much (even though I've got one), but I do like the Black Country Accent . . . It sounds singy and old-fashioned. (S. Murphy, Birmingham)
(b) Yorksher accent and more specifically 'ull accent rules!. (red badger, 'ull)
(c) I never notice my accent until someone points it out and the way we shorten words and the rs we sound here in Bristol. (Debbie Smith, Bristol)

Many of these comments are straightforwardly attitudinal, and are expressed in terms of liking one accent and not liking another; and typically respondents are positive about their own varieties (though with distressing frequency this is not the case for the denizens of Birmingham). More interesting are the cases where respondents attempt to ground their comments with reference to particular phonetic or phonological

characteristics of varieties, as shown in (1) by 'red badger', who encodes [h]-dropping in his/her response, and in Debbie Smith's comment on Bristol 'r'. These mentions, however, are hardly very specific, and are consequently hard to interpret. What is the issue with 'r' in Bristol here? For instance, is it the phonetic quality that counts, or its distribution? And does it matter whether someone is commenting on a feature as particularly distinctive relative to the accent next door, as it were, or with a more global picture of English varieties in mind?

The obvious solution to questions like these is to call in the experts: in this case, to ask linguists to comment on such value judgements and informal descriptions, and to pick out the features speakers are responding to. This is a perfectly reasonable approach, and linguists are clearly able to locate the relevant features in a more sophisticated and more technically informed descriptive system, such as the IPA or Wells' (1982) Standard Lexical Sets. Likewise, we can attempt to explain why particular phonetic or phonological characteristics have the shape they do, from the perspective either of phonological theory, or of the different histories of each variety, invoking the motivations and patterns of sound change. Nonetheless, the uncomfortable truth remains that even linguists are essentially responding to features we intuitively perceive as salient differences between varieties. What we currently lack is a clear and agreed means of reinforcing and replicating these intuitions by comparing them with objective measures of linguistic similarity or difference.

Furthermore, there are good reasons for attempting to answer both of the questions in the title of this section. It is one thing to say how accents are different: as we have seen, respondents to the 'Voices' poll were clearly able in at least some cases to identify and localize particular distinguishing features, and linguists are typically well trained in spotting and systematizing such differences too. But it is something else again to be able to produce, alongside such descriptions, a measure of the degree of difference or similarity between varieties. It is not enough to know that Edinburgh and Glasgow Scots are different, or that both are different from Birmingham, Liverpool, Newcastle, and Bristol English: speakers can go further than this, in recognizing broad categories like 'Scottish accents', and even in misidentifying a Scottish accent as Irish or vice versa. Similarly, we as linguists should also be asking which varieties cluster together as more similar, and for each pair or group, *how* similar they are.

Answering this second question crucially requires the development of quantitative methods. The intention in pursuing quantitative approaches is not to supplant other linguistic approaches, but to support them. Clearly, describing variation in detailed phonetic terms, embedding the resulting facts into phonological models, and exploring dialect histories are all vital steps in understanding the differences between varieties. But when we move from the question of how (and why) varieties differ, to the related but distinct issue of how different they are, we require above all else robust and sensitive quantitative approaches to allow us to measure degrees of difference in objective, testable, and repeatable ways. These results in turn should then be compared with what we might know from linguistic history, or suspect from phonological theory, or be told by nonlinguist speakers.

Quantitative methods for language are by no means new, and over the past ten years or so there has been a particularly lively interest in developing and testing them. From our present perspective, however, the problem is that the focus of research on quantitative approaches so far has been predominantly historical. Questions of classification and subgrouping have featured heavily in work by, for instance, Ringe and his co-workers, whose computational cladistics project has taken a perfect phylogeny approach to first-order branching in Indo-European (Ringe, Warnow & Taylor, 2002; Nakhleh, Warnow, Ringe & Evans, 2005). There has been (rather controversial) work on dating protolanguages, or assessing when languages may have begun to diverge (Gray & Atkinson, 2003; papers in Forster & Renfrew, 2006). There have also been attempts to distinguish common inheritance from borrowing (McMahon & McMahon, 2005; McMahon, Heggarty, McMahon & Slaska, 2005; Heggarty, 2005). Increasingly, work on all these issues has begun to explore the utility of networks as well as trees in representing the often complex histories of languages (see the papers in McMahon, 2005).

In the discussion below, we begin to explore the prospects for developing different quantitative solutions to different problems. Here, the focus will be on the calculation, analysis, and representation of measures of phonetic similarity. Our work is essentially synchronic and variety-based, rather than a historical investigation at the language or family level; and we concentrate on the phonetics, whereas most work to date (leaving aside Ringe et al.) has involved lexical comparisons over basic meaning lists, with judgements of whether items in different languages are cognate or not. This does not mean that these different methods for different linguistic levels are necessarily incompatible in the longer term: it will be both interesting and vital at a later stage to compare the outcomes of phonetic similarity measures with assessments of how closely related particular varieties might be in genealogical terms. However, we make no mention of these historical measures and issues to begin with. If we keep these separate from our phonetic similarity scoring, we can maintain them as independent procedures, and perhaps use them later to validate one another by correlating their results and trying to interpret any differences or mismatches we might find (see also McMahon & McMahon, 2005: chapter 8).

2 Phonetic comparison

2.1 Why use quantitative methods?

It is clear what might in principle be gained from quantitative methods, but attaining these goals means being able to apply and assess such methods. It is important to understand that this will never involve a single stage: in fact, there are three interlinked but separate steps.

First, the linguistic data must be converted into numbers. For lexical data, this has typically meant selecting a basic meaning list; filling the slots on this list for a pair of languages; assessing which forms in the same slot are cognate; and assigning a code

https://doi.org/10.1017/S1360674306002139 Published online by Cambridge University Press

to reflect this. Such codes could simply be 1 for cognate and 0 for noncognate; or they can be rather more complex, as in the much-used Dyen, Kruskal & Black (1992) database, where different ranges of codes are also used to signal unique forms or likely borrowings. This first step of turning language data into meaningful numbers can be inherently much more complex than meets the eye, and will inevitably involve a number of stages of selection, programming, and design.

Second, the resulting numerical data must be processed. One important aspect of this stage involves generating and selecting appropriate visual representations by means of tree- and network-drawing programs. These outputs can in turn be tested further through statistical postprocessing.

Finally, there is a third stage of interpretation. Finding statistical significance does not necessarily tell us what the numbers involved mean, or what has motivated them; and we can see pictures without understanding why they are meaningful. It is therefore essential at this stage to involve linguists who are specialists in the particular language or languages concerned, if we are to have any realistic prospect of understanding what lies behind the patterns we have uncovered.

Of course, we might legitimately ask, if linguists have to interpret the results at this third stage in any case, why should we not expect them simply to work with the linguistic data directly in the first place, bypassing the steps of coding and processing altogether? We see three reasons for maintaining the three-stage quantitative analysis. First, numerical approaches can reveal patterns which are real but marginal, or which involve a relatively small number of data points in a large overall volume; this argument is familiar from corpus linguistics and sociolinguistics. Secondly, it is a core principle of scientific work in any domain that hypotheses are not only to be made, but also to be tested; and this means being able to replicate results and confirm or disconfirm initial ideas. Using quantitative methods to confirm something we already thought we knew is not futile, but on the contrary is one of the most important steps we can take towards confirming our linguistic intuitions and findings. Finally, if we find repeatedly that the same effect correlates with the same signal in the data, we can then assume the same interpretation in cases where the linguistic situation alone is less clear, allowing us to generalize from the known to the unknown. For all these reasons, quantitative approaches are valuable additions to linguistics. It follows that we must pay due attention to all the stages of coding, processing, and interpretation, involving colleagues with the appropriate expertise at each stage.

2.2 Coding the data

As we have seen, the most familiar type of coding for linguistic data involves so-called Swadesh lists of basic meanings, which are translated into different languages (a process in itself much more complex than is often assumed; see Slaska, 2005), and scored according to whether or not they are cognate. Such coding is of strictly limited usefulness for work at the accent and dialect level, since there will rarely be enough lexical distance in the basic vocabulary to provide fine, dialect-level classification. We therefore turn to the phonetics, where a single word potentially

provides a great deal more variation between accents, and therefore a great deal more information.

How, then, are we to convert phonetic data into numbers? We believe that the initial, coding stage *must* be linguistics-led. To transform phonetic data appropriately and meaningfully, it is essential to understand how phonetic and phonological systems work, and to appreciate that not all features are equal in terms of their salience or crosslinguistic distribution. An approach motivated principally by computational simplicity is likely to miss many of the apparently minor but actually rather important aspects of sounds and sound systems. On the other hand, at the second stage of analysis, we argue that tried and tested methods developed for analysing and visualizing biological data are eminently extendable to accent comparisons (as also demonstrated for language classification in McMahon & McMahon, 2005). We see no rationale at all for developing processing methods and models from scratch, since such programs (whether they are producing visual representations or undertaking statistical analyses) will simply be dealing with numbers, regardless of whether those are transformations of linguistic, molecular genetic or sociological data.

This is not to say that coding of phonetic data is absolutely untried: on the contrary, there have been numerous proposals for methods of phonetic comparison, and many of these are reviewed in Kessler (2005). Techniques and applications range from the extremely computationally simple use of Levenshtein distances (where strings are matched according to the shortest possible distance between them), through computation over feature bundles, to matching outputs with reference strings of some kind; the last category might include comparing a child's utterances to an adult target, or second language pronunciations with first language targets, or diagnosing and quantifying articulatory difficulties. Many of these approaches have proved useful in a particular domain, but would not be readily applicable to dialect comparison. Those which have been applied to quantifying similarity between dialects, like Nerbonne & Heeringa's (1997, 2001) Levenshtein-distance-based approach to Dutch, are arguably lacking in phonetic detail: this is discussed at length in Heggarty, McMahon & McMahon (2005) and Heggarty (forthcoming).

2.3 A linguistics-led approach

We have developed, therefore, our own purpose-designed stage 1 method for expressing phonetic similarity meaningfully in numbers. This has already been applied to a range of languages from four subfamilies within Indo-European in Heggarty (2000); and to a selection of Romance languages and dialects in Heggarty, McMahon & McMahon (2005), which also specifically contrasts the method with others developed from the standpoint of computational linguistics. Heggarty (forthcoming), moreover, gives a very full description, starting out from first methodological principles.

Here, our focus is more on stage 2 processing techniques, so this article is not the place to enter into the technical details of our method, already set out in those other publications. Moreover, our method takes its analysis and quantification of

https://doi.org/10.1017/S1360674306002139 Published online by Cambridge University Press

phonetic similarity to a higher level of complexity and detail than can be dealt with in the space available here. This in itself is somewhat atypical of most computational approaches to language hitherto, which typically set great store by the computational simplicity and 'elegance' of the algorithms that they borrow into linguistics from disciplines outside it. We, on the other hand, make no bones about our method being rather complex. For this is with good cause: the multifarious relationships between sounds that make for all their different degrees of similarity to each other are not in themselves particularly simple. Logically then, an overly simple, nonlinguistic model will not be able to represent those relationships properly, nor measure them usefully and meaningfully.

Devising any stage 1 coding method inevitably means confronting two methodological challenges, namely the quantification problem (how do we put meaningful numbers on aspects of language?) and the compatibility problem (how do we ensure we are comparing like with like between different languages or varieties?). Our method consists of two main components, each corresponding broadly to one of these challenges. First, we propose an analysis model for measuring phonetic similarity, which can take any two sounds in isolation, analyse the relationships between them, and convert that analysis into a numerical expression of the degree of phonetic similarity they exhibit. Secondly, to make use of that quantification model, it needs to be applied to real data sets that represent the phonetics of real languages. To compare like with like in this part of the analysis, we need some way of determining which sounds in two different languages or dialects we can meaningfully compare against each other within the word. That is, our method also needs a matching mechanism.

2.3.1 The quantification mechanism

Every language-specific or indeed dialect-specific system may differ from others in a number of ways. So in order for our method to compare and measure these different systems against each other we need to base both the comparison and the quantification around some common reference points, to which all of these different systems can be related. This common reference framework is to be found firstly in those concepts in our analysis of language that are *universal* rather than language-specific. An important consequence of our emphasis on universality is that our method is crucially distance-based, rather than character-based. In a distance-based method like ours, phonetic data will be transformed into measures of distance from one variety (over an agreed set of segments or sequences) to another, whereas in character-based approaches, numbers are assigned to different states of the same, preselected set of phonetic or, more often, phonological characters.

It is worth mentioning that this focus on universality also has a bearing on a fundamental question that much previous work in quantifying difference in sound has failed to address explicitly, let alone answer: do we compare languages at the phonemic or the phonetic level? The short answer is that phonemic systems and phonemes differ from one language or dialect to the next; so if we simply compare strings of phonemes, we cannot ensure that we are comparing like with like. If we are to have universal

https://doi.org/10.1017/S1360674306002139 Published online by Cambridge University Press

reference points, then we really need to compare phones, not phonemes. In any case, at the dialect level it is especially important to compare varieties to a high level of phonetic detail, since many dialects can have 'identical' phonemes, which however differ radically in their allophony and distribution. Many varieties of English have the 'same' phoneme /l/, for instance, but differ markedly in their realizations of it: Tyneside English has widespread clear [l]; much of Yorkshire has widespread dark [ɫ]; and Standard Southern British English (SSBE) and Standard American (Standard US) have both clear [l] and dark [ɫ] in different phonological contexts. Of course, this does not prevent a subsequent, additional comparison at the phonemic level, and Heggarty (forthcoming) outlines a method of this kind, but one that again necessarily begins with the phonetics. Nor is the concept of phonemic distinctiveness ignored in our method: on the contrary, it proves essential, as we shall see below – but on a different level.

How, then, do we go about putting meaningful numbers on phonetic similarity, in a principled way? The crudest summary of our quantification component is that it is a 'phonetic version' of distinctive feature analysis, in that sounds are compared in how many different *phonetic* features they share or differ in. As demanded by the need for compatibility, our model keeps as close as possible to universal, phonetic reality. Our guiding principle for ensuring that our figures are truly meaningful is that they must express in numbers the *significance* of each of the differences between sounds, *relative* to each other. Thus, we appeal to certain clear norms observable crosslinguistically, as evident in Maddieson's (1984) UPSID database – for example, we ask which of the possible phonetic differences are used most heavily for phonemic contrasts. These norms stand as proxy for the relative significance of phonetic differences crosslinguistically; and this shows also how our model makes use of the concept of phonemic distinctiveness.

For now our model is primarily articulatory, particularly for consonants, though it does also include a number of *ad hoc* mechanisms to balance cases where acoustic similarity departs significantly from articulatory similarity, as for example with [f] and [x] or bunched vs retroflex /r/. In general the method takes analysis to a level of phonetic depth far beyond most computational methods proposed hitherto. For instance, our study of English dialects has called for a refined system for transcribing length differences and assigning corresponding relative weightings to them. This is needed in order to distinguish, for instance, the longer and shorter forms of the 'phonemically long' English stressed monophthongs and diphthongs, as found before voiced and voiceless consonants respectively: the well-known allophonic length contrasts in *bead* vs *beat* and *strive* vs *strife*. Naturally, the analysis must be applied consistently across transcriptions of all varieties, to represent correctly the differences between varieties that do exhibit such differential lengths and those that do not. Of course, as we apply our model to finer and finer accent differences, our quantifications can still gain in accuracy from taking the analysis to even greater levels of detail; further refinement of our method is already underway, so that the model can interpret more diacritics in transcriptions, for instance.

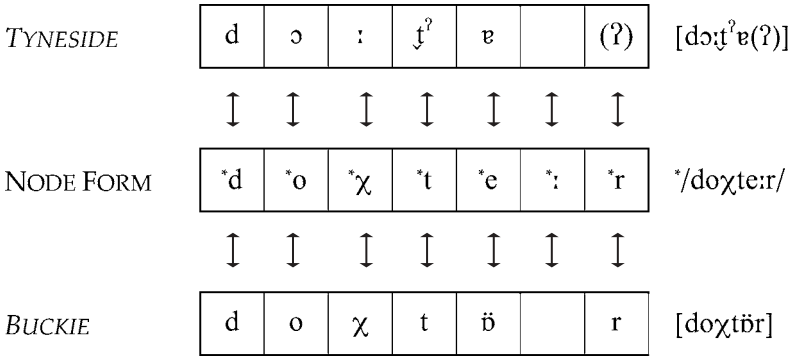


Figure 1. Matching up cognates in detail through an ancestor node representation for Tyneside and Buckie cognates of 'daughter' (originally Proto-Germanic ∗/doχteːr/

2.3.2 The matching mechanism

We must also ensure that we compare like with like in determining which sounds from one language or dialect are to be measured for similarity against which sounds in another. When the language varieties to be compared are from the same family, as is by definition the case with dialect comparison, then we can relate them all to the *common origin* they sprang from. We can make use of our knowledge about their common origin to allow us to compare not just any sounds considered in isolation, but the particular corresponding sounds in actual words from different related varieties. For each word in our database, we consider the set of its cognate forms across all these related language varieties, all descended from the same original form in the protolanguage. In Romance, for example, we would consider together the pronunciations of Italian *otto*, Spanish *ocho*, Portuguese *oito*, and French *huit*, all descended from the Latin numeral *octō* 'eight'. Our matching mechanism appeals to this ancestor form as a common reference point against which to match up, for each of its sounds, whatever phonetic reflex survives in the cognate form in each daughter language.

This is illustrated in figure 1 for two varieties of English, the Tyneside and Buckie cognates of Proto-Germanic */doχteːr/, *daughter*. The ancestor form of the cognate in fact serves only as a skeleton of articulations (or 'gestures') and lengths, to which we match up all the articulations and timing slots in each cognate form descended from it. These sounds in each modern cognate form can thus all ultimately be matched up against each other, through their relationships to the same sound in the ancestral form. That ancestor thus fulfils the role of a common reference point and 'node' through which to ensure that we always compare like with like between the varieties in that language family.

The system is in fact considerably more complex than it is possible to display in figure 1, so that sounds can be matched up to as detailed a level as possible. For example, multiple slots in one language's cognate may be matched up to a single slot in another: simple stops to affricates or fricative release stops; or short pure vowels to long vowels

https://doi.org/10.1017/S1360674306002139 Published online by Cambridge University Press

or diphthongs; and so on. In all cases, different relative lengths are always weighted and distributed appropriately, as is particularly important in cases of diphthongization, compensatory lengthening, and the precise comparison of English varieties that show significant contrasts in fine length differences. Surviving articulations can also be matched up correctly even where others have been lost from the sequence (or indeed new ones inserted by epenthesis), not only in cases of segment loss, but also nasalizations, palatalizations, etc. Indeed in these latter cases, *contemporaneous* articulations in one language can be matched up to what are *sequential* timing slots in another: the nasal gesture in a nasalized vowel ṽ, derived from original V+N; or the palatal gesture in C^{j}, coalesced from original C+[i]. Consonants can also be matched to corresponding vowels, and their phonetic similarity compared as appropriate.

Since the role and status of the node form are sometimes misunderstood, it is worth reiterating that it is only there to match up which sounds in each different language's cognate are the modern reflexes of each other – or in other words, the reflexes of the same original sound in the ancestor form. Once we have thus identified which two modern sounds are to be compared with each other, the operation to measure *their* phonetic similarity is made *directly* between them, by applying the method's quantification component to these two sounds as if in isolation. In these similarity calculations, the ancestor form is no longer involved at all. In other words, the comparison is established *through* the node form, but no quantification of phonetic similarity is ever made against it, so its close phonetic transcription is not a concern. Indeed, the node is not a true phonetic form at all, just a 'placeholder' skeleton of articulation types and slots, empty of any further phonetic details. What all this means is that the method is not dependent on us having a detailed phonetic transcription of the ancestor form. For language families for which we do feel confident in our knowledge of the ancestor language's phonetic details rather than just its reconstructed phonemic form, then we can add phonetic transcriptions of its cognates to the database just as we do for any other modern data language. But even if we do not, a simple skeleton representation of each ancestor form is enough for us to compare modern languages against each other.

2.4 The data set

2.4.1 The cognate list

The data set our method requires is made up of a set of words that are cognate in all the languages to be covered; our list of sixty Germanic cognates is shown in figure 2. That is, we compare German *Blume* with its true English cognate *bloom*, and not with *flower*, even though this is its modern semantic equivalent. *Flower* is of course a loanword from French *fleur*, and while this is itself ultimately a cognate root in Indo-European, the loan event broke in English the chain of normal continuous transmission that defines a true cognate.

Our quantified results therefore represent measures of the net divergence between a pair of language varieties in their phonetic realizations of the same cognate; or in other words, measures of the net differences which have arisen since their common ancestor

https://doi.org/10.1017/S1360674306002139 Published online by Cambridge University Press

	PROTO-GERMANIC	ENGLISH
1	boːkjoːn	*beech*
2	bloːðam	*blood*
3	bloːmon	*bloom*
4	broːθar	*brother*
5	doχteːr	*daughter*
6	daɣaz	*day*
7	auzoːn	*ear*
8	aχtoː	*eight*
9	augon	*eye*
10	fadeːr	*father*
11	fuːir	*fire*
12	fimfi	*five*
13	foːt	*foot*
14	petwor	*four*
15	fullaz	*full*
16	ɣeβan	*give*
17	ɣoðaz	*good*
18	ɣrasam	*grass*
19	ɣroːnjaz	*green*
20	handus	*hand*

	PROTO-GERMANIC	ENGLISH
21	haldan	*hold*
22	χornaz	*horn*
23	χaitaz	*hot*
24	χundam	*hundred*
25	eka	*I*
26	knewam	*knee*
27	mæːnon	*moon*
28	moːðar	*mother*
29	muːs	*mouse*
30	namoːn	*name*
31	neujaz	*new*
32	nokt	*night*
33	niɣun	*nine*
34	nasus	*nose*
35	ainaz	*one*
36	saltam	*salt*
37	saɣjan	*say*
38	seχwan	*see*
39	seβun	*seven*
40	seŋɣwan	*sing*

	PROTO-GERMANIC	ENGLISH
41	swestr	*sister*
42	seks	*six*
43	snaiwaz	*snow*
44	sunuz	*son*
45	sterron	*star*
46	sunnoːn	*sun*
47	teχan	*ten*
48	θu	*thou (you s.)*
49	θrijiz	*three*
50	tuŋɣoːn	*tongue*
51	tanθuz	*tooth*
52	twai	*two*
53	watar	*water*
54	χwat	*what?*
55	windaz	*wind*
56	wulfaz	*wolf*
57	wulloː	*wool*
58	wurmiz	*worm*
59	jæːram	*year*
60	juŋɣaz	*young*

Figure 2. List of sixty cognates for comparisons of varieties of English and other Germanic languages

form. Comparing cognates means comparing word forms that we know go back to the same common '*phonetic* origin', which is what guarantees that we are matching up like with like in phonetics. Meaning does not enter into the calculations; and this is exactly as it must be. For, if we allow ourselves to mix the semantic and phonetic levels, we end up with a hybrid measure partly of similarity in phonetics, partly of overlaps in semantics and cognate form. Indeed, if we were to match words by meaning, not cognacy, this would sometimes have us compare noncognates against each other (e.g. German *Hund* against English *dog*, rather than *hound*), in which case any measure of their phonetic similarity is effectively quantifying little more than the arbitrariness of the sound-to-meaning relationship.

The fact that we are limited to cognates might at first sight seem a weakness, but bear in mind that our method is not designed to diagnose whether or not languages are related. In work at the dialect level in particular, focusing on cognates is scarcely a hardship: there will typically be no shortage of cognates, while their validity is effectively guaranteed where the comparative method has been successfully applied

https://doi.org/10.1017/S1360674306002139 Published online by Cambridge University Press

(and if it has not been, then we need a different, diagnostic method in any case). For English in particular, we can have considerable confidence in the cognate status of the words we have selected, and for many of them a detailed knowledge of their phonetic histories to boot; and we can actively make use of this knowledge to ensure that our method matches like with like to a very high degree of precision. It follows that comparing by cognates is far from a limitation, but is the very essence of what we are trying to measure: net difference in phonetics, untangled from signals from different levels such as semantics. This approach also allows linguists to make full use of our hard-earned linguistic knowledge about sound correspondences, to contribute directly to the accuracy and reliability of our quantifications.

2.4.2 Language varieties

As input, our model requires a fairly close phonetic transcription of each of our sixty cognate words in each variety in the study. We would like to express our thanks to the colleagues who provided the transcriptions, all linguists who are native speakers of and/or specialists in the varieties concerned.[1] For the purposes of this article, our data set serves principally to demonstrate the operation and potential of our stage 1 method; and to provide some sample results that can be fed into some stage 2 methods. It is this primarily illustrative aim that has governed our selection of the twenty language varieties covered so far: these are Proto-Germanic, Old and Modern Icelandic, Standard German (Hochdeutsch), West Saxon Old English, Standard US, Australian (Victoria), Sheffield, Liverpool, Berwick, Standard Scottish English, Glasgow, Buckie, Tyrone (both Traditional and Standard), RP, Middlesbrough, Tyneside (Traditional), Wisbech and Derby. Our ongoing research[2] will soon greatly extend the database to a more extensive word list and to further varieties of English, as well as to a number of other Germanic languages.

Our stage 1 method has been purposely designed to be as flexible as possible, so that we can use it to compare as widely as possible across a range of domains in which language varieties can differ. The varieties included span all levels from the most similar *accents*, through more different *dialects*, to quite different *languages*. The inclusion of varieties of Icelandic and German entails that the common ancestor to be used as our matching skeleton has to be Proto-Germanic, not an earlier attested or reconstructed form of English. We also wish to compare *different historical stages* of those varieties against each other, hence the inclusion of Old and Modern Icelandic, as well as West Saxon Old English. Older varieties have been entered as data languages like any modern varieties; but clearly, their transcriptions cannot be as precise or reliable in phonetic detail, so quantifications for earlier varieties are necessarily somewhat more

[1] Our thanks, for the varieties indicated, go to: Joan Beal (Tyneside), Gavan Breen (Australia), David Britain (Wisbech), Jayne Carroll (Old and Modern Icelandic), Karen Corrigan (Tyrone), Paul Foulkes (Derby), Patrick Honeybone (Liverpool), Mark Jones (Sheffield), Carmen Llamas (Middlesbrough), Warren Maguire (Tyrone), Kim Schulte (Standard German), Jennifer Smith (Buckie), Jane Stuart-Smith (Glasgow), Dominic Watt (Berwick).

[2] We gratefully acknowledge the financial support of the AHRC for project 112229 ‘Sound comparisons: dialect and language comparison and classification by phonetic similarity’.

tentative (as we indicate by showing them in *italics* in the results tables). This applies *a fortiori* to results from our assumed phonetic transcriptions for Proto-Germanic, for of course reconstruction works essentially only at the phonemic level. In any case, as discussed above, this entry of Proto-Germanic as a data language in its own right does not in any way affect the validity of our matching node skeleton. This is an entirely separate structure in the database which takes the form only of the articulation and slot structure assumed for the Proto-Germanic form, and for this purpose narrow phonetic detail is not at all required. The method can equally well be used to compare *different stylistic* or *sociolinguistic varieties* too, though so far we have only included a single illustration of this, in the shape of two variants for the English of Southwest Tyrone, a more traditional and a more standardized one.

Clearly, there are some limitations inherent in the methods we are using, which may also require revisions to be made in future; at the very least, they need to be borne in mind when we are interpreting our results. Although we hope to have established that using cognates does not in itself constitute a problem, we must ensure that the specific cognates we select allow coverage of as many segments and contexts as possible for the languages and varieties we are comparing. This requires constant review as we add new varieties, with their own phonological systems and phonotactic restrictions. The fact that a number of colleagues have provided transcriptions introduces the possibility of inconsistencies in transcription practice, and this may be all the more salient since we are dealing with both standard and nonstandard varieties: it is at least conceivable that transcribers might be disposed to produce closer transcriptions and use more diacritics when they are dealing with nonstandard varieties, whose descriptions are much less frequently reduced to broad phonemic terms. We have attempted to pre-empt these problems by producing extremely detailed and fully exemplified instructions to volunteer transcribers, and in future work will be conducting more crosschecks on transcriptions, as well as involving a member of the project team who will discuss the exercise personally with all those producing transcriptions. Finally, there is something of a mismatch in any attempt to produce very detailed phonetic transcriptions from what is effectively a single speaker: if we take our aims to their logical limit, we should be at least as interested in variation within an accent as in variation between accents. Again, this is an issue we will be addressing explicitly in the next stage of our research; but there would be no point in proceeding to that further level of detail if we could not demonstrate that there are useful insights to be had for even a relatively minimal data set.

3 Results: output from stage 1 coding

3.1 Similarity ratings

Our stage 1 method (which was written by Paul Heggarty as a program in Visual Basic for Microsoft Excel) calculates, for every pairwise combination of language varieties in the study, a measure of the similarity between their particular phonetic

https://doi.org/10.1017/S1360674306002139 Published online by Cambridge University Press

reflexes of the same cognate. Figure 3 shows these overall results for the full database of sixty cognates, which form the input to the stage 2 methods discussed in section 4 below. These sixty word forms together serve as simply a sample of the phonetics of each variety, and in this sense their cognate status is to ensure that this sample is strictly equivalent from one variety to the next, in that all go back to the same common phonetic starting point. In combining together the results from each cognate to produce the overall results, the word actually plays no role as a weighting unit, since it is largely a phonological and indeed a grammatical concept, and again, we must not mix levels. Our calculations are in *phonetics*, so the basic weighting unit is a phonetic one: a default-type phone (see Laver, 1994: 571–2), with a single articulation, which we take also as a 'standard-length' segment. This ensures that the same sound difference always contributes equally to the overall results, whatever the length of any particular words it occurs in.

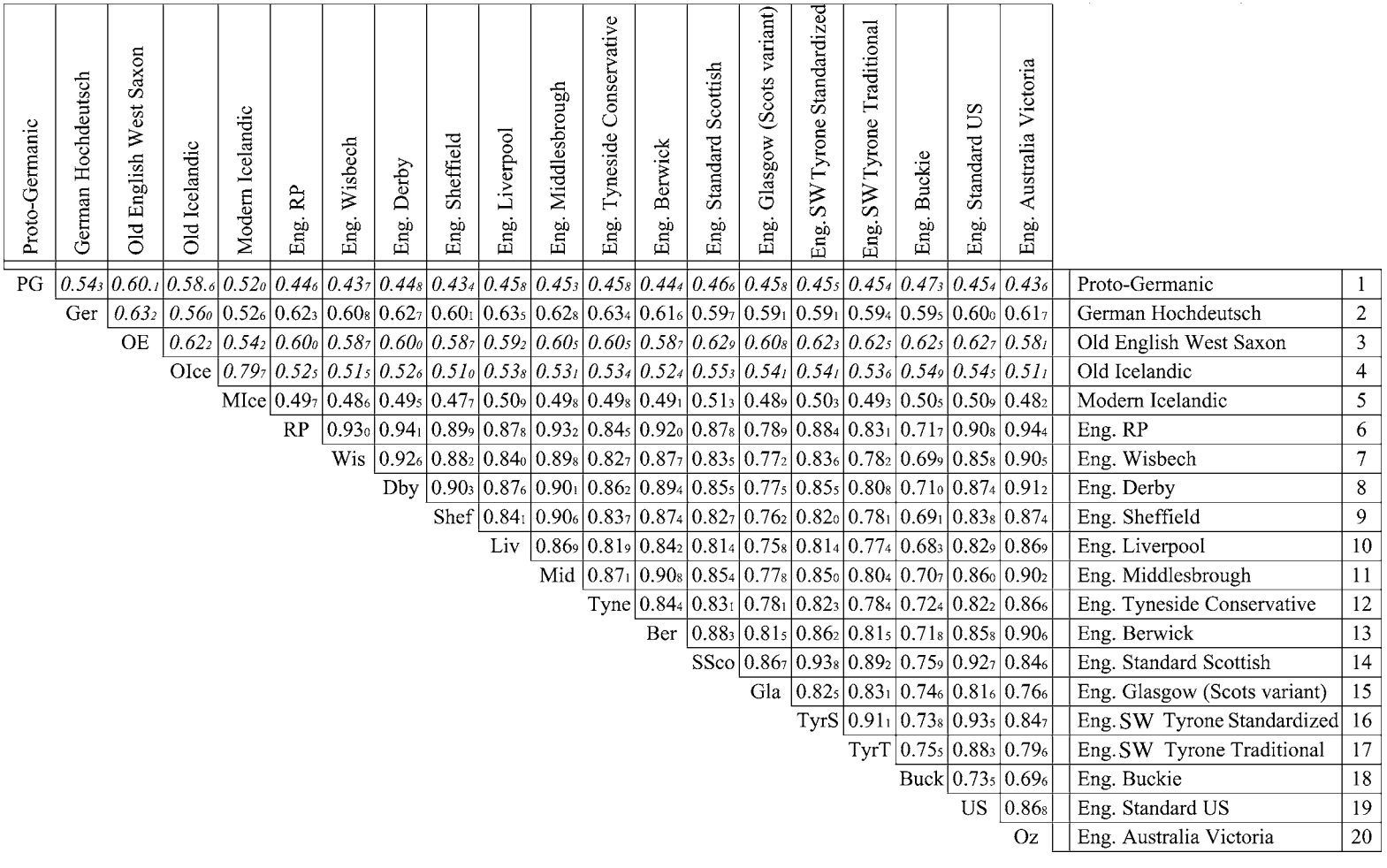
However, it should be noted that not every cognate set displays the same level of phonetic variation. In the cognates of *bloom*, pronunciations are highly similar in all varieties of English, with all results for comparisons between them having similarity ratings of 0.8 or above (where 1 is the value for forms that are phonetically identical, at least to the depth of phonetic detail the model is taken to at present). On the other hand, the cognates of *daughter* show considerably more difference from variety to variety: consider, for example, Liverpool [dˠɔˈθ͇ɛ] vs Tyneside [dɔːt̬ˀɐ] which would come out only 0.47 similar.

Rather than impressions and hitherto unquantified judgements, we now have actual hard and precise *measures* of similarities and differences; and in a numerical format that allows us to input them to further mathematical processing. But this further processing is particularly important, because the results from our stage 1 method in figure 3 end up as an intimidating mass of figures. While we know there must be crucial signals in them, we cannot easily and reliably tell apart all the complex relationships of each variety to every other one by simply looking at the figures and juggling them in our minds all at once. The more varieties we add, the more impractical this becomes; in principle we should of course be gaining from having more data, but we need to be able to grasp what they are telling us. Evidently, what is required is some means of synthesizing the complex signals of relatedness contained in all these figures, and representing them graphically for us to interpret more easily, and in a more objectively balanced way. It is at this point, then, that we turn to the prospect of adopting and adapting stage 2 processing tools.

4 Processing using network programs

In the remainder of this article, we shall focus on processing (that is, stage 2 activity in our terms) using resources beyond the phonetic comparison program itself. For us, this means programs originally developed for biological data. It is true, at least in principle, that linguistic data might behave differently from biological data; and if so, then new programs tailored to linguistic data might appropriately be developed

https://doi.org/10.1017/S1360674306002139 Published online by Cambridge University Press



Proto-Germanic	German Hochdeutsch	Old English West Saxon	Old Icelandic	Modern Icelandic	Eng. RP	Eng. Wisbech	Eng. Derby	Eng. Sheffield	Eng. Liverpool	Eng. Middlesbrough	Eng. Tyneside Conservative	Eng. Berwick	Eng. Standard Scottish	Eng. Glasgow (Scots variant)	Eng. SW Tyrone Standardized	Eng. SW Tyrone Traditional	Eng. Buckie	Eng. Standard US	Eng. Australia Victoria		
PG	*0.543*	*0.60.1*	*0.58.6*	*0.520*	*0.446*	*0.437*	*0.448*	*0.434*	*0.458*	*0.453*	*0.458*	*0.444*	*0.466*	*0.458*	*0.455*	*0.454*	*0.473*	*0.454*	*0.436*	Proto-Germanic	1
	Ger	*0.632*	*0.560*	0.526	0.623	0.608	0.627	0.601	0.635	0.628	0.634	0.616	0.597	0.591	0.591	0.594	0.595	0.600	0.617	German Hochdeutsch	2
		OE	*0.622*	*0.542*	*0.600*	*0.587*	*0.600*	*0.587*	*0.592*	*0.605*	*0.605*	*0.587*	*0.629*	*0.608*	*0.623*	*0.625*	*0.625*	*0.627*	*0.581*	Old English West Saxon	3
			OIce	*0.797*	*0.525*	*0.515*	*0.526*	*0.510*	*0.538*	*0.531*	*0.534*	*0.524*	*0.553*	*0.541*	*0.541*	*0.536*	*0.549*	*0.545*	*0.511*	Old Icelandic	4
				MIce	0.497	0.486	0.495	0.477	0.509	0.498	0.498	0.491	0.513	0.489	0.503	0.493	0.505	0.509	0.482	Modern Icelandic	5
					RP	0.930	0.941	0.899	0.878	0.932	0.845	0.920	0.878	0.789	0.884	0.831	0.717	0.908	0.944	Eng. RP	6
						Wis	0.926	0.882	0.840	0.898	0.827	0.877	0.835	0.772	0.836	0.782	0.699	0.858	0.905	Eng. Wisbech	7
							Dby	0.903	0.876	0.901	0.862	0.894	0.855	0.775	0.855	0.808	0.710	0.874	0.912	Eng. Derby	8
								Shef	0.841	0.906	0.837	0.874	0.827	0.762	0.820	0.781	0.691	0.838	0.874	Eng. Sheffield	9
									Liv	0.869	0.819	0.842	0.814	0.758	0.814	0.774	0.683	0.829	0.869	Eng. Liverpool	10
										Mid	0.871	0.908	0.854	0.778	0.850	0.804	0.707	0.860	0.902	Eng. Middlesbrough	11
											Tyne	0.844	0.831	0.781	0.823	0.784	0.724	0.822	0.866	Eng. Tyneside Conservative	12
												Ber	0.883	0.815	0.862	0.815	0.718	0.858	0.906	Eng. Berwick	13
													SSco	0.867	0.938	0.892	0.759	0.927	0.846	Eng. Standard Scottish	14
														Gla	0.825	0.831	0.746	0.816	0.766	Eng. Glasgow (Scots variant)	15
															TyrS	0.911	0.738	0.935	0.847	Eng. SW Tyrone Standardized	16
																TyrT	0.755	0.883	0.796	Eng. SW Tyrone Traditional	17
																	Buck	0.735	0.696	Eng. Buckie	18
																		US	0.868	Eng. Standard US	19
																			Oz	Eng. Australia Victoria	20

Figure 3. Phonetic similarity ratings for a range of varieties of English and other Germanic languages for a set of sixty cognates. 1 = identical pronunciations (to the level of phonetic depth covered so far). Figures below 1 denote progressively less phonetic similarity

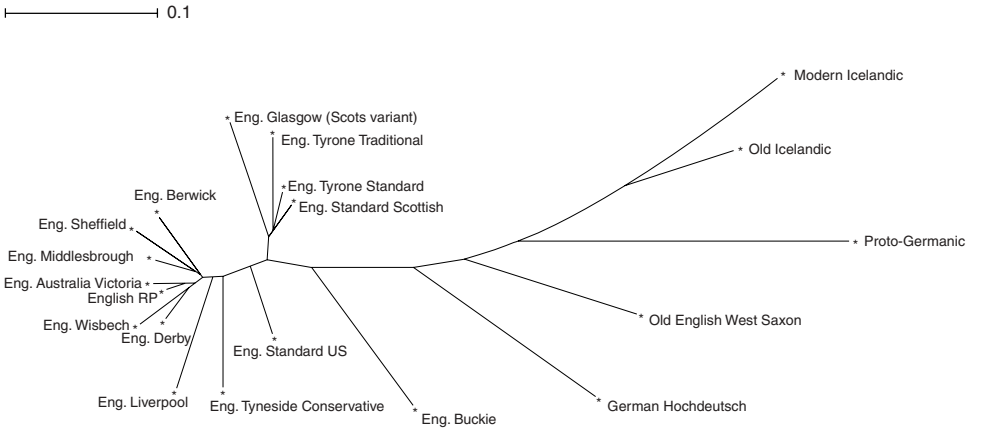


Figure 4. Neighbour-joining tree, sixty cognates, English and Germanic

in the future. However, we cannot tell whether that further step is necessary without checking the outputs we find from existing programs. It is also essential to understand the requirements and assumptions of different models as they apply to biological data, in order to assess whether they are suitable for use with linguistic data: the programs on offer are already diverse and are rapidly becoming more so, and it is certainly not valid to think of them together as if they were all essentially the same. In the following sections, we shall therefore discuss some preliminary results from a range of programs.

4.1 Tree-drawing and tree-selection programs

The simplest programs first developed for the processing and visual representation of biological data were tree-drawing and tree-selection programs. These are familiar to linguists in that they produce diagrams reminiscent of traditional linguistic family trees: for the most part, these will be unrooted and thus more star-like in appearance, but they can easily be converted into rooted trees by selecting a particular language to serve as the root (see also McMahon & McMahon, 2005, where tree and network programs are discussed extensively). Figure 4 shows the output from the Neighbour-Joining algorithm in the SplitsTree 4 package (Huson & Bryant, 2005), which has been selected here because it most closely approximates the steps that would be taken by a linguist drawing a tree: the two closest languages or groups are clustered, then the next closest is added, and so on up through the tree. This is a robust and computationally very tractable way of drawing trees; the difference the program makes is in the additional step of generating all or many of the possible trees for the data, then selecting the best. So figure 4 shows not a random tree, but one of those trees that fits the phonetic similarity data for our sixty cognates and twenty varieties best. The tree in figure 4 is unrooted, and branch lengths are meaningful, so that longer branches mean more change: hence, as we might expect, the distance between the modern varieties and

https://doi.org/10.1017/S1360674306002139 Published online by Cambridge University Press

Proto-Germanic, away on the far right, and the location of Old Icelandic and Old English nearest to the protolanguage. As for relationships within English, we find the English English varieties (plus that of Victoria, Australia) all in a cluster; a further cluster of Scots and Irish varieties (with the exception of Buckie, which is alone and closer to the protolanguage, suggesting that it is archaizing in some respects); and Standard US lying at the root of the Scottish cluster.

One obvious but problematic aspect of tree-drawing programs is that they are designed only to draw trees. They will therefore find and recommend the best tree for the data, even where this does not fit all the data; and we risk missing information that is not consistent with *any* tree. For example, if a variety shares certain features with one cluster of varieties, and others with a second cluster, the tree may represent it as intermediate (as with Standard US in figure 4). However, it may also appear within one cluster or the other, as the program may prioritize one set of similarities and effectively disregard the other. The essential problem here is that relationships between varieties are multidimensional, and when such complexity is forced into two dimensions, which are all we are permitted given a binary branching tree structure with no connections between branches, then distortions may occur. Trees cannot show the effects of contact either, and this may be vitally important at both the language and dialect levels. The fact that a tree-drawing program selects a tree does not guarantee that all the characteristics of the varieties depicted are completely tree-like, but reflects the fact that such programs have no option but to draw a tree, even if that is an idealization and conceals certain aspects of the real situation in which linguists might be very interested indeed. When we talk about a program selecting the 'best' tree, then, this does not mean the perfect or ideal representation, but only the tree that is *least incompatible* with the signals in the language data.

One way of checking for disparities of this kind is bootstrapping, a further stage 2 processing technique which involves altering the data set slightly (for example, by randomly removing 5 of the 60 cognates, and either reprocessing with the remaining 55, or resampling 5 more from the full list). Such resampling may reveal a highly consistent, single-consensus tree, which would then be extremely well supported; but it might also show that different trees are selected on different runs, suggesting that there are non-tree-like factors at work. This is precisely what we find in the case of Berwick, which appeared in figure 4 within the English English cluster. However, in fully 25 out of 35 resampled bootstrap iterations, the quite different pattern shown in figure 5 emerged instead: here, Berwick (and also Tyneside) appear on the margins of the English English cluster, and both appear to be inclining towards the Scots and Irish varieties. Note that in this rooted phylogram, horizontal distances are meaningful but distances up and down are meaningless and included for visual clarity only.

Bootstrapping reveals instability or incompatibilities in the data, but cannot change the essential limitations of tree-based analyses. Fortunately, network programs are now readily available: their primary innovation involves assessing whether the data are really fundamentally tree-like in the first place, or whether they are characterized by overlapping and conflicting patterns.

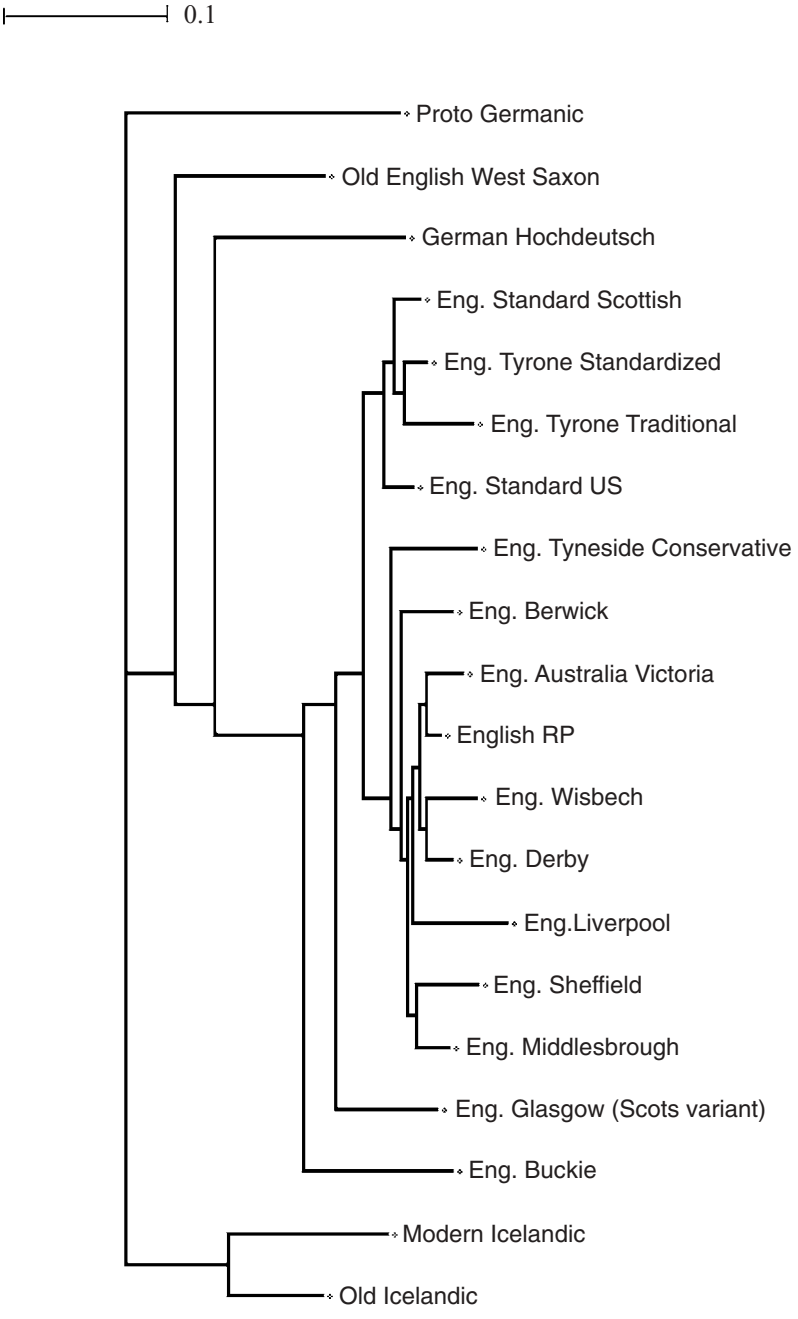


Figure 5. Neighbour-joining tree, resampled (pattern found in 25 of 35 runs). Tree has been rooted by setting Proto-Germanic as an outgroup

https://doi.org/10.1017/S1360674306002139 Published online by Cambridge University Press

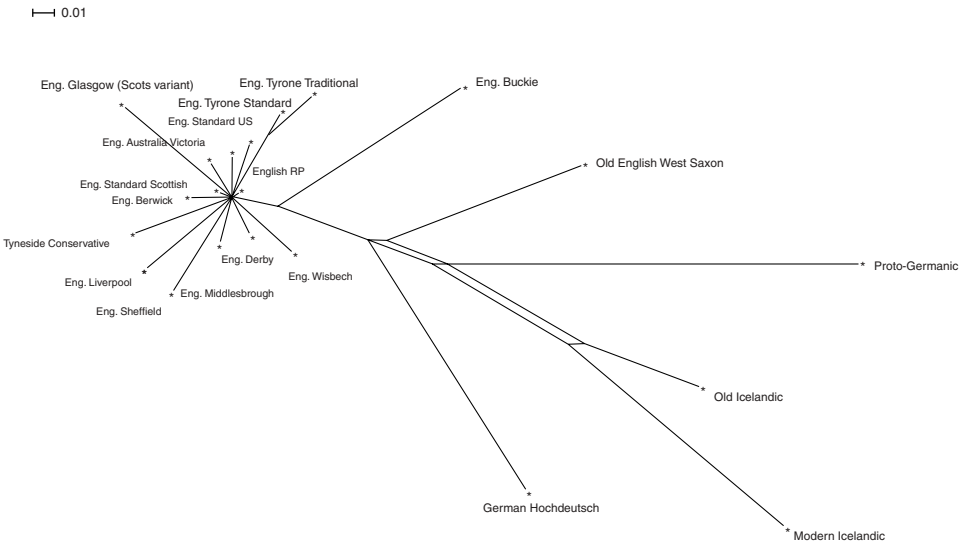


Figure 6. SplitsTree, sixty cognates, English and Germanic

4.2 Network programs

The distinct advantage of network-type programs is that they draw a tree only in cases where the relationships between the cases depicted do indeed create a tree-like pattern, whereas when these are compatible with more than one tree they construct a network that features reticulations between cases. A reticulation is a line drawn between varieties, which gives the appearance of a box rather than a straightforward branch and converts a strict tree into a network configuration, as can be seen for Old Icelandic and Old English in figure 6 (drawn using the network program SplitsTree[3]). Crucially, then, network-type programs need not ever force a tree on non-tree-like data, but can represent graphically both the tree-like and the non-tree-like aspects of relatedness in the same diagram. In biology, reticulations mean either homoplasy (that is, independent parallel development), or 'mixing', whether by recombination or gene transfer. Both are also relevant to linguistics, where parallel innovation is fairly frequent with certain types of natural sound change (as with the case of syllable-final /l/ vocalization discussed by Heggarty, 2006: 187, for example), and borrowing is likewise widespread.

Two network programs which are suitable for the continuous numerical distance data output from our stage 1 program are SplitsTree and NeighbourNet (Huson, 1998; Huson & Bryant, 2005; the processes of split decomposition involved in network construction

[3] Note that the first network-type program used for analysis of linguistic data was Network (Bandelt et al., 1995; Bandelt, Forster & Röhl, 1999, http://www.fluxus-engineering.com); however, this is appropriate for character data only, while we use distance-based data.

https://doi.org/10.1017/S1360674306002139 Published online by Cambridge University Press

are discussed in detail in McMahon & McMahon, 2005; chapters 7 and 8). An initial output from SplitsTree is shown in figure 6; but although this provides good resolution at the language level, it appears to have diagnosed so many interconnections between varieties of English that this part of the diagram has simply collapsed into a star-like shape. This does tell us that we have interrelations between dialects; but the program does not cope well with large data sets including complex and crosscutting signals. There is currently no way of resetting the resolution on the program, so we must conclude that SplitsTree is currently likely to be unworkable for analyses of phonetic similarity at the dialect level.

More promising results, however, come from applications of NeighbourNet (implemented in SplitsTree4, available from http://www.splitstree.org), which operates at a level of resolution suitable for both language and dialect data. NeighbourNet, like the tree-drawing program discussed earlier, employs a neighbour-joining algorithm, but is extended to make it possible to handle larger data sets that show more complex and potentially inconsistent signals in patterns of similarity.

We can make a distinction between two sets of methods for comparing and classifying systems, and between two types of visual representations they generate. On the one hand, cladistic methods classify on the basis of signals of common ancestry, prioritizing features which show descent with divergence from an original common source; and they generate phylograms, diagrams intended to mirror the order of historical branching within a group. The comparative method in historical linguistics is a cladistic method, and its results, appropriately enough, are represented in family trees. On the other hand, however, NeighbourNet is a phenetic method, designed to diagnose signals of similarity, regardless of their origin and significance; and these are shown in phenograms, which again depict similarity, or distance, without prejudice to whether that results from common ancestry, contact or parallel developments. NeighbourNets can therefore be read in terms of relative distance between varieties, but the clusters the program produces will not necessarily reflect historical affiliations; they may, but further investigation would be required to demonstrate whether this is the case. Since our priority is to investigate synchronic phonetic similarity, with the causes of that similarity being investigated at the third stage of interpretation, phenetic methods of this kind are well suited to our data, which of course are measures of similarity in the first place. However, we must take care not to evaluate the outputs of NeighbourNet by expecting them to match trees we might anticipate finding on the basis of historically motivated groupings. Phylograms and phenograms can validly be compared, and we can hope to learn a great deal from such comparisons; but we cannot either support or reject one type of representation on the basis of matches or mismatches with the other. For illustration, an initial NeighbourNet for our data is shown in figure 7. This figure shows the same general, overall structure as our initial Neighbour-Joining tree in figure 4, but with the addition of plentiful reticulations to show differential feature sharings across varieties and clusters. Notably, Berwick here emerges as clearly intermediate, with reticulations linking it in some cases with the Scots and Irish varieties, and in other cases with the English Englishes.

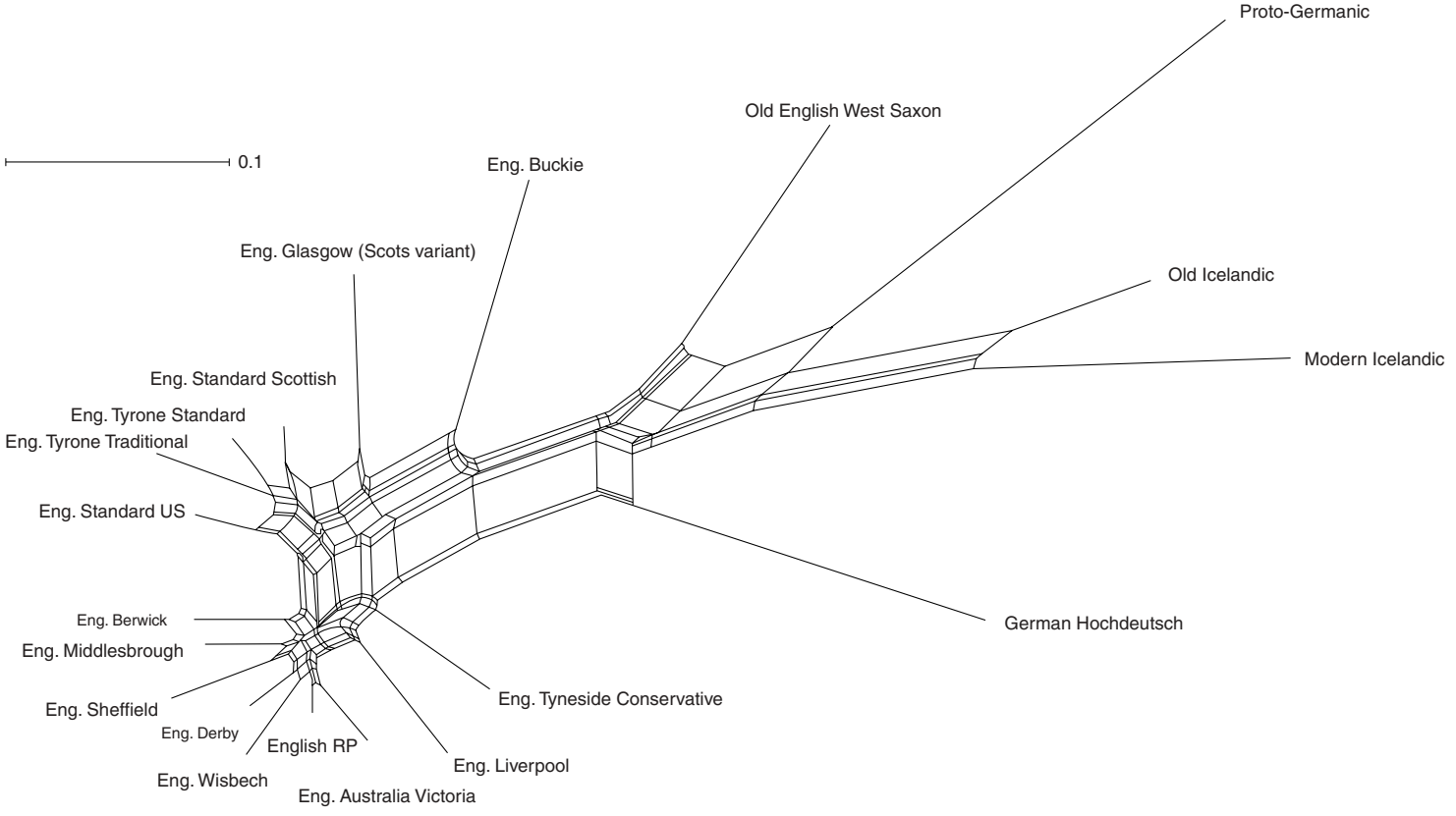


Figure 7. NeighbourNet output for average distance for sixty cognates between twenty varieties

These networks have the considerable advantage of encapsulating different possible tree structures in a single diagram. When a reticulation appears in a network, it shows similarities incompatible with a single tree. In network A of figure 8, we see a reticulation, or box, joining up four of our varieties. However, each of the trees B and C represents a loss of information from our data, effectively ignoring either the similarities between RP and Berwick, and Standard US and SSE (B), or between Berwick and SSE, and Standard US and RP (C). The price of drawing a tree is the prioritization of some similarities over others, rather than the depiction of all of them, which is what the network allows. We would predict that, in such cases, bootstrapping of trees B or C would reveal shifting allegiances among the varieties, as shown already in section 4.1 above.

Indeed, we can also carry out bootstrapping for network programs; this can demonstrate that the patterns we find are robust, and rank the splits or reticulations in the diagram for consistency and robustness. Three bootstrap resamplings for just our twenty varieties of English are shown in figure 9. The top right and bottom diagrams show Berwick intermediate between the two major clusters, as in figure 7; the top left network has Berwick much closer to the other English Englishes, but this is a pattern seen in only 6 of 35 iterations.

4.3 Further processing and interpretation

As we have seen, part of stage 2, the processing aspect of quantitative work on linguistic data, involves the generation and selection of visual representations. However, further work is required to establish the meaning and significance of the various splits in these diagrams. Resampling and bootstrapping already provide some preliminary statistical testing, and identify the splits and configurations of varieties that are most robust and therefore need to be prioritized at the subsequent stage of interpretation: there is clearly no sense in attempting to interpret and explain a signal which turns out to be completely artefactual. The Network program referred to earlier has an advantage here, in that it generates, alongside its selected best diagram for the data, a list of those characters which are behaving in a non-tree-like way, so that these can immediately be prioritized for further consideration. Programs like SplitsTree and NeighbourNet are unable to generate such lists, because they are operating with distance data, and therefore with composite scores over a range of data points. Unlike Network, which works directly on character data, these programs consequently cannot identify the particular points which have contributed non-tree-like patterns to the overall distance measures. However, there are alternative ways of identifying and conducting more in-depth analyses of particular splits in NeighbourNet diagrams.

Figure 10 shows one major split which appears in 100 per cent of our bootstrapped runs of NeighbourNet. Such individual splits can be highlighted and then investigated further, to elucidate them by establishing which of our sixty cognate items, and therefore which phonetic features, are contributing to their appearance.

https://doi.org/10.1017/S1360674306002139 Published online by Cambridge University Press

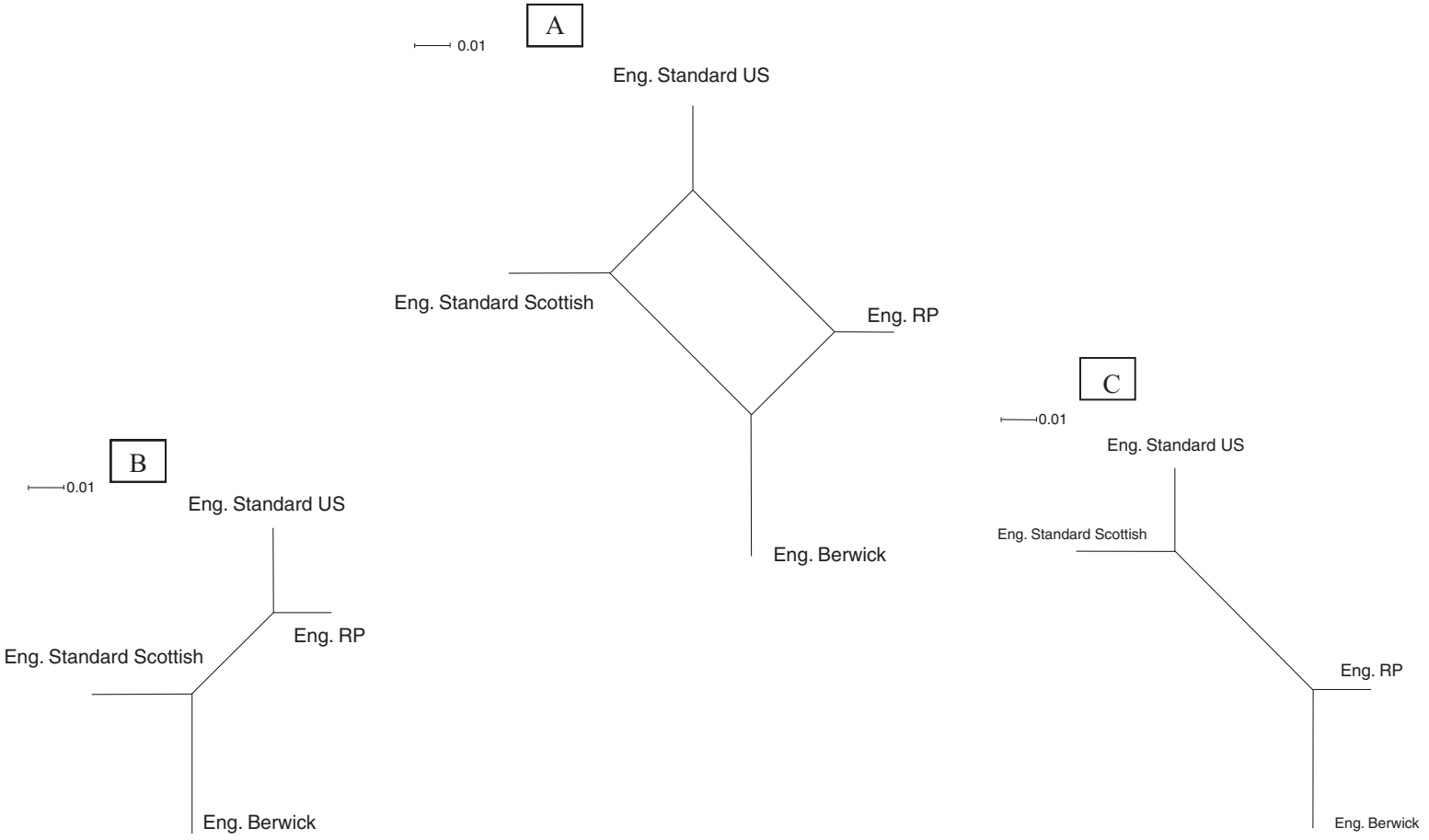


Figure 8. Relationship between network with reticulations and encapsulated trees

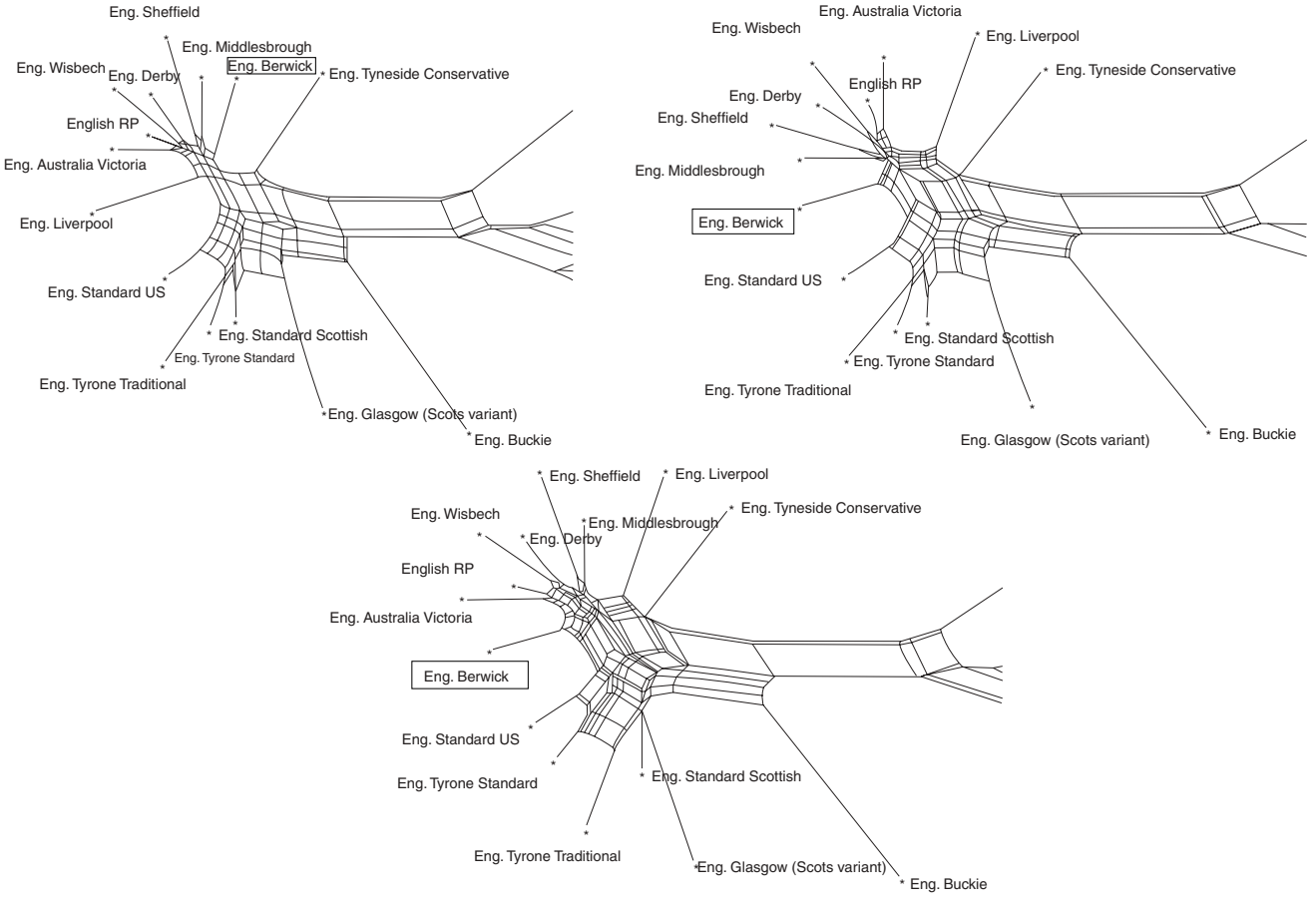


Figure 9. Three bootstrap iterations of NeighbourNet, sixty cognates, English varieties

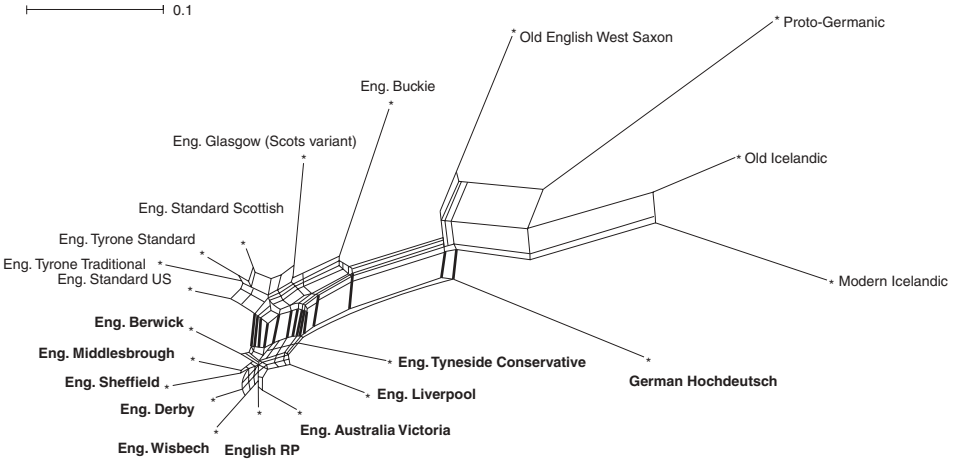


Figure 10. A major split in the data

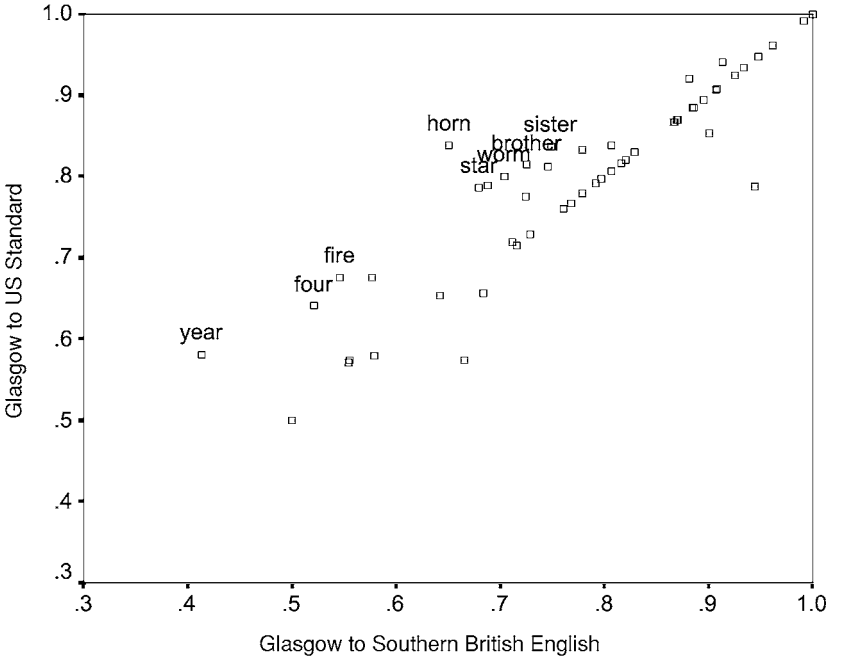


Figure 11. Cognates contributing to the major split

In this case, we contrasted two varieties taken from the same side of the major split in figure 10, here Glasgow and Standard US; and then further compared one of these varieties, Glasgow, with a third variety from the other side of the split, here RP or Standard Southern British English. The resulting scatter plot appears as figure 11.

https://doi.org/10.1017/S1360674306002139 Published online by Cambridge University Press

The sixty data points corresponding to our sixty cognates are distributed around the diagonal separating the two two-way comparisons: the items that lie furthest off the diagonal and *above* it are those in which Glasgow and Standard US are more similar to each other than either is to RP. If we isolate these items, the majority, as labelled in figure 11, have postvocalic /r/. Clearly the major split is the well-known division between rhotic and nonrhotic varieties of English.

In the same way, we can ask which features are involved in the shifting position of Berwick, which our bootstraps of trees and network representations have both established as lying between the English English cluster, and the Scots and Irish one. Clearly Berwick is nonrhotic, and it appears consistently on the nonrhotic side of the major split in figure 10; but which features are pulling it towards the Scots and Irish cluster? A similar scatter plot analysis has been carried out to produce figure 12, which displays comparisons between Berwick and Middlesbrough, and Berwick and Scottish Standard English (SSE). The relevant forms here are those furthest off the diagonal line and *below* it, which have the characteristics shared by Berwick and SSE, but not by Berwick and Middlesbrough. These include items like *tooth*, *green*, *three*, *beech*, *hold*, *see*, *new*, *day*, *blood*, *bloom*, *moon,* and *good*; and although this set would require further, detailed, stage 3 linguistic interpretation, the strong indication is that high vowels, diphthongs, and perhaps vowel length are implicated in the connection between Berwick and the Scots and Irish varieties. To produce a full picture, the same analysis would have to be replicated for all the varieties in the Scots and Irish cluster against Berwick; and it would also be interesting to assess whether the same forms link Tyneside with this cluster. For comparison, items above the diagonal in figure 12 are those where Berwick and Middlesbrough share more than Berwick and SSE; and here again, it is no surprise to find a concentration of items with orthographic postvocalic ⟨r⟩.

Another split present in 100 per cent of runs distinguishes Buckie from every other variety of English: moreover, the highlighting in figure 13 shows that this primarily involves features shared by Buckie and all the older languages – namely Proto-Germanic, Old Icelandic and West Saxon Old English – as well as by Modern German and Modern Icelandic.

Again, a scatter plot identifies which cognates are contributing to the isolation of Buckie from all other modern varieties of English: these are given in figure 14 with their transcriptions for Buckie. It is clear that these are indeed very different from the other Englishes, including the other varieties of Scots. Some cognates do show innovations in Buckie (*eye*, and *what*, for example), but many seem particularly archaic in Buckie, with transcriptions notably close to those for Proto-Germanic in figure 2 above. This accounts for the position of Buckie in the networks, where it was the variety of English closest to the root in 100 per cent of runs. Please note that the transcriptions in figure 14 include a system of differentiated diacritics for length intended for recognition by our phonetic comparison program, which may look slightly unfamiliar from the perspective of Scots or English dialectology.

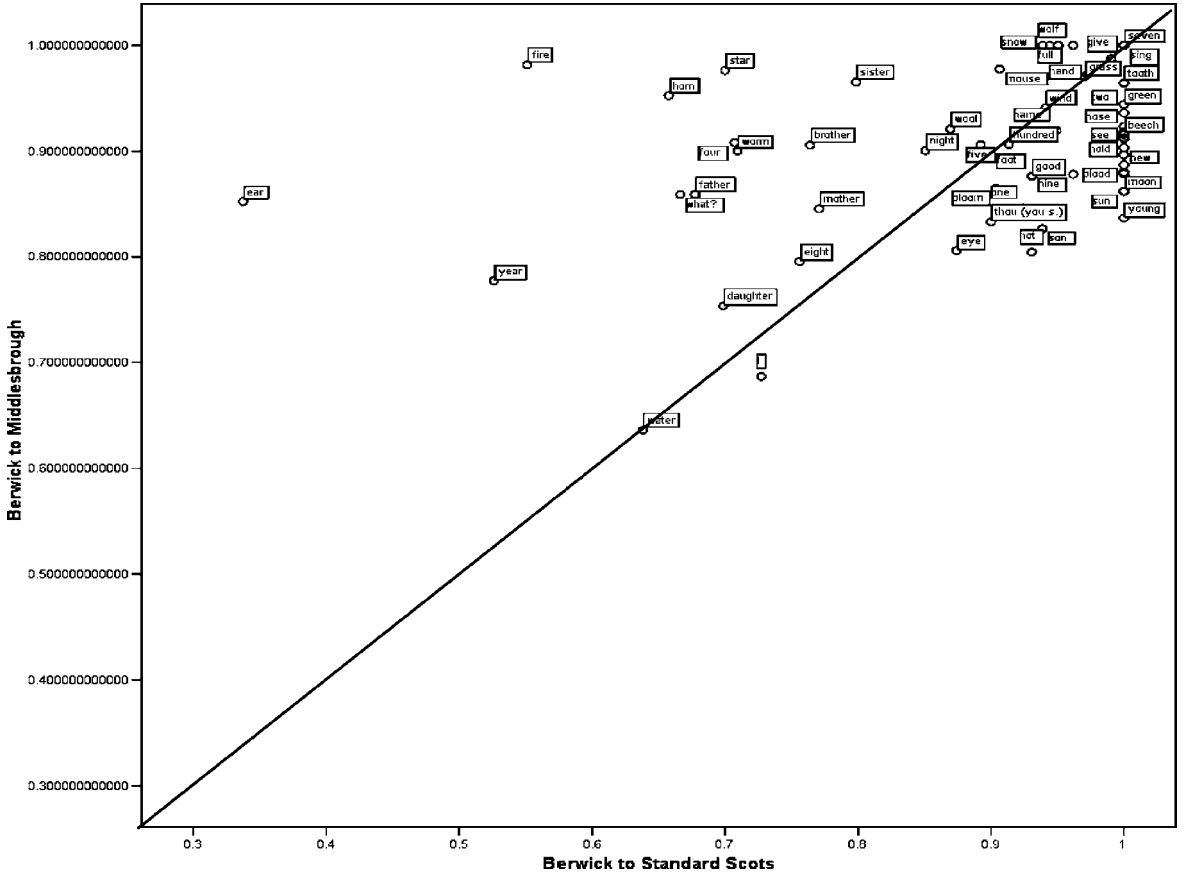


Figure 12. Berwick to Middlesbrough compared with Berwick to SSE

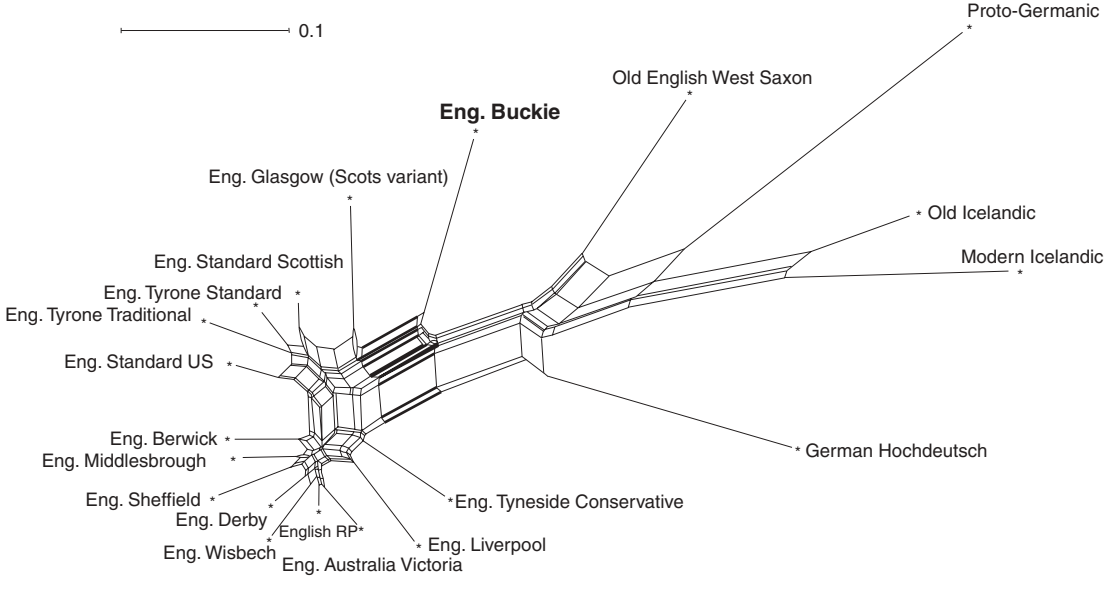


Figure 13. NeighbourNet, split isolating Buckie highlighted

	English orthographic	Proto-Germanic	SSBE	Buckie
8	*eight*	aχtoː	eɪt	ɑχt
9	*eye*	augon	aˑɪ	iˑ
14	*four*	petwor	fɔː(ɹ)	fawɵ̈r
16	*give*	ɣeβan	gɪv	giˑ
26	*knee*	knewam	niː	kəniˑ
32	*night*	nokt	naɪt	naçt
36	*salt*	saltam	sɒɬt	sa̤t
52	*two*	twai	t^{h}uː	t^{h}wɑˑ
54	*what?*	χwat	wɒt	f̬it
55	*wind*	windaz	wɪnd	wɒˑn
57	*wool*	wulloː	wʊɬ	wʊˑwəɬ
59	*year*	jæːram	jɪˑə(ɹ)	jijɵ̈ɾ

Figure 14. Cognates of the sixty in which Buckie is particularly dissimilar to other English varieties, as picked out by SPSS cross-comparisons of Glasgow ~ Buckie vs. Middlesbrough ~ Glasgow

5 Future prospects

Varieties of English have been described and compared for centuries, but until now the focus of the great majority of comparative dialectological work has been on how varieties differ, whether that involves comparing phonemic systems, phonotactic constraints, patterns of allophony, lexical incidence, or sociolinguistic and situational preferences. In this article, we hope to have demonstrated that appropriate use of quantitative methods and computational techniques allow us to take initial steps towards answering the related but separate question of how different varieties are. It has been possible to devise a linguistics-led program to produce measures of the phonetic similarity between languages and varieties; these figures can subsequently be input to computational tools developed initially for other disciplines, to produce diagrammatic representations and analyses of the particular features that shape those measures of similarity.

https://doi.org/10.1017/S1360674306002139 Published online by Cambridge University Press

Clearly, the work reported here is preliminary: there is much more to be done in terms of improving data, analysis, programming and processing. In our own research, we hope to proceed to further refinements of our phonetic similarity matching, to allow us to move beyond cognates in possible future comparisons across families. In the shorter term, we will include more varieties, and a range of speakers for at least one variety. Furthermore, we aim to extend our method from Germanic and Romance (see Heggarty, forthcoming; Heggarty, McMahon & McMahon, 2005) to the main surviving indigenous language family of the Americas, Quechua, where there are numerous questions over dialect-level subgroupings that we hope to elucidate. The results presented here are clearly not definitive, but we hope they are indicative of the prospects quantitative tools and techniques bring to comparative linguists and dialectologists in the field of phonetic similarity.

Authors' addresses:
April McMahon, Paul Heggarty, Warren Maguire
Linguistics and English Language
PPLS, University of Edinburgh
14 Buccleuch Place
Edinburgh
EH8 9LN
April.McMahon@ed.ac.uk

Robert McMahon
Molecular Genetics Laboratory
Molecular Medicine Centre
Western General Hospital
Crewe Road
Edinburgh
EH4 2XU

References

Bandelt, H.-J., P. Foster, B. C. Sykes & M. B. Richards (1995). Mitochondrial portraits of human populations using median networks. *Genetics* **141**: 743–53.

Bandelt, H.-J., P. Forster & A. Röhl (1999). Median-joining networks for inferring intraspecific phylogenies. *Molecular Biology and Evolution* **16**: 37–48.

Dyen, I., J. B. Kruskal & P. Black (1992). An Indoeuropean classification: a lexicostatistical experiment. *Transactions of the American Philosophical Society* **82**, part 5. Data available at http://www.ldc.upenn.edu, or http://www.ntu.edu.au/education/langs/ielex/.

Forster, P. & C. Renfrew (eds.) (2006). *Phylogenetic methods and the prehistory of languages*. Cambridge: McDonald Institute for Archaeological Research.

Gray, R. & Q. Atkinson (2003). Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* **426**: 435–9.

Heggarty, P. (2000). Quantifying change over time in phonetics. In Renfrew, C., A. McMahon & L. Trask (eds.), *Time depth in historical linguistics*. Cambridge: McDonald Institute for Archaeological Research. 531–62.

Heggarty, P. (2005). Enigmas en el origen de las lenguas andinas: aplicando nuevas técninas a las incógnitas por resolver. *Revista Andina* **40**: 9–80.

https://doi.org/10.1017/S1360674306002139 Published online by Cambridge University Press

Heggarty, P. (2006). Interdisciplinary indiscipline? Can phylogenetic methods meaningfully be applied to language data – and to dating language? In Forster, P. and C. Renfrew (eds.), *Phylogenetic methods and the prehistory of languages*. Cambridge: McDonald Institute for Archaeological Research: 183–94.

Heggarty, P. (forthcoming) *Measured language*. Oxford: Blackwell.

Heggarty, P., A. McMahon & R. McMahon (2005). From phonetic similarity to dialect classification: a principled approach. In Delbecque, N., J. van der Auwera & D. Geeraerts (eds.), *Perspectives in variation*. Amsterdam: Mouton De Gruyter. 43–91.

Huson, D. H. (1998). SplitsTree: a program for analysing and visualizing evolutionary data. *Bioinformatics* **14**(10): 68–73.

Huson, D. H. & D. Bryant (2005). Appplication of phylogenetic networks in evolutionary studies. To appear in *Molecular Biology and Evolution*.

Kessler, B. (2005). Phonetic comparison algorithms. In McMahon (ed.) (2005): 243–60.

Laver, J. (1994). *Principles of phonetics*. Cambridge: Cambridge University Press.

Maddieson, I. (1984). *Patterns of sounds*. Cambridge: Cambridge University Press.

McMahon, A. (ed.) (2005). *Quantitative methods in language comparison*. Special issue of *Transactions of the Philological Society* **103**(2).

McMahon, A., P. Heggarty, R. McMahon & Natalia Slaska (2005). Swadesh sublists and the benefits of borrowing: an Andean case study. In McMahon, A. (ed.) (2005): 147–70.

McMahon, A. & R. McMahon (2005). *Language classification by numbers*. Oxford: Oxford University Press.

Nakhleh, L., T. Warnow, D. Ringe & S. Evans (2005). A comparison of phylogenetic reconstruction methods on an Indo-European dataset. In McMahon, A. (ed.) (2005): 171–92.

Nerbonne, J. & W. Heeringa (1997). Measuring dialect difference phonetically. In Coleman, J. (ed.), *Workshop on computational phonology*. Madrid: Special Interest Group of the Association for Computational Linguistics. 11–18.

Nerbonne, J. & W. Heeringa (2001). Dialect areas and dialect continua. *Language Variation and Change* **13**: 375–400.

Nerbonne, J., W. Heeringa & P. Kleiwig (1999). Edit distance and dialect proximity. In Sankoff, D. & J. Kruskal (eds.), *Time warps, string edits and macromolecules: The theory and practice of sequence comparison*. Stanford: CSLI. v–xv.

Ringe, D., T. Warnow & A. Taylor (2002). Indo-European and computational cladistics. *Transactions of the Philological Society* **100**: 59–129.

Slaska, N. (2005). Lexicostatistics away from the armchair: handling people, props and problems. In McMahon, A. (ed.) (2005): 221–42.

Warnow, T. (2005). Detecting language contact in Indo-European. Unpublished talk given at Edinburgh University, 2005 available at http://www.cs.u.texas.edu/users/tandy/edinburgh-ling.pdf. Accessed 13/10/2005.

Wells, J. C. (1982). *Accents of English*. 3 volumes. Cambridge: Cambridge University Press.