Feature-based versus aggregate analyses of the DECTE corpus: Phonological and morphological variability in Tyneside English

Karen P. Corrigan, Adam Mearns, and Hermann Moisl (Newcastle University)

Abstract

The *Diachronic Electronic Corpus of Tyneside English* (DECTE) is a naturalistic spoken corpus of interviews with residents of Tyneside and surrounding areas of North East England. It updates the earlier *Newcastle Electronic Corpus of Tyneside English* (NECTE), which combined two sub-corpora dating from the late 1960s and mid 1990s, and supplements these with materials from an ongoing monitor corpus established in 2007. The first part of this paper outlines the background and development of the DECTE project. It then reviews research that has already been conducted on the corpus, comparing the different feature-based and aggregate analyses that have been employed. In doing so, we hope to highlight the crucial role that aggregate methods, such as hierarchical cluster analysis, can have in identifying and explaining the parameters that underpin aspects of language variation, and to demonstrate that such methods can and do work well in combination with feature-centric approaches.

Acknowledgments

This paper evolved from Karen Corrigan’s presentation at the "Workshop on Cross-Linguistic and Language-Internal Variation in Text and Speech" at Freiburg University. She is grateful to the organisers for the invitation and to the audience for helpful comments. The written version of the paper, which is co-written by her and other members of the DECTE team (Adam Mearns and Hermann Moisl), greatly benefitted from comments by two reviewers, which we would also like to acknowledge. Thanks are likewise due to the funder of the DECTE project, the Arts and Humanities Research Council (grant number: AH/H037691/1).

1. Introduction

The *Diachronic Electronic Corpus of Tyneside English* (DECTE, http://research.ncl.ac.uk/decte) is, like many of the data-sets discussed in this volume, a naturalistic spoken corpus. It incorporates dialectal English of the twentieth and twenty-first centuries collected from residents of Tyneside and the surrounding areas of North East England. The principal locations represented are the city of Newcastle upon Tyne, on the north side of the River Tyne, and the town of Gateshead, on the south side. However, as the project continues to progress, the geographical reach of the corpus is being extended to include speakers from other areas of the North East, such as County Durham, Northumberland and Sunderland. In this paper, we begin by tracking the development of this electronic corpus initiative over the last decade, briefly describing the history and structure of the three sub-corpora and the underlying principles which have guided the construction of DECTE. We will then discuss examples of research that has already been carried out on different facets and phases of the material, with a focus on variation in aspects of the phonetics and phonology of the dialect, and in some of its discourse and grammatical features. Through a review of these previous studies, we will illustrate some of the ideas pertinent to this volume by evaluating and comparing different feature-based and aggregate analyses that have been applied to the corpus. In doing so, we will argue that aggregate methods, such as hierarchical cluster analysis, have a crucial role to play in uncovering the nature and significance of the parameters that underpin the variability of Tyneside English, and of languages in general. In light of this, we will briefly outline our plans for new aggregate analyses of the DECTE material when the current phase of corpus building is complete. We will also suggest that aggregate analyses can and do work well in combination with feature-centric approaches, and with systematic observation and native speaker intuition, especially where the outcomes based on these different methods exhibit significant areas of congruity.

2. DECTE: background and development

The DECTE project updates the existing *Newcastle Electronic Corpus of Tyneside English* (NECTE, http://research.ncl.ac.uk/necte), which was created between 2000 and 2005 and consists of two pre-existing sub-corpora of audio-recorded Tyneside speech (Allen et al. 2007). One of these was assembled as part of the *Tyneside Linguistic Survey* (TLS) which was carried out in the late 1960s and early 1970s (Strang 1968; Pellowe et al. 1972; Pellowe and Jones 1978; Jones-Sargent 1983). The TLS was designed to determine whether systematic phonetic variation could be correlated with social characteristics. To this end, one-to-one interviews averaging 30 minutes were conducted with Tyneside speakers who were encouraged to talk about their lives, as well as being asked for judgements on certain language features and constructions. The interviews were represented in the corpus by analogue reel-to-reel audio recordings, orthographic and phonetic transcriptions of the first ten minutes or so of the recordings, and detailed social data files. The NECTE team was able to identify components relating to 114 interviews, with 37 full sets of recordings, transcriptions and social data. The other constituent part of NECTE is the corpus that was collected between 1991 and 1994 for the *Phonological Variation and Change in Contemporary Spoken English* (PVC) project (Milroy et al. 1997; Docherty and Foulkes 1999; Watt and Milroy 1999). As the name indicates, the PVC project investigated patterns of phonological variation and change. The core of its materials consists of 18 digital audio-taped interviews, of up to one hour in length, with self-selected dyads of friends or relatives, matched in terms of age and social class, who had freedom to converse on a wide range of subjects with minimal interference from the fieldworker. Only selective phonetic transcriptions of lexical items of interest were produced, and records of social data were limited to the gender, age and broadly defined socio-economic class of the participants. The focus of the NECTE project was to preserve and transform all of the available TLS and PVC material, amalgamating the two corpora into a single enhanced electronic resource that conformed to the guidelines and standards established by the Text Encoding Initiative (TEI, http://www.tei-c.org/index.xml) for the digital representation of documents in XML format. The aim was to produce a combined corpus that would be available, both to researchers and to the wider public, in a variety of formats: digitized sound, phonetic and standard orthographic transcriptions, and grammatically tagged texts, all of which were aligned and made accessible online.

There have been two important subsequent developments since NECTE was launched in 2005. Firstly, NECTE has become a partner in a collaborative programme which has created a new web-based portal called ENROLLER (http://www.gla.ac.uk/enroller). This portal is designed to provide access to an integrated and interoperable online repository which will facilitate searches within and across a range of electronic data-sets and resources,[1] thereby allowing language and literature researchers to retrieve and download comparative materials

[1] As well as NECTE, the ENROLLER portal currently incorporates a range of other resources, including *The Scottish Corpus of Text and Speech* (SCOTS), *The Dictionary of the Older Scottish Tongue*, and *The Historical Thesaurus of English* (http://www.gla.ac.uk/departments/stella/enroller/resources).

for the kind of aggregate analyses under discussion in this volume. Due to the nature of the federated resources combined in the ENROLLER scheme, the portal will thus permit comparative investigations that can explore English across regional, social and temporal space.

The second and most recent development in our corpus-building activities has been the augmentation of NECTE with the NECTE2 corpus, which was begun in 2007. NECTE2 (http://research.ncl.ac.uk/necte2) consists of digitized audio recordings and orthographic transcriptions of dyadic interviews, together with records of informant social details and other supplementary material, collected by undergraduate and postgraduate students and researchers at Newcastle University as part of a learning and teaching initiative that encompasses courses in areas such as linguistic variation and change, sociolinguistics and discourse analysis. The interviews record the language use of a variety of local informants from a range of social groups, and – as indicated above – extend the geographical domain covered in the earlier collections to include other parts of the North East of England. Successive cohorts of students add their own interviews to NECTE2, and then draw on the full set of materials that they and students of previous years have collected, using it as the basis for the analysis of a range of language features. In effect, therefore, the addition of NECTE2 means that the combined NECTE resource has become a monitor corpus, growing annually by between 100 and 150 new 60 minute interviews, while stretching back nearly five decades to the earliest material, recorded in the 1960s. The current stage of the project involves the full incorporation of the NECTE2 materials into the existing NECTE collection, to create the new *Diachronic Electronic Corpus of Tyneside English* (Figure 1). We have chosen the term “Diachronic” for this new federated corpus, not only because of the time span covered in terms of the periods in which interviews have been and continue to be collected, but also because of the span covered in terms of the lifetimes of participants, encompassing as it does almost a century from 1895, when the first speakers were born, to the early 1990s, when the youngest were born.

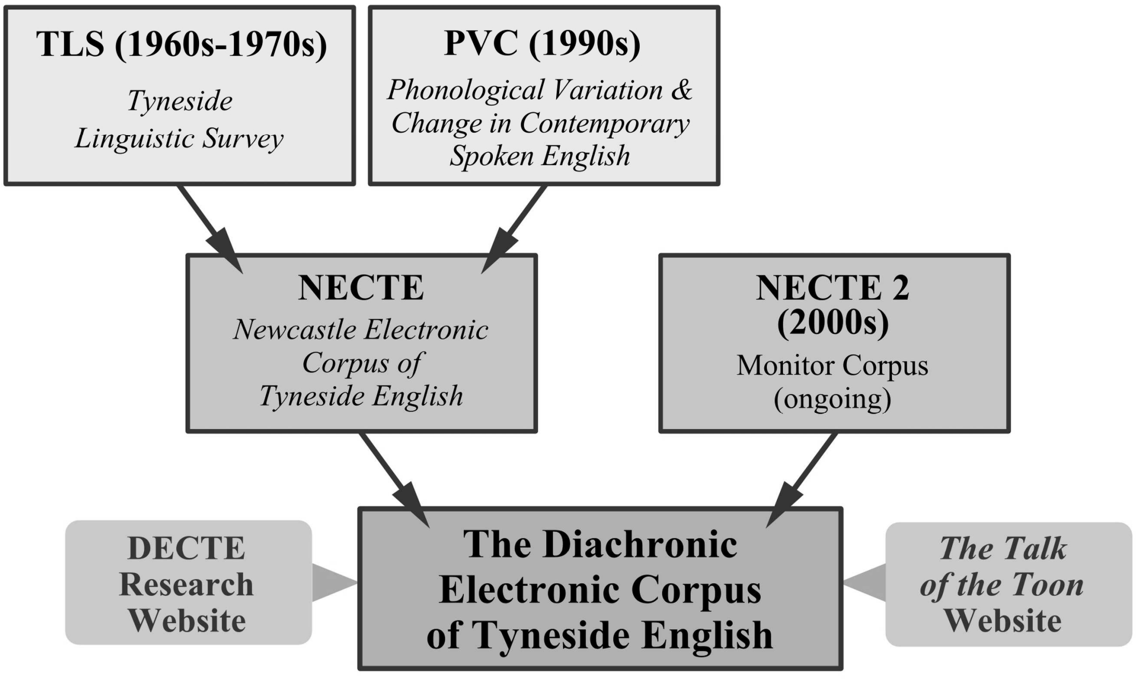

TLS (1960s-1970s)
Tyneside Linguistic Survey
PVC (1990s)
Phonological Variation & Change in Contemporary Spoken English
NECTE
Newcastle Electronic Corpus of Tyneside English
NECTE 2 (2000s)
Monitor Corpus (ongoing)
DECTE Research Website
The Diachronic Electronic Corpus of Tyneside English
The Talk of the Toon Website

Figure 1. The constituent elements of DECTE (http://research.ncl.ac.uk/decte).

This latest phase of the project will allow us to amalgamate and update all of the materials, a process which is important for two reasons. Firstly, it will enable the structure and organization of the corpus as a whole to be streamlined. Secondly, it will involve revision of the XML files around which the electronic corpus is based, so that individually and collectively they comply with the latest instantiation of the Text Encoding Initiative Guidelines, P5 (http://www.tei-c.org/Guidelines/P5), which were released in 2007, after the launch of NECTE. This is crucial for DECTE since, unlike NECTE, the intended audience is much wider than end-users only from academic domains. The intention for this enhanced version of the corpus is to engage a diverse range of user groups, with different interests and requirements. To this end, there will be two web portals offering access to the material, each with its own focus. The DECTE research website will be configured for purely academic purposes, while the *Talk of the Toon*[2] website will be geared towards users in school and museum contexts, as well as members of the general public. With this broader set of users in mind, this public-facing portal will integrate the DECTE audio and text files with still and moving images related to the sociocultural themes and aspects of local history that informants touch upon in the corpus interviews. The multimedia nature of the *Talk of the Toon* site is one feature of the current phase of corpus development that ties in with the important new encoding features of the TEI P5 Guidelines. In particular, the integration of the different elements of the site will benefit from the fact that the P5 revision involves significant changes that provide "new support for manuscript description, multimedia and graphics, standoff annotation, and representation of data pertaining to people and places", as well as "changes to the way that linking mechanisms are expressed, so that pointing to other documents will be easier" (http://www.tei-c.org/Guidelines/P5).

Having described the background to and composition of DECTE, we now move on to explore how the corpus can be exploited as a research tool for determining the nature and trajectory of variation and change in the Tyneside region, looking first at examples of feature-based approaches to the material, and then at studies that employ aggregate analysis methods.

3. Analyses of DECTE with a "feature-based" orientation

3.1. The GOAT vowel

We briefly indicated earlier that the form *toon*, in the website name *Talk of the Toon*, represents a variant [u;] pronunciation characteristic of the Tyneside English accent. This is an example of a relic feature exhibited by speakers whose phonological systems do not reflect historical changes principally associated with the Great Vowel Shift of the later Middle Ages and Early Modern period. There are other associated phonological variants occurring in Tyneside English, some of which have been subjected to feature-based accounts that display exactly the characteristics described in Szmrecsanyi and Kortmann (2009). A good case in point of research into these variants that uses data now incorporated in DECTE are the analyses undertaken by Watt and Milroy (Watt 1999, 2000, 2002; Watt and Milroy 1999), focusing on the so-called GOAT vowel (after Wells 1982).

[2] *Toon* reflects a characteristic Tyneside pronunciation of the word *town* – a relic pre-GVS pronunciation of the [aU] diphthong. It has become a synonym for Newcastle itself, and is particularly associated with the city in the context of football, with supporters of Newcastle United being known both locally and nationally as 'The Toon Army'.

Present-Day English *goat* has its origins in Old English *gât*. During the Middle English period, pronunciations in the different regional dialects that had emerged became distinctive, particularly in terms of the division between the regions to the north and to the south of the River Humber (Figure 2)

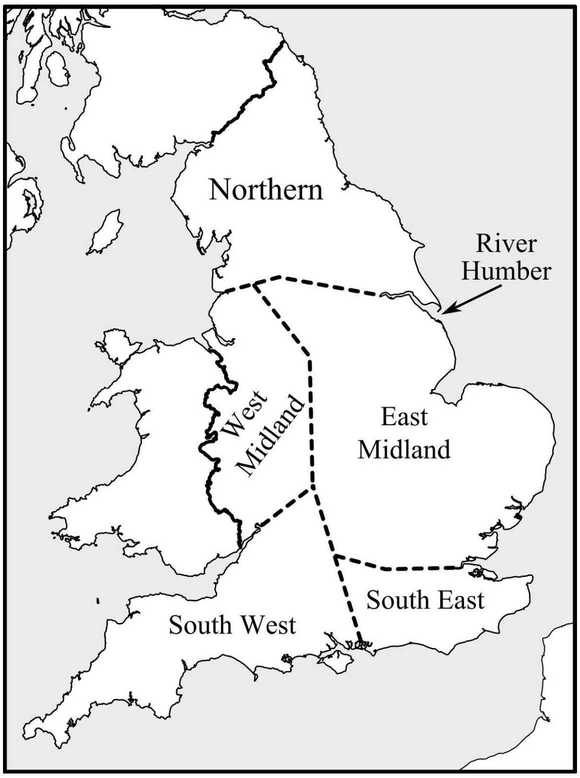


Figure 2. Middle English dialect areas, with the River Humber indicating the boundary of the Northern region (Outline image: NordNordWest 2011 CC-BY-SA-3.0, via Wikimedia Commons, http://commons.wikimedia.org/wiki/File:United_Kingdom_location_map.svg).

In Old English, the long vowel of *gât* apparently had an open back quality, /A:/. By the Middle English period, this had been fronted to /a;/ north of the Humber, that is in Northumbrian and Scots usage. Subsequently, it was generally raised through the operation of the Northern Great Vowel Shift, yielding /e;/. By contrast, in the Southumbrian area, there was a raising in Middle English from /A;/ to /O;/, and then a further raising through the operation of the Southern Great Vowel Shift to /o;/. Eventually, this latter form gave way in many places to innovative diphthongal pronunciations originating in Southern England, so that traditional RP speakers had [oU] by the early twentieth century. This variant has itself been subject to change in the post-World War II period, such that among younger speakers the starting point of the diphthong is now [@] (see Corrigan in press; Moisl et al. 2011 and Upton 2004). As with the *toon* variant, relics of pre-shift pronunciations can be found among speakers of the Tyneside dialect. Thus, in addition to those who exhibit the prestige pronunciations associated with RP, which we will see shortly are preferred among some social strata of the population in the North East of England, there are speakers recorded in DECTE who retain the older [a;] residualism of the GOAT vowel. It is, however, rare, being favoured only by the oldest generation of speakers. Indeed, even among this group it is lexically restricted to items like *know*, which has the traditional eye dialect spelling *knaa* discussed in this context in Beal et al. (2007).

The full set of GOAT variants, as isolated in feature-based accounts of DECTE's materials, is listed below (Table 1), with the first four of these variants being the most productive.

Table 1. Variants of GOAT in Tyneside English.

	GOAT Vowel
Variant 1	o;
Variant 2	U@
Variant 3	ø;
Variant 4	oU
Variant 5	a;

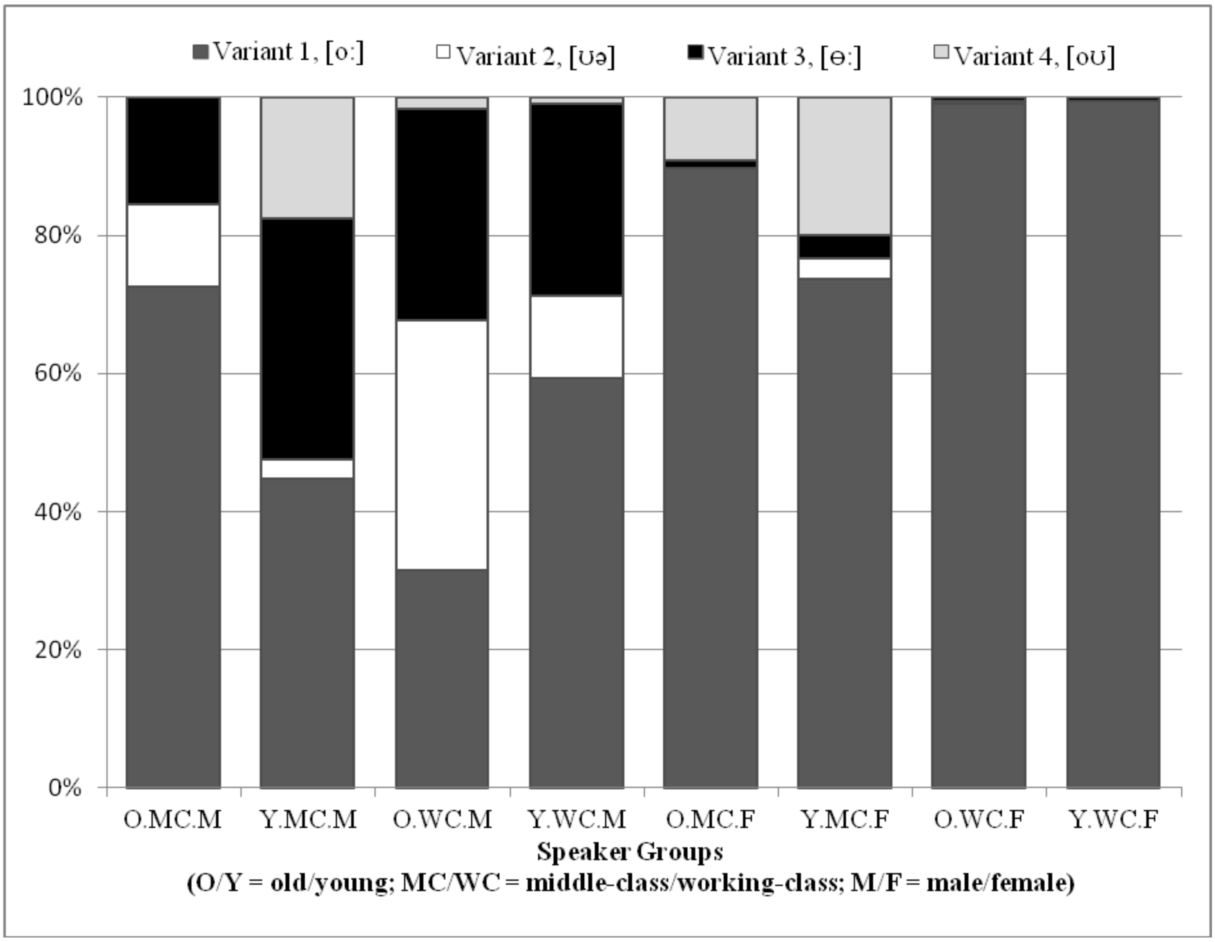
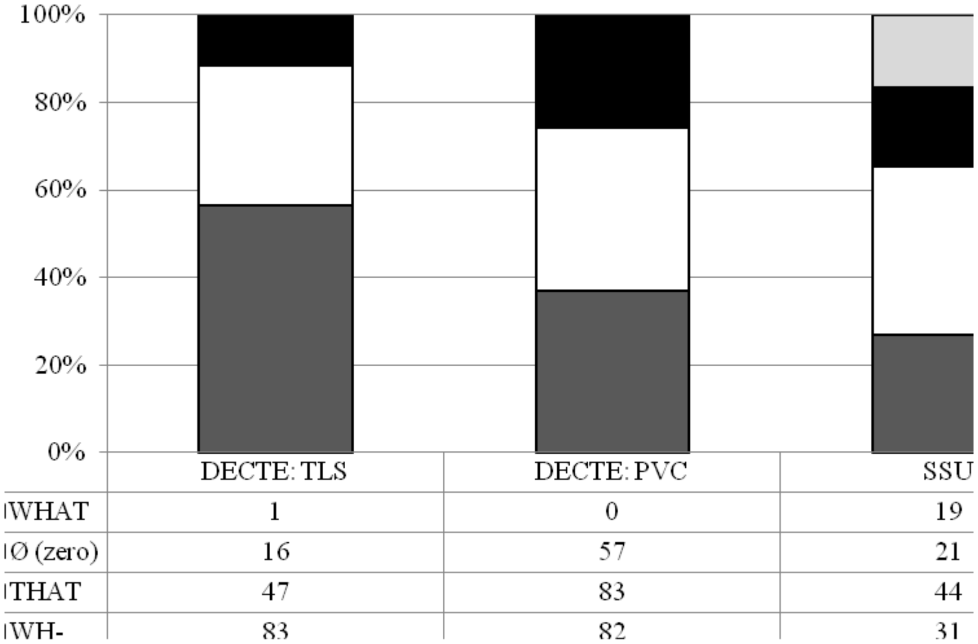


Figure 3. Percentage distribution of GOAT variants by speaker group.

Figure 3 displays the distribution of the first four variants across the community, based on Watt and Milroy’s various analyses of material from the PVC sub-corpus of DECTE. It demonstrates quite clearly that most social groups prefer Variant 1, [o;]. This suggested to Watt and Milroy (1999: 36) that [o;] is the “unmarked” Tyneside form of the GOAT vowel. Variant 4, the traditional RP diphthong pronunciation, [oU], is clearly the most prestigious form in the Tyneside community too, since it is preferred by young middle-class speakers of both genders. Watt (2000: 83) observes that [U@] and [ø;], Variants 2 and 3, can be considered “male forms”, since the former is most closely associated with males of the older working-class group, while the later is generally avoided by females of all groups. Variant 3 is also notable for being highly localized and lacking in overt prestige. Watt and Milroy (1999: 37) suggest, therefore, that Variant 3 is typically linked with those males who want to assume and project a clear sense of “local identity”.

3.2. Grammatical marking in DECTE and the *Survey of Sheffield Usage* (SSU)

The second example of feature-based research that draws on DECTE is the study of grammatical marking in the dialects of Tyneside and Sheffield by Beal and Corrigan (2007, 2011). This work investigated the distribution of a very limited set of features across time, as well as in social and regional space. It compared different speaker groups in two of the sub-corpora of DECTE (TLS and PVC), as well as in the *Survey of Sheffield Usage* (SSU), which was constructed along similar lines to the *Tyneside Linguistic Survey*, though 20 years later. Figure 4 illustrates the findings for subject and object relative clause marking in both restrictive and non-restrictive contexts.



	DECTE: TLS	DECTE: PVC	SSU
WHAT	1	0	19
Ø (zero)	16	57	21
THAT	47	83	44
WH-	83	82	31

Figure 4. Relative clause marking strategies in the dialects of Tyneside and Sheffield (number of tokens as a percentage of overall rates of occurrence).

The later corpora – the SSU of the 1980s and the PVC of the 1990s – show an increase in the usage of the zero relative variant, by comparison with the 1960s TLS data, with the figure for the most recent of the three sub-corpora (PVC) being three and a half times greater than that for the TLS, collected some thirty years earlier. There are other clear differences too. Both of the Tyneside samples have higher numbers and proportions of WH- relatives, with the SSU figures being noticeably lower especially in comparison with the frequency of this variant in the TLS sub-corpus of DECTE. In the PVC sub-corpus, TH- and WH- relatives are fairly evenly distributed, while speakers in the SSU subsample employ the TH- variant slightly more frequently than they do WH-. The use of *what* as a relative pronoun is almost nonexistent in the two Tyneside sub-corpora, but accounts for almost as many relative clauses as the zero variant in the SSU database.

Another focus of Beal and Corrigan (2011) was the distribution across regional and social space of dual form adverbs, that is forms with and without *-ly*, as in the following examples from Tagliamonte's York corpus as well as from DECTE and the SSU:

(1) *I mean I was **real-ø** small and everything you-know **really** tiny built and I was small in stature as well* (Tagliamonte and Ito 2002: 236)

(2) *that's one thing I **really** love ... getting on the back of a motorbike ... you know, **real-ø** fast, really fast* (DECTE-tlsg17)

(3) *Well it's not changing **rapidly**, it's changing **gradual-ø*** (1981: SSU/011)

As Figure 5 demonstrates, with regard to the relative frequencies of *real* – as opposed to Standard English *really* – the patterns of usage in the Tyneside and Sheffield communities appear to be similar to one another. Moreover, education, in particular, appears to play an important role in both regions, with dramatically higher levels of *real* being used by young school leavers, irrespective of their gender.

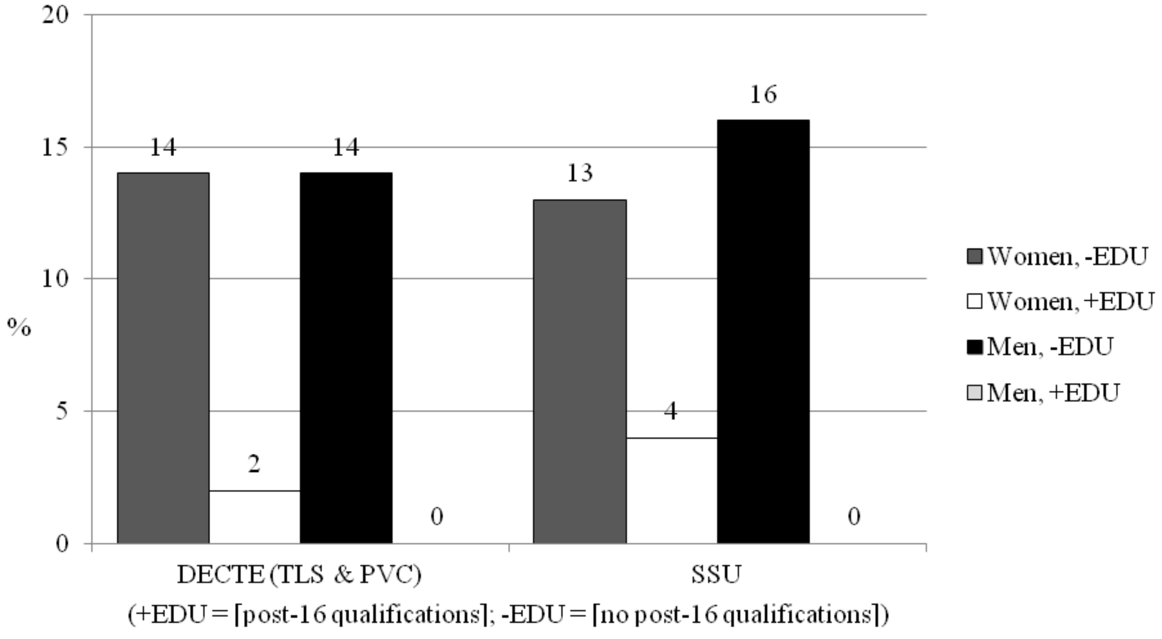


Figure 5. Percentage distribution of {Ø} *real* by gender and education in the dialects of Tyneside and Sheffield.

Figure 6 shows that education plays a similar role when looking at the zero morphological marking of other adverbs, such as *gradual* in (3), with a clear demarcation in both regions between early and late school-leavers once more apparent. However, it also reveals that in other respects the patterns of distribution are somewhat different from those we saw with *real*. Firstly, all speakers in the community (including educated men this time) use zero variants at least some of the time, and this is true of the informants from Tyneside as well as those from Sheffield. Secondly, there are a couple of interesting patterns that differentiate the two regions. Educated females from Tyneside show near categorical use of {-ly} variants, but unlike the patterning of *real*, this is not as pronounced a tendency for the same cohort of speakers in Sheffield. A similar pattern is found among men, with those in the less-educated group from Sheffield having a slightly higher tendency towards zero marking than their peers on Tyneside.

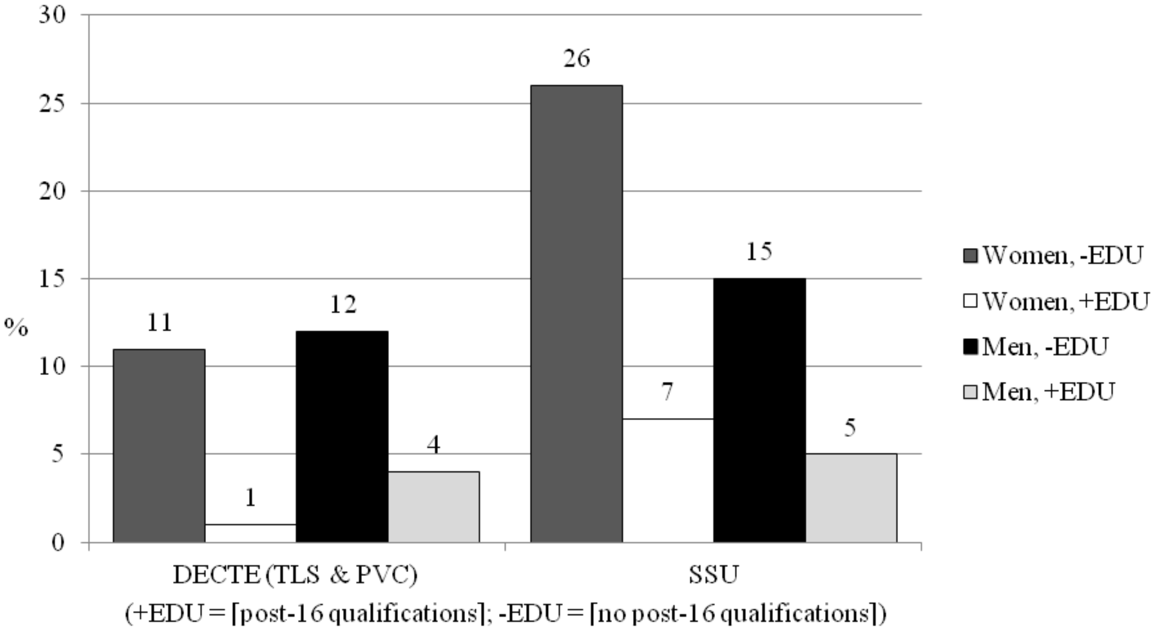


Figure 6. Percentage distribution of {Ø} in other adverbs by gender and education in the dialects of Tyneside and Sheffield.

3.3. Intensifiers

Our third example of research that involves a feature-centric approach, by Barnfield and Buchstaller (2010), also examines adverbials. In this case, the focus is on their function in discourse as intensifiers, which is another feature that seems to vary globally among English dialects. Unlike the previous examples, this study also covers the more recent NECTE2 sub-corpus of DECTE, so that we can view change across the widest time-depth possible for this database, tracing the way in which different intensifiers wax and wane through the latter part of the twentieth century and into the early part of the twenty-first.

The patterning of five of the most frequent intensifier variants across the entire period illustrates the considerable changes in the expression of intensification that have been captured in DECTE (Figure 7). Longitudinal observation of this variable clearly shows that not all variants have changed in the same way and that different types of variation can be discerned. Firstly, there is the long term competition between the high frequency variants *very* and *really*. *Very* was the dominant form in the 1960s by a considerable margin, but had declined in popularity by the 1990s to such an extent that it had fallen below a slightly rising *really*. The data from the 2000s shows it regaining some ground and once again becoming more frequent than *really*, though now only by a small margin. Barnfield and Buchstaller (2010: 273) note that the longitudinal real time analysis afforded by DECTE "reveals that the competition between these two forms is ongoing and firmly embedded in the systemic interaction within the variable as a whole", therefore offering a perspective on the rivalry between the two variants that is different from that found in apparent time studies, which suggest a more straightforward displacement of *very* by *really*. A different and rather striking pattern of change is seen with respect to *dead*. This variant did not figure at all in the TLS sub-corpus of the 1960s, but emerged abruptly in the PVC of the 1990s to supplant *very* as the most popular intensifier. With a subsequent dramatic decline between the 1990s and 2000s, the sharp rise and fall pattern it follows is the behaviour of the quintessential linguistic fad. Finally, and in stark contrast with the sudden but relatively short-lived spike associated with a fad, there has been a slow but steady rise in the usage of the two lower frequency variants, *so* and *pure*, with the result that the incidence of the former in the most recent sub-corpus is on a par with the formerly prevalent *dead*, and may well overtake it as this monitor corpus is updated over the next few years.

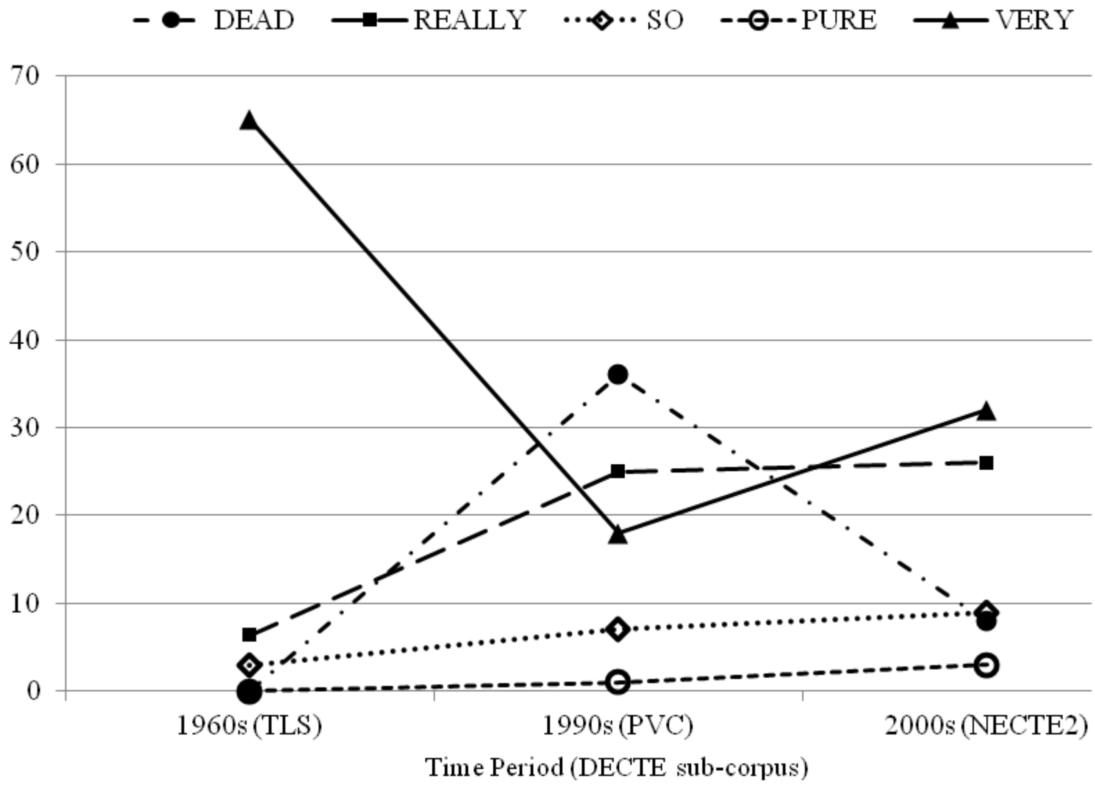


Figure 7. Trajectory of five intensifier variants across the three DECTE sub-corpora (Barnfield and Buchstaller 2010: 273).

3.4. Some characteristic limitations of feature-based analyses

Even from this brief review, it is clear that researchers have conducted quite a range of variationist studies of DECTE and its various sub-corpora of the type described in the words of Nerbonne (2008: 365) as proceeding to characterize differences in regional, social or temporal space in a "bottom-up" fashion. They focus on single or modest combinations of features, and the choice of which variables to examine is determined in large part by prior knowledge of the speech community to be observed. We would agree with Nerbonne (2008: 366) that doing so rarefies "informed intuition" over "analytic techniques" in a manner that – in some senses – is not that different from the approach of a generativist when faced with the problem of accounting for microparametric variation. It is not surprising, then, that the features in question are often well-known shibboleths, as is the case with the GOAT vowel on Tyneside (Beal 2000; Pearce 2009).

Inevitably, of course, one cannot avoid being in the position of having to make choices as to which of a potentially vast number of features should comprise one's "feature portfolio" (Szmrecsanyi, this volume). However, we would argue that the key principle here is not necessarily to avoid shibboleths, as such, but to include these alongside a wide range of other features, the function of which as markers of variation is less clearly signalled at the outset of the research. This strategy is in keeping with Nerbonne's (2006: 464) view that the portfolio should consist of a "large number of variables, even though they will contain a great deal of variation irrelevant to questions of geographic or social conditioning" so as to "provide the most accurate picture of the relations among the varieties examined."

At this point, we will briefly return to the analysis of the morphological marking of dual form adverbs in the dialects of Tyneside and Sheffield, to draw out another problematic issue raised by a feature-centric approach to geolinguistic variation. Comparing again the findings for the adverb *real* (Figure 5) with the results for other adverbs (Figure 6), we recall that there are clear differences, for example, in relation to the usage of zero variants by women in Sheffield and by educated men in both regions. It is evident, then, that the data illustrate another relevant point made by Nerbonne (2009: 185), namely that "individual features are often at odds with one another in detail, making any one of them unsuitable as a sole defining element in linguistic geography." Thus, neither of these adverb features in isolation has a convincing claim to being a good indicator of provenance.

With this in mind, we would argue that what is needed instead is an approach to variation – be it geographic, social or temporal – in which the "white noise" is eliminated. In this case, "white noise" refers to the missing data, exceptions and conflicting tendencies which arise, and which we are all familiar with from single-feature methodologies of the kind illustrated above. Signals of provenance, of diachronic change, and of social difference in language are, in fact, so complex that aggregate analyses are required in order to "see the wood for the trees", as Szmrecsanyi (this volume) has put it. Thus, we will turn now to aggregate approaches to data in DECTE, so as to demonstrate both the degree of complexity involved and some of the findings this research has produced.

4. Aggregate analyses of DECTE's TLS sub-corpus

Since the second half of the twentieth century, digital information technology has generated vast amounts of electronic data across the range of science and engineering disciplines, and, in response, these disciplines have developed mathematically and statistically based computational technologies for data interpretation. One of these technologies, cluster analysis (Gan et al. 2007; Xu and Wunsch 2009; Everitt et al. 2011), is a family of mathematically-based computational methods for identification and graphical display of structure in data where the data is too large either in terms of the number of variables or of the number of objects described, or both, for it to be readily interpretable by direct human inspection. It has long been used for this purpose in applications like hypothesis generation, hypothesis confirmation, and dimensionality reduction across a broad range of science and engineering disciplines.

Corpus linguistics has historically made little use of cluster analysis, but the recent development of the field (O'Keefe and McCarthy 2010) now makes it a potentially very useful tool for corpus-based linguistic analysis. In addition to data, the advent of information technology has generated huge amounts of electronic text in a wide range of world languages, both as corpora created specifically for linguistic research and as a result of text creation in business, government, cultural activity, and personal communication. This body of electronic text offers the linguistics research community a rich source of information about the structure and use not only of well studied languages like English but also of less intensively studied ones, dialects, endangered languages, and historically documented forms. Cluster analysis is mainly useful at the initial stages of research, particularly where the language or the linguistic phenomenon of interest is not well understood, as a way of discovering theoretically interesting structure in data abstracted from corpora which can then be used to generate linguistic hypotheses.

The *Tyneside Linguistic Survey* (Strang 1968; Pellowe et al. 1972; Jones-Sargent 1983) saw the potential of cluster analysis for corpus-based sociolinguistic and dialectological research at a time when the methodology of its application in the so-called "hard" sciences, together with the underlying mathematical theory, were in their infancy, and the computational technology necessary for its implementation was just barely up to the task. The research question the TLS team asked was: *Is there systematic phonetic variation in the Tyneside speech community as represented by NECTE, and, if so, does that variation correlate systematically with social variables?* In contrast to the then-universal and even now dominant feature-centric approach to variationist study, it proposed a fundamentally empirical methodology for finding the answer, in which salient factors were extracted from corpus data and then served as the basis for hypothesis generation. To this end, a phonetic transcription scheme analogous to the IPA was defined, and samples of the TLS audio interviews were transcribed using that scheme. These phonetic transcriptions were then cluster-analyzed and correlated with speaker-specific social data associated with the interviews, with a view to deriving and relating to one another the most important linguistic and social determinants of linguistic variation in the Tyneside area. Figure 8 shows a sample TLS cluster tree (Jones-Sargent 1983).[3]

[3] Detailed accounts of the TLS research aims, methodology, and results are in Strang (1968), Pellowe et al. (1972), and Jones-Sargent (1983).

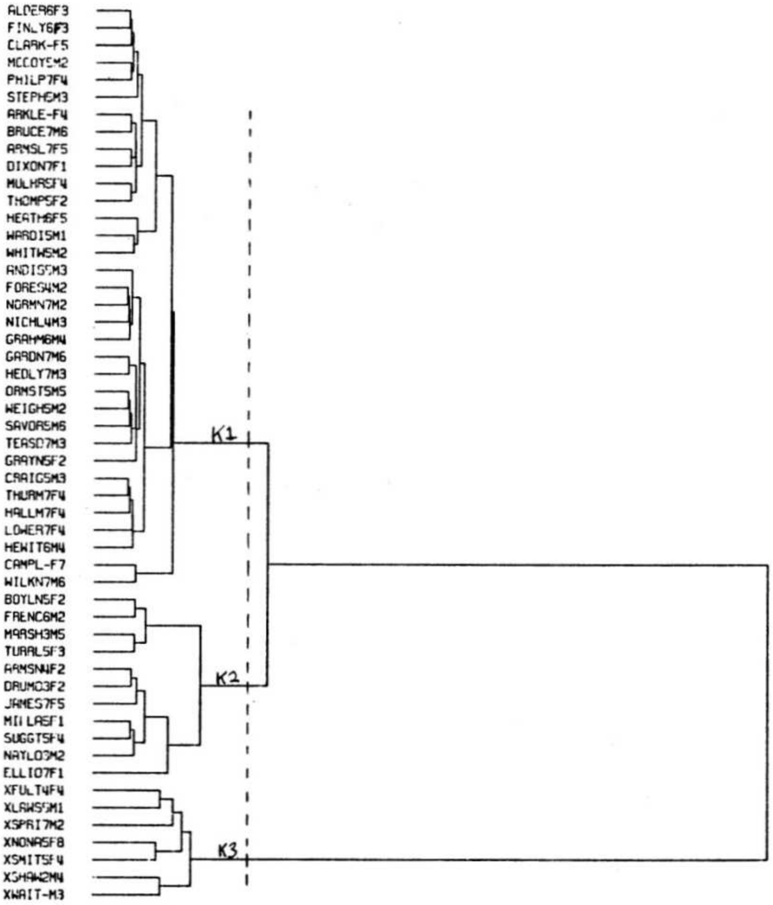


Figure 8. TLS cluster tree for segmental phonological variables in group %FON1: monophthongs.

The remainder of this section outlines the development of the TLS methodology by members of the teams that created NECTE and DECTE. The discussion is in two main parts: the first part deals with data abstraction from these corpora, and the second with the application of cluster analysis to that data.

4.1. Data creation

Data is constructed from observation of real world objects, and the process of construction raises a range of issues that determine the amenability of the data to analysis and the interpretability of the analytical results. The importance to cluster analysis of understanding such data issues can hardly be overstated. On the one hand, nothing can be discovered that is beyond the limits of the data itself. On the other, failure to understand and, where necessary, to emend relevant characteristics of data can lead to results and interpretations that are distorted or even worthless. For these reasons, an outline of data issues is given before moving on to discussion of cluster analytical methods.

Data is a description of objects from a domain of interest in terms of a set of variables (or: features) such that each variable is assigned a value for each of the objects. Given *m* objects described by *n* variables, the standard representation of data for computational analysis across the sciences generally is a matrix M in which each of the *m* rows represents a different object, each of the *n* columns represents a different variable, and the value at $M_{i.j}$ describes object *i* in terms of variable *j*, for *i* = 1..*m*, *j* = 1..*n*. The matrix thereby makes the link between the researcher's conceptualization of the domain in terms of the semantics of the variables s/he has chosen and the actual state of the world, and allows the resulting data to be taken as a representation of the domain based on empirical observation. This representation of data is the vector-space model extensively used in language technologies such as information retrieval and data mining, and assessed in terms of its applicability to dialectometry by Heeringa (2004).

The TLS component of NECTE includes 64 phonetic transcriptions of about 10 minutes from each of the 64 audio recordings (Allen et al. 2007). The data representing a single transcription is a 156-element vector *t*, each of whose elements represents a different phonetic segment in the TLS transcription scheme, and the value at any given element t_j (for *j* = 1..156) is the frequency of occurrence of segment *j* in the transcription. The vector *t* is taken to be a description of the phonetic usage of the speaker corresponding to the transcription; Table 2 gives an example.

Table 2. Vector representation of a single NECTE speaker's phonetic usage.

J	1	2	3	...	156
Phonetic segment	g	i	T	...	3;
Transcription frequency	31	28	123	...	0

The speaker represented by the vector in Table 2 uses phonetic segment *g* 31 times, *i* 28 times, and so on. The set of speaker vectors is assembled into a matrix M in which the rows *i* (for *i* = 1..*n*, where *n* is the number of speakers) represent the 64 speakers, the columns *j* (for *j* = 1..156) represent the phonetic segment variables, and the value at $M_{i,j}$ is the number of times speaker *i* uses the phonetic segment *j*. A fragment of this 64 × 156 matrix M is shown in Table 3.

Table 3. Fragment of the NECTE data matrix M.

J	1	2	3	...	156
Phonetic segment	g	i	t	...	3;
Speaker 1 transcription frequency	31	28	123	...	0
Speaker 2 transcription frequency	22	8	124	...	0
⋮	⋮	⋮	⋮		⋮
Speaker 64 transcription frequency	19	3	73	...	0

The matrix in Table 3 was the basis for the TLS phonetic analysis and remains so in the more recent work described in what follows. That more recent work has, however, identified the need for two types of data transformation prior to clustering: length normalization and dimensionality reduction.[4]

4.1.1. Length normalization

The transcriptions from which M is abstracted vary substantially in length, as shown in Figure 11. The horizontal axis represents the transcriptions 1..64 and the vertical axis the number of codes per transcription; it is clear from Figure 9 that there are a few relatively long transcriptions and a few relatively short ones, with most fairly constant in length between them. This variation in interview length can skew cluster analysis for an intuitively straightforward reason: frequencies for a longer interview will tend to be larger than those for a shorter one, and, because cluster analysis of M is based on these frequencies, there will be a tendency to cluster by transcription length rather than by the more interesting pattern of phonetic variation.

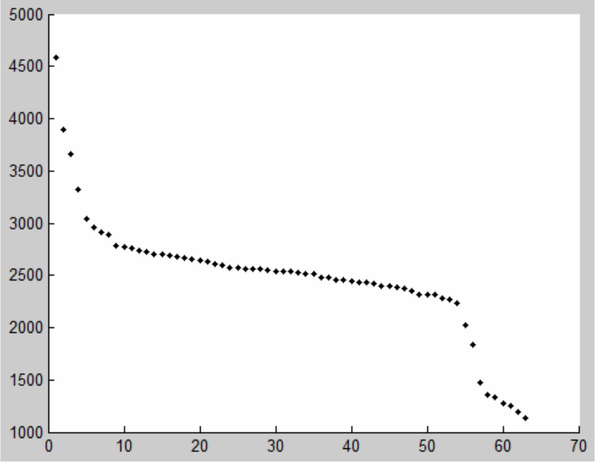


Figure 9. Number of phonetic segments in each of the 64 NECTE transcriptions

The solution is to normalize the values in the data matrix (Spärck-Jones et al. 2000; Moisl 2010b), which involves transformation of the row vectors of the data matrix in relation to some normalization factor. Three such factors are briefly considered here with reference to our data matrix M.

(i) Normalization by mean transcription length

The values in each row vector M_i, for i = 1..the number of transcriptions m in the corpus, are multiplied by the ratio of the mean number of segments per transcription across all m transcriptions to the number of segments in transcription t_i:

$$M_i = M_i \left(\frac{\mu}{nrofsegments(t_i)} \right) \qquad (1)$$

where

[4] A third data issue is increasingly recognised in data processing research generally: data nonlinearity. The NECTE data matrix M has been found to be substantially nonlinear (Moisl and Jones 2005), and current work is investigating the implications of this (Moisl 2007).

- M_i is the data matrix row representing the frequency profile of transcription $t_{i,}$.
- *nrofsegments*(t_i) is the total number of phonetic segments in t_i.
- μ *is* the mean number of segments across all m transcriptions:

$$\mu = \sum_{i=1..m} \frac{nrofsegments(t_i)}{m} \qquad (2)$$

The longer the transcription, the numerically smaller the ratio, and vice versa; the effect is to decrease the values in the vectors that represent longer transcriptions, and increase them in vectors that represent shorter ones, relative to average transcription length.

(ii) Normalization by individual transcription length

The values in each row vector M_i are divided by the number of segments in the corresponding transcription t_i:

$$M_i = M_i \left(\frac{1}{nrofsegments(t_i)}\right) \qquad (3)$$

This scales the values in M_i in relation to the number of segments in the interview that M_i represents, thereby eliminating variation in interview length as a factor.

(iii) Cosine normalization

The values in each row vector M_i are divided by its length:

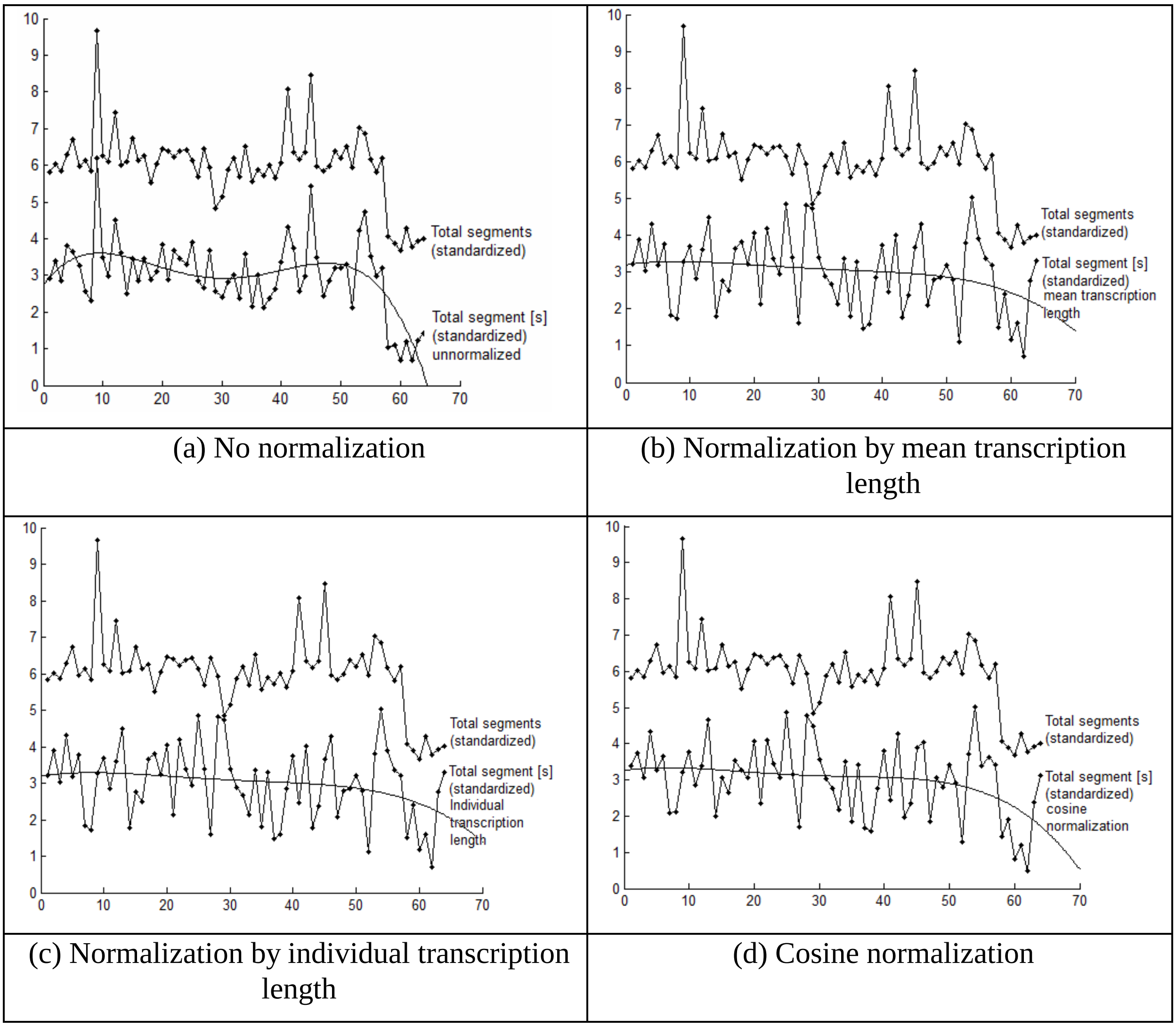
$$M_i = M_i \left(\frac{1}{\|M_i\|}\right) \qquad (4)$$

where, for j = 1..the number of columns / phonetic variables in M,

$$\|M_i\| = \sqrt{\sum_{j=1..n} M_{i,j}{}^2} \qquad (5)$$

This normalization transforms M_i to a unit vector, that is, a vector of length 1, and, because all the M_i are now equally long, variation in interview length can no longer be a factor. This normalization method is called “cosine” because the inner product of unit vectors is the cosine of the angle between them, which is standardly used in the Information Retrieval community as a measure of the distance between the vectors in a data space (for example Singhal et al. 1996; Manning et al. 2008: 110–113).

The effect of these normalizations on the values in the data variables can be seen by examining the column vectors of M. Figure 10 shows this for column 4 of M, the segment [s], though any other column would have done as well. The normalizations generate values on substantially different numerical scales, and, to permit comparison, all vectors were standardized using Student’s *t*-statistic (Boslaugh and Watters 2008: chapter 19), which reduces everything to a common scale but leaves the shapes of the distributions of values in the various vectors unaltered.



(a) No normalization

(b) Normalization by mean transcription length

(c) Normalization by individual transcription length

(d) Cosine normalization

Figure 10. Effect of various normalization measures on values of the variable [s] in the NECTE data matrix M.

In Figures 10(a) to 10(d), the horizontal axis indicates the interviews 1..64, the vertical axis represents a (standardized) count of phonetic segments, the upper line depicts the (standardized) number of phonetic segments in each of the interviews 1-64, and the lower line plots the (standardized) number of instances of the segment [s], that is, the values in column 4 of M. The curves through the lower plots are fourth-degree polynomial lines of best fit, whose purpose is explained below.

Figure 10(a) exemplifies the earlier observation that frequencies for a longer interview will tend to be larger than those for a shorter one, in that visual inspection shows the shape of the lower plot following that of the upper plot quite closely; this is especially clear for variable 10 and variables 57-64. By comparison, visual inspection shows that the shape of the lower plot in Figure 10(b) follows the upper one far less closely, with formerly lower frequencies increased and formerly higher frequencies diminished; again, variables 10 and 57-64 exemplify this most clearly. The lines of best fit confirm these visual impressions: the one in Figure 10(b) is much flatter than the one in 10(a). Finally, visual inspection of Figures 10(c) and 10(d) suggests that the corresponding normalizations yield results which look very similar or even identical to that of 10(b).

Which normalization is best? Mean transcription length and individual transcription length are linear variants of each other, as shown in equations (1) and (3) above, and they consequently give structurally identical results, as the corresponding Figures 10(b) and 10(c) suggest. Reference to equation (4) suggests that it too is simply a linear variant of (1) and (3), but careful inspection of Figure 10(d) reveals that it differs slightly from 10(b) and 10(c). These differences arise from nonlinearities which cosine normalization introduces. These nonlinearities distort the data unnecessarily, and hence play havoc with subsequent cluster analysis (see Moisl 2010a for a more detailed discussion). The choice, therefore, is between mean transcription length and individual transcription length normalizations. In this paper, we arbitrarily select the former.

4.1.2. Dimensionality reduction

Sparsity is a major issue in data analysis generally because obtaining enough data is usually difficult or even intractable as the dimensionality of the dataset grows. “Dimensionality” refers to the number of variables / columns used to describe objects in a data matrix (Lee and Verleysen 2007); this is an aspect of the “curse of dimensionality” so often cited in the data processing literature. The problem is that the space in which the data is embedded grows very quickly with dimensionality and, to keep the data from becoming disfunctionally sparse, more and more data is required until, equally quickly, obtaining enough data becomes impossible.

Assume, for example, some bivariate data in which both variables record frequency in the range 0..9: the number of possible vectors like (0,9), (3,4), and so on is $10 \times 10 = 100$. For trivariate frequency data the number of possible vectors like (0,9,2) and (3,4,7) is $10 \times 10 \times 10 = 1000$. In general, in the case of integer data, the number of possible vectors is r^d, where r is the measurement range (here 0..9) and d the dimensionality. The r^d function generates an extremely rapid increase in data space size with dimensionality: even a modest $d = 8$ for a 0..9 range allows for 100,000,000 different vectors. This is a problem because the larger the dimensionality, the more difficult it becomes to maintain a degree of data density sufficient to yield reliable analytical results.

To see why, assume that we want to analyse, say, 24 speakers in terms of their usage frequency of 2 phonetic segments; assume also that occurrence of these segments is rare, so a range of 0..9 is sufficient. The ratio of actual to possible vectors in the space is 24 / 100 = 0.24, that is, the vectors occupy 24% of the data space. If one analyses the 24 speakers in terms of 3 phonetic segments, the ratio of actual to possible vectors is 24 / 1000 = 0.024 or 2.4 % of the data space. In the 8-dimensional case it is 24 / 100,000,000, or 0.00000024 %. A fixed number of vectors occupies proportionately less and less of the data space with increasing dimensionality. In other words, the data space becomes so sparsely inhabited by vectors that any relationships among them are increasingly difficult to discern by analysis. What about using more data? Let’s say that 24% occupancy of the data space is judged to be adequate for reliable analysis. To achieve that for the 3-dimensional case one would need 240 vectors (that is, speakers), 2400 for the 4-dimensional case, and 24,000,000 for the 8-dimensional one. This may or may not be possible. And what are the prospects for dimensionalities higher than 8?

Because provision of additional data to improve the definition of a sparse manifold is not always, or even usually, possible, much research has addressed ways of reducing data dimensionality.[5] The dimensionality of the NECTE data matrix is 156 because 156 variables are used to describe each speaker, but there are only 64 speakers in the 156-dimensional space. In other words, the NECTE data is extremely sparse and needs to be dimensionality-reduced as much as possible. Various reduction methods have been experimented with in earlier work on the NECTE data to achieve this, but in the present instance one of the intuitively most straightforward reduction method is used. As will be seen in the next part of the discussion, cluster analysis groups data objects on the basis of the degree to which they differ with respect to the variables used to describe them. For cluster analytic purposes, therefore, a variable is useful in direct proportion to the amount of variation in the values that it takes: a variable with substantial variability will be very useful as a clustering criterion, one with moderate variability will be moderately useful, and one with little or no variability will be of little or no use at all. One of the ways to reduce dimensionality, therefore, is to eliminate from the data matrix variables which are only marginally useful in this sense.

An obvious way to do this is to calculate the statistical variance of the data matrix columns and to discard those which fall below some predefined threshold of usefulness. In this spirit, the column variances of the mean-transcription-length normalized NECTE data matrix M were calculated, sorted in descending order of magnitude, and plotted; the plot is shown in Figure 11.

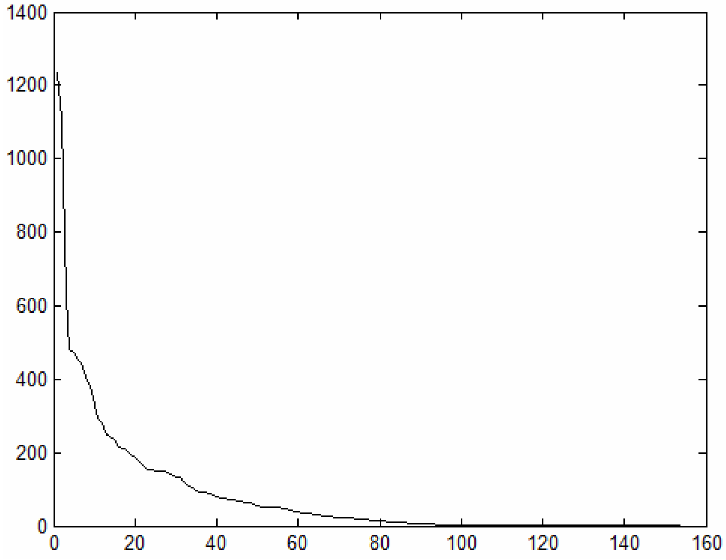


Figure 11. Plot of column variances of the NETE data matrix M sorted in descending order of magnitude.

There are a few relatively high-variance variables, a large number of relatively low-variance variables, and a moderate number of intermediate-valued ones in between. The variances of the variables to the right of, say, the 50th seem so low relative to the high- and intermediate-variance ones that they can be eliminated, thereby achieving a very substantial dimensionality

[5] Numerous such methods have been developed and the associated literature is extensive. As such, there is no hope even of outlining the topic in a brief discussion like this one. For recent overviews, see Lee and Verleysen (2007) and Carreira-Perpinan (2011).

reduction from 156 to 50. Why 50 and not, say, 45 or 70? There is no definitive answer; where to place the threshold is a matter of researcher judgment.

4.2. Cluster analysis

Cluster analysis is primarily a tool for data exploration and subsequent hypothesis generation, and it was and is used as such both by the TLS researchers and more recently for analysis of the NECTE phonetic transcriptions (Moisl et al. 2006; Moisl and Maguire 2008). It identifies structure latent in data, and awareness of such structure can be used to draw the inferences on the basis of which hypotheses are formulated. To see how this works, assume that the research question is the original TLS one: *Is there systematic phonetic variation in the Tyneside speech community, and, if so, does that variation correlate systematically with social structure?*

To keep the examples tractable in what follows, a random subset of 24 rows of M representing the phonetic usage of 24 NECTE speakers is selected. To start, only one of the available 156 phonetic variables is used to differentiate the speakers, as shown in Figure 14.

Table 4. Frequency data for $@_1$

Speaker	**$@_1$**
tlsg01	3
tlsg02	8
tlsg03	3
tlsn01	100
tlsg04	15
tlsg05	14
tlsg06	5
tlsn02	103
tlsg07	5
tlsg08	3
tlsg09	5
tlsg10	6
tlsn03	142
tlsn04	110
tlsg11	3
tlsg12	2
tlsg52	11
tlsg53	6
tlsn05	145
tlsn06	109
tlsg54	3
tlsg55	7
tlsg56	12
tlsn07	104

It is easy to see by direct inspection of the data that the speakers fall into two groups: those that use $?_1$ relatively frequently and those that use it infrequently. Based on this result, the obvious hypothesis is that there is systematic variation in phonetic usage with respect to $?_1$ in the speech community.

If two phonetic variables are used, as in Figure 15, direct inspection again shows two groups, those that use both $@_1$ and $@_2$ relatively frequently and those that do not, and the hypothesis is analogous to the one just stated.

Table 5. Frequency data for $@_1$ and $@_{2.}$

Speaker	**$@_1$**	**$@_2$**
tlsg01	3	1
tlsg02	8	0
tlsg03	3	1
tlsn01	100	116
tlsg04	15	0
tlsg05	14	6
tlsg06	5	0
tlsn02	103	93
tlsg07	5	0
tlsg08	3	0
tlsg09	5	0
tlsg10	6	0
tlsn03	142	107
tlsn04	110	120
tlsg11	3	0
tlsg12	2	0
tlsg52	11	1
tlsg53	6	0
tlsn05	145	102
tlsn06	109	107
tlsg54	3	0
tlsg55	7	0
tlsg56	12	0
tlsn07	104	93

There is no theoretical limit to the number of variables that can be used. As the number of variables and observations grows, so does the difficulty of generating hypotheses from direct inspection of the data. In the present case, the selection of $@_1$ and $@_2$ in Tables 4 and 5 was arbitrary, and the speakers could have been described using more phonetic segment variables. Table 6 shows twelve.

Table 6. Frequency data for a range of phonetic segments.

Speaker	**$@_1$**	**$@_2$**	**o;**	**$@_3$**	**ī**	**eī**	**n**	**$a;_1$**	**$a;_2$**	**aī**	**r**	**w**
tlsg01	3	1	55	101	33	26	193	64	1	8	54	96
tlsg02	8	0	11	82	31	44	205	54	64	8	83	88
tlsg03	3	1	55	101	33	26	193	64	15	8	54	96
tlsn01	100	116	5	17	75	0	179	64	0	19	46	62
tlsg04	15	0	12	75	21	23	186	57	6	12	32	97
tlsg05	14	6	45	70	49	0	188	40	0	45	72	79
tlsg06	5	0	40	70	32	22	183	46	0	2	37	117
tlsn02	103	93	7	5	87	27	241	52	0	1	19	72
tlsg07	5	0	11	58	44	31	195	87	12	4	28	93
tlsg08	3	0	44	63	31	44	140	47	0	5	43	106
tlsg09	5	0	30	103	68	10	177	35	0	33	52	96
tlsg10	6	0	89	61	20	33	177	37	0	4	63	97
tlsn03	142	107	2	15	94	0	234	15	0	25	28	118
tlsn04	110	120	0	21	100	0	237	4	0	61	21	62
tlsg11	3	0	61	55	27	19	205	88	0	4	47	94
tlsg12	2	0	9	42	43	41	213	39	31	5	68	124
tlsg52	11	1	29	75	34	22	206	46	0	29	34	93
tlsg53	6	0	49	66	41	32	177	52	9	1	68	74
tlsn05	145	102	4	6	100	0	208	51	0	22	61	104
tlsn06	109	107	0	7	111	0	220	38	0	26	19	70
tlsg54	3	0	8	81	22	27	239	30	32	8	80	116
tlsg55	7	0	12	57	37	20	187	77	41	4	58	101
tlsg56	12	0	21	59	31	40	164	52	17	6	45	103
tlsn07	104	93	0	11	108	0	194	5	0	66	33	69

What hypothesis would one formulate from inspection of the data in Table 6, taking into account all the variables? And what about, say, all 64 NECTE speakers and 156 variables? These questions are clearly rhetorical, and there is a straightforward moral: human cognitive makeup is unsuited to seeing regularities in anything but the smallest collections of numerical data. To see the regularities we need help, and that is what cluster analysis provides.

Cluster analysis is a family of computational methods for identification and graphical display of structure in data when the data is too large either in terms of the number of variables or of the number of objects described (or both) for it to be readily interpretable by direct inspection, as already noted. All the members of the family work by partitioning a set of objects in the domain of interest into disjoint subsets in accordance with how relatively similar those objects are in terms of the variables that describe them. The objects of interest in Figures 14-16 are speakers, and each speaker’s phonetic usage is described by a set of variables. Any two speakers’ phonetic usage will be more or less similar depending on how similar their respective variable values are: if the values are identical then so are the speakers in terms of their phonetic usage, and the greater the divergence in values the greater the differences in usage. Cluster analysis of the data groups 24 speakers in terms of how similar their frequency of usage of 12 phonetic segments is. There are various kinds of cluster analysis (Gan et al. 2007; Xu and Wunsch 2009; Everitt et al. 2011) and Figure 12 shows the results from the one most often used, namely, hierarchical cluster analysis.

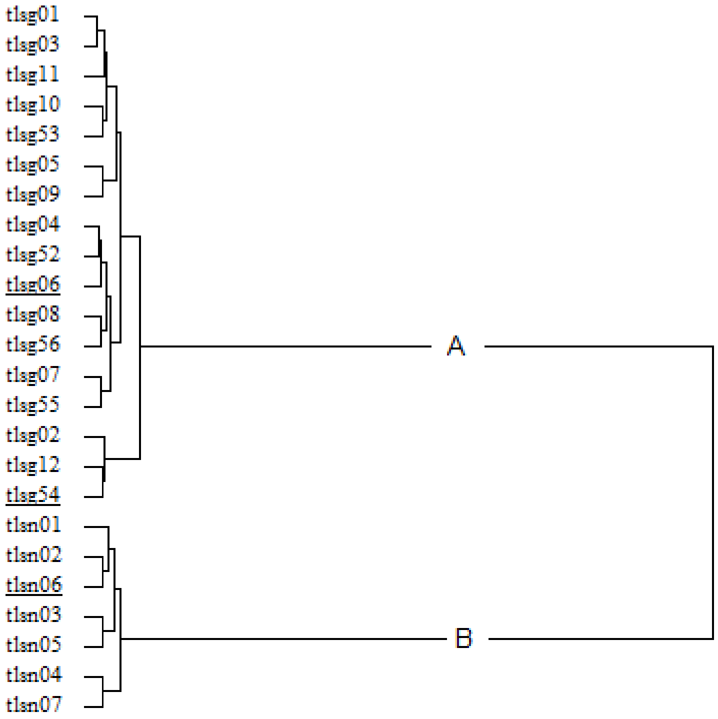


Figure 12. Hierarchical cluster analysis of the data in Table 6, using squared Euclidean distance and Ward's Method.

Figure 12 shows the cluster structure of the speaker data as a hierarchical tree (a "dendrogram"). To interpret the tree one has to understand how it is constructed, so a short intuitive account is given here. The labels at the leaves of the tree are the speaker-identifiers. These labels are partitioned into clusters in a sequence of steps. Initially, each speaker is taken to be a cluster on his or her own. At the first step, the data is searched to identify the two most similar clusters which, when found, are joined into a superordinate cluster in which their degree of similarity is graphically represented as the length of the horizontal lines joining the subclusters: the more similar the subclusters, the shorter the lines. This procedure is then applied recursively in a sequence of steps to cluster pairs until only a single cluster remains, which is the completed cluster tree.

In the actual clustering procedure assessment of similarity is done numerically, but for present expository purposes visual inspection of Figure 12 is sufficient, and, to judge by the shortness of the horizontal lines, the singleton clusters tlsg01 and tlsg03 at the top of the tree are the most similar. These are joined into a composite cluster (tlsg01 tlsg03). At the second step the data is searched again to determine the next-most-similar pair of clusters. Visual inspection indicates that these are tlsg04 and tlsg52 about a third of the way down the tree, and these are joined into a composite cluster (tlsg04 tlsg52). At step 3, the two most similar clusters are the composite cluster (tlsg04 tlsg52) constructed at step 2 and tlsg06. These are joined into a superordinate cluster ((tlsg06 tlsg56) tlsg06). The sequence of steps continues in this way, combining the most similar pair of clusters at each step, and stops when there is only one cluster remaining, which contains all the subclusters.

The resulting tree gives an exhaustive graphical representation of the similarity relations in the speaker data. It shows that there are two main groups of speakers, labelled A and B, which differ greatly from one another in terms of phonetic usage, and, though there are differences in usage among the speakers in those two main groups, these are minor relative to that between A and B.

Once the structure of the data has been identified by cluster analysis it can be used for hypothesis generation. Based on the analysis in Figure 12 the obvious hypothesis is that, with respect to the selected phonetic variables, the speakers in the community from which the data was drawn fall into two distinct groups. This hypothesis can, moreover, be elaborated in accordance with particular research aims. To a dialectologist, for example, the interest might lie not only in knowing that there is systematic variation in linguistic usage among speakers, but also in identifying the most important phonetic determinants of that variation. The main interest in Figure 12 is what differentiates clusters A and B. One approach to finding out is to create summary descriptions of the phonetic characteristics of these two main clusters and then to compare them (Moisl and Maguire 2008). This is done by taking the mean of the variable values for the speakers in each cluster, thereby creating cluster centroid vectors, as in Table 7.

Table 7. Centroids for clusters A and B in Figure 12

Cluster A												
Speaker	**@$_1$**	**@$_2$**	**o;**	**@$_3$**	**ī**	**eī**	**N**	**a;$_1$**	**a;$_2$**	**aī**	**r**	**w**
tlsg01	3	1	55	101	33	26	193	64	1	8	54	96
tlsg03	3	1	55	101	33	26	193	64	15	8	54	96
tlsg04	15	0	12	75	21	23	186	57	6	12	32	97
tlsg55	7	0	12	57	37	20	187	77	41	4	58	101
tlsg07	5	0	11	58	44	31	195	87	12	4	28	93
tlsg11	3	0	61	55	27	19	205	88	0	4	47	94
tlsg06	5	0	40	70	32	22	183	46	0	2	37	117
tlsg56	12	0	21	59	31	40	164	52	17	6	45	103
tlsg08	3	0	44	63	31	44	140	47	0	5	43	106
tlsg10	6	0	89	61	20	33	177	37	0	4	63	97
tlsg53	6	0	49	66	41	32	177	52	9	1	68	74
tlsg05	14	6	45	70	49	0	188	40	0	45	72	79
tlsg09	5	0	30	103	68	10	177	35	0	33	52	96
tlsg52	11	1	29	75	34	22	206	46	0	29	34	93
tlsg02	8	0	11	82	31	44	205	54	64	8	83	88
tlsg12	2	0	9	42	43	41	213	39	31	5	68	124
tlsg54	3	0	8	81	22	27	239	30	32	8	80	116
Centroid A	6.53	0.53	34.18	71.71	35.12	27.06	189.88	53.82	13.41	10.94	54.00	98.24

Cluster B												
Speaker	**@$_1$**	**@$_2$**	**o;**	**@$_3$**	**ī**	**eī**	**N**	**a;$_1$**	**a;$_2$**	**aī**	**r**	**w**
tlsn01	100	116	5	17	75	0	179	64	0	19	46	62
tlsn04	110	120	0	21	100	0	237	4	0	61	21	62
tlsn07	104	93	0	11	108	0	194	5	0	66	33	69
tlsn02	103	93	7	5	87	27	241	52	0	1	19	72
tlsn06	109	107	0	7	111	0	220	38	0	26	19	70
tlsn03	142	107	2	15	94	0	234	15	0	25	28	118
tlsn05	145	102	4	6	100	0	208	51	0	22	61	104
Centroid B	116.14	105.43	2.57	11.71	96.43	3.86	216.14	32.71	0.00	31.43	32.43	79.57

All the speakers whom the cluster tree assigns to A are collected in the cluster A table in Table 7. The mean of each column in cluster A is calculated and shown at the bottom of the table, and the vector of 12 values then represents the average phonetic characteristics of the speakers in A. The same is done for B. A and B can now be compared and the bar plot in Figure 13 shows the result graphically. The relative degrees of disparity in phonetic usage are shown by the differences in the heights of the bars representing A and B.

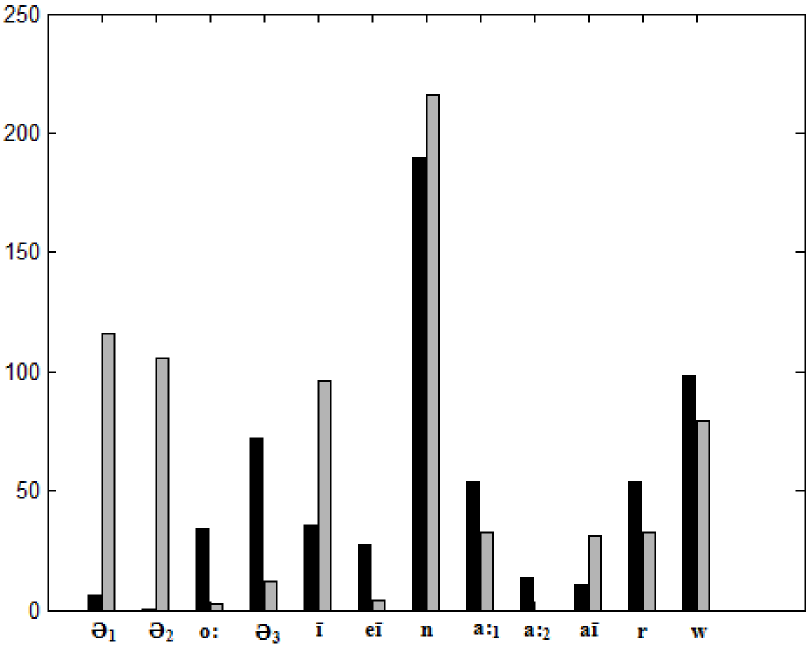
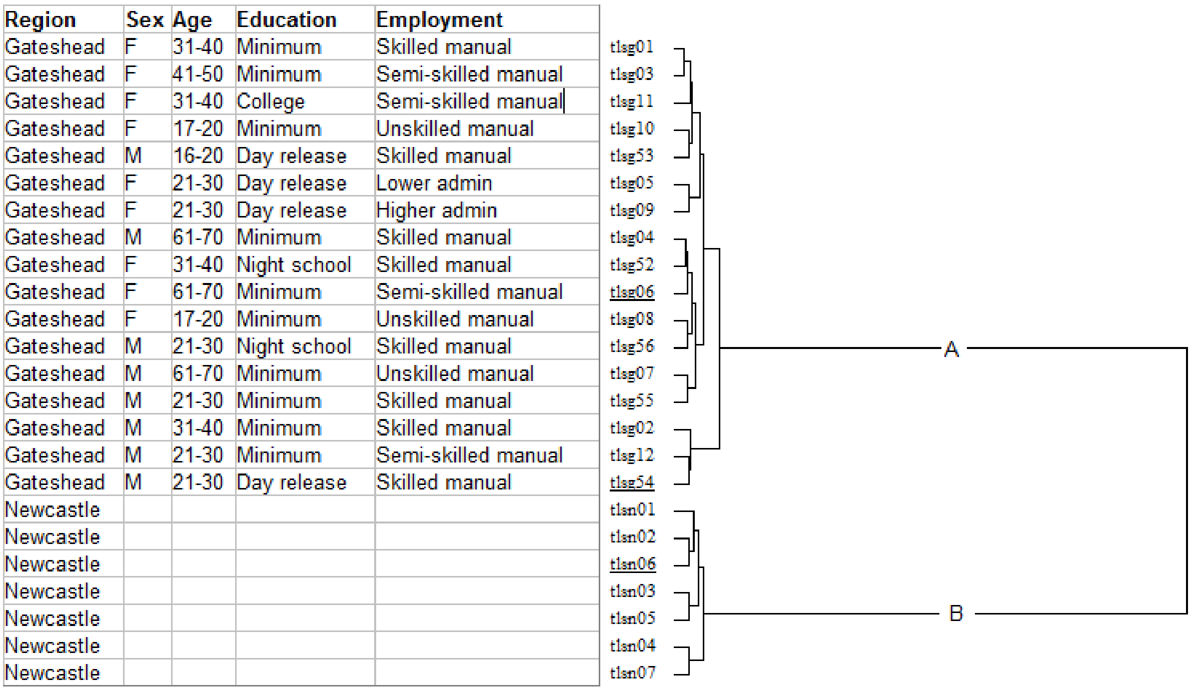


Figure 13. Centroids from Table 7

Alternatively, a sociolinguist might want to extend the hypothesis by determining if the phonetically-based cluster structure correlates systematically with any available social data, which the TLS provides. NECTE incorporates much of the TLS social data, and, as Figure 14 shows, there is indeed a striking correlation: phonetic variation among speakers in the Newcastle and Gateshead areas of Tyneside is relatively small compared to the relatively much larger difference between them.[6]



Region	Sex	Age	Education	Employment
Gateshead	F	31-40	Minimum	Skilled manual
Gateshead	F	41-50	Minimum	Semi-skilled manual
Gateshead	F	31-40	College	Semi-skilled manual
Gateshead	F	17-20	Minimum	Unskilled manual
Gateshead	M	16-20	Day release	Skilled manual
Gateshead	F	21-30	Day release	Lower admin
Gateshead	F	21-30	Day release	Higher admin
Gateshead	M	61-70	Minimum	Skilled manual
Gateshead	F	31-40	Night school	Skilled manual
Gateshead	F	61-70	Minimum	Semi-skilled manual
Gateshead	F	17-20	Minimum	Unskilled manual
Gateshead	M	21-30	Night school	Skilled manual
Gateshead	M	61-70	Minimum	Unskilled manual
Gateshead	M	21-30	Minimum	Skilled manual
Gateshead	M	31-40	Minimum	Skilled manual
Gateshead	M	21-30	Minimum	Semi-skilled manual
Gateshead	M	21-30	Day release	Skilled manual
Newcastle				
Newcastle				
Newcastle				
Newcastle				
Newcastle				
Newcastle				
Newcastle				

Figure 14: Cluster tree correlated with social data.

Analyses and results of the full NECTE phonetic data matrix M are available in Moisl et al. (2006) and Moisl and Maguire (2008). Projected cluster analytic work on DECTE will on the one hand take account of data nonlinearity, and on the other will extend its application to grammatical features. Finally, it has to be noted that cluster analysis in general and hierarchical cluster analysis in particular are not without their problems. Consider, in particular, the following issues:

- There are numerous ways of measuring the degree of similarity between data objects, and of clustering on the basis of such measures. Also, different combinations of similarity measurement and clustering algorithm applied to the same dataset can and will generate different cluster structures. Which if any of these is optimal, and how is optimality judged?
- Clustering can be unstable in the sense that small changes to the data, such as leaving out even as few as a single row or column of the data matrix, can lead to fundamentally different analyses. Does this instability reflect a substantive change in the structure of the data, or is it an artefact of the clustering method?
- Interpretation of any given analysis can be subjective. How many clusters are there, for example, and how well do they relate to what is known about the domain that the data describes?

[6] No social information apart from regionality is available for the Newcastle speakers, so the speaker comparisons cannot be taken further.

These problems are recognized in the cluster analysis community, and a wide range of solutions to them has been proposed; see for example Haldiki et al. (2001), Gan et al. (2007, ch. 17), Xu and Wunsch (2009, ch. 10), Everitt et al. (2011, chapter 9), though they will not be explored further in this paper..

5. Conclusion

Our general argument in the foregoing discussion has been that, in view of the rapid development of digital electronic natural language corpora on the one hand and of mathematically and statistically based computational methods for analysis of such corpora on the other, the time has come to take aggregate analysis as seriously as the traditional feature-centric approach in corpus-based variationist linguistics.

We are not, of course, the only research group to have come to this conclusion. In an extensive series of publications, John Nerbonne, Wilbert Heeringa, Martijn Wieling, Peter Kleiweg and their co-workers have argued convincingly for the application of aggregate analysis in dialectometry using, among other things, a variety of cluster analytic techniques to identify the distribution of significant dialectal features in languages as disparate as Dutch, German, Norwegian, Bulgarian, and Catalan (see, for example, Heeringa and Nerbonne 2012; Nerbonne 2006 and 2008; Nerbonne, Kleiweg, Heeringa and Manni 2008; Nerbonne 2010; Wieling and Nerbonne 2010; Wieling, Shackleton and Nerbonne submitted). As we hope to have shown, the combination of aggregate analysis with feature-centric approaches is important for unlocking the extent to which Tyneside English is variable depending on who the speaker is and where they are situated temporally and spatially.

We would go so far as to say that aggregate methods, in particular, are crucial for unlocking the secrets of variability in languages about which we have no instincts with respect to variation, because they have not received the same scholarly attention. This is especially true of ancient or endangered languages, where variation can be uncovered using techniques such as cluster analysis without the need for access to native speaker intuitions of the kind we are familiar with in both the generative tradition and, as we hope to have shown here, in feature-centric accounts in the variationist literature.

References

Allen, Will, Joan C. Beal, Karen P. Corrigan, and Hermann Moisl 2007 A Linguistic ‘Time Capsule’: The Newcastle Electronic Corpus of Tyneside English. In: Joan C. Beal, Karen P. Corrigan and Hermann Moisl (eds.), *Creating and Digitizing Language Corpora, Vol. 2: Diachronic Databases*, 16–48. Basingstoke: Palgrave Macmillan.

Barnfield, Kate and Isabelle Buchstaller 2010 Intensifiers on Tyneside: Longitudinal developments and new trends. *English World-Wide* 31(3): 252–287.

Beal, Joan C. 2000 From Geordie Ridley to Viz: Popular literature in Tyneside English. *Language and Literature* 9(4): 343–359.

Beal, Joan C., Karen P. Corrigan, Nicholas Smith, and Paul Rayson 2007 Writing the Vernacular: Transcribing and Tagging the Newcastle Electronic Corpus of Tyneside English (NECTE). In: Anneli Meurman-Solin and Arja Nurmi (eds.), *Studies in Variation, Contacts and Change in English*, *Volume 1. Annotating Variation and Change.* Research Unit for

Variation, Contacts and Change in English (VARIENG), University of Helsinki. <http://www.helsinki.fi/varieng/journal/volumes/01/beal_et_al>.

Beal, Joan C. and Karen P. Corrigan 2007 'Time and Tyne': A corpus-based study of variation and change in relativization strategies in Tyneside English. In: Stephan Elspass, Nils Langer, Joachim Scharloth and Wim Vandenbussche (eds.), *Germanic Language Histories 'from Below' (1700–2000), Proceedings*, 99–114. (Studia Linguistica Germanica 86.) Berlin / New York: Walter de Gruyter.

Beal, Joan C. and Karen P. Corrigan 2011 Inferring syntactic variation and change from the *Newcastle Electronic Corpus of Tyneside English* (NECTE) and the *Corpus of Sheffield Usage* (CSU). In: Terttu Nevalainen and Susan Fitzmaurice (eds.), *Studies in Variation, Contacts and Change in English, Volume 7. How to Deal with Data: Problems and Approaches to the Investigation of the English Language over Time and Space* Research Unit for Variation, Contacts and Change in English (VARIENG), University of Helsinki. <http://www.helsinki.fi/varieng/journal/volumes/index.html>.

Boslaugh, Sarah and Paul Watters 2008 *Statistics in a Nutshell*. Beijing:O'Reilly.

Carreira-Perpinan, Miguel 2011 Dimensionality Reduction. London: Chapman & Hall.

Corrigan, Karen P. in press GOAT vowel variants in the Diachronic Electronic Corpus of Tyneside English (DECTE). In: Elizabeth Closs Traugott and Terttu Nevalainen (eds.), *The Oxford Handbook of the History of English.* Oxford: Oxford University Press.

Docherty, Gerry and Paul Foulkes 1999 Sociophonetic variation in 'glottals' in Newcastle English. *Proceedings of the 14th International Congress of Phonetic Sciences,* 1037–1040. Berkeley: University of California.

Everitt, Brian, Sabine Landau, Morven Leese, and Daniel Stahl 2011 *Cluster Analysis*. 5th edition. Chichester: Wiley.

Gan, Guojun, Chaoqun Ma, and Jianhong Wu 2007 *Data Clustering. Theory, Algorithms, and Applications*. Philadelpha: Society for Industrial and Applied Mathematics.

Haldiki, Maria, Yannis Batistakis, and Michalis Vazirgiannis 2001 On clustering validation techniques. *Journal of Intelligent Information Systems* 17: 107–145.

Heeringa, Wilbert 2004 Measuring Dialect Pronunciation Differences using Levenshtein Distance, Ph.D. thesis, University of Groningen.

Heeringa, Wilbert and John Nerbonne 2012 Dialectometry. Accepted to appear in: Frans Hinskens & Johan Taeldeman (eds.), *Language and Space. An International Handbook of Linguistic Variation, Volume III: Dutch.* (Series: Handbook of Linguistics and Communication Science (HSK)). Berlin and New York: Walter de Gruyter.

Jones, Val 1978 Some problems in the computation of sociolinguistic data. Ph.D. thesis, University of Newcastle upon Tyne.

Jones-Sargent, Val 1983 *Tyne Bytes. A computerised sociolinguistic study of Tyneside*. Frankfurt am Main: Peter Lang.

Lee, John A. and Michael Verleysen 2007 *Nonlinear Dimensionality Reduction*. Berlin: Springer.

Manning, Christopher, Prabhakar Raghavan, and Hinrich Schütze 2008 *Introduction to Information Retrieval*. Cambridge: Cambridge University Press.

Milroy, Lesley, James Milroy, Gerry Docherty, Paul Foulkes and David Walshaw 1997 Phonological variation and change in contemporary English: evidence from Newcastle-upon-Tyne and Derby. *Cuadernos de Filologia Inglesa* 8(1): 35–46.

Moisl, Hermann and Val Jones 2005 Cluster analysis of the Newcastle Electronic Corpus of Tyneside English: a comparison of methods. *Literary and Linguistic Computing* 20: 125–146.

Moisl, Hermann, Warren Maguire, and Will Allen 2006 Phonetic variation in Tyneside: Exploratory multivariate analysis of the Newcastle Electronic Corpus of Tyneside English. In: Frans Hinskins (ed.), *Language Variation: European Perspectives. Selected Papers from the Third International Conference on Language Variation in Europe (ICLaVE 3), Amsterdam, June 2005*, 127–141. Amsterdam: John Benjamins.

Moisl, Hermann 2007 Data nonlinearity in exploratory multivariate analysis of language corpora, Computing and Historical Phonology. In: John Nerbonne, Mark T. Ellison, and Grzegorz Kondrak (eds.), *Proceedings of the Ninth Meeting of the ACL Special Interest Group in Computational Morphology and Phonology, Prague, June 2007*, 93–100. Association for Computational Linguistics. <http://aclweb.org/anthology-new/W/W07/W07-1312.pdf>.

Moisl, Hermann and Warren Maguire 2008 Identifying the main determinants of phonetic variation in the Newcastle Electronic Corpus of Tyneside English. *Journal of Quantitative Linguistics* 15(1): 46–69.

Moisl. Hermann 2010a Variable scaling in cluster analysis of linguistic data. *Corpus Linguistics and Linguistic Theory* 6: 75–103.

Moisl, Hermann 2010b Finding the minimum document length for reliable clustering of multi-document natural language corpora. *Journal of Quantitative Linguistics* 18: 23–52.

Moisl, Hermann, Karen P. Corrigan, Isabelle Buchstaller, and Adam Mearns 2011 The phonetics of Tyneside speech: a diachronic study of the 'goat' vowel. Paper presented at Methods in Dialectology XIV, August 2011, University of Western Ontario, London, Ontario.

Nerbonne, John 2006 Identifying linguistic structure in aggregate comparison. *Literary and Linguistic Computing* 21(4): 463–475.

Nerbonne, John, Peter Kleiweg, Wilbert Heeringa, and Franz Manni 2008 Projecting Dialect Differences to Geography: Bootstrap Clustering vs. Noisy Clustering. In: Christine Preisach, Lars Schmidt-Thieme, Hans Burkhardt, and Reinhold Decker (eds.), *Data Analysis, Machine Learning, and Applications. Proc. of the 31st Annual Meeting of the German Classification Society*, 647–654. Berlin: Springer.

Nerbonne, John 2009 Data-Driven Dialectology. *Language and Linguistics Compass* 3(1): 175–198.

Nerbonne, John 2010 Mapping Aggregate Variation. In: Alfred Lameli, Ronald Kehrein, and Stephan Rabanus (eds.), *Language and Space. International Handbook of Linguistic Variation. Vol. 2: Language Mapping*, 476–495. Berlin: Mouton De Gruyter.

O'Keefe, Anne and Michael McCarthy (eds.) 2010 *The Routledge Handbook of Corpus Linguistics*. Abingdon: Routledge.

Pearce, Michael 2009 A perceptual dialect map of North East England. *Journal of English Linguistics* 37(2): 162–192.

Pellowe, John and Val Jones 1978 On intonational variety in Tyneside speech. In: Peter Trudgill (ed.), *Sociolinguistic Patterns in British English*, 101–121. London: Arnold.

Pellowe, John, Graham Nixon, Barbara Strang, and Vince McNeany 1972 A dynamic modelling of linguistic variation: the urban (Tyneside) linguistic survey. *Lingua* 30: 1–30.

Singhal, Amit, Gerard Salton, Mandar Mitra, and Chris Buckley 1996 Document length normalization. *Information Processing and Management* 32: 619–633.

Spärck Jones, Karen, Steve Walker and Stephen E. Robertson 2000 A probabilistic model of information retrieval: development and comparative experiments, part 2. *Information Processing and Management* 36: 809–40. <http://www.soi.city.ac.uk/~ser/blockbuster/pmir-pt2-reprint.pdf>

Szmrecsanyi, Benedikt and Bernd Kortmann 2009 The morphosyntax of varieties of English worldwide: A quantitative perspective. *Lingua* 119(11): 1643–1663.

Strang, Barbara 1968 The Tyneside Linguistic Survey. Paper presented at the International Congress on Dialectology, Marburg 1965. *Zeitschrift für Mundartforschung, Neue Folge* 4: 788–794.

Tagliamonte, Sali A. and Rika Ito 2002 Think really different: Continuity and specialization in the English adverbs. *Journal of Sociolinguistics* 6(2): 236–266.

Upton, Clive 2004 Received Pronunciation. In: Bernd Kortmann, Edgar Schneider, Kate Burridge, Rajend Mesthrie, and Clive Upton (eds.), *A Handbook of Varieties of English*, 217–230. Berlin: Mouton de Gruyter.

Watt, Dominic1999 Phonetic variation in two Tyneside vowels: Levelling in lockstep. In: John J. Ohala, Yoko Hasegawa, Manjari Ohala, Daniel Granville, and Ashlee C. Bailey (eds.), *Proceedings of the 14th International Congress of Phonetic Sciences, San Francisco, August 1999*, 1621–1624. Berkeley: University of California Press.

Watt, Dominic2000 Phonetic parallels between the close-mid vowels of Tyneside English: Are they internally or externally motivated? *Language Variation and Change* 12: 69–101.

Watt, Dominic2002 "I don't speak with a Geordie accent, I speak, like, the Northern accent": Contact-induced dialect levelling in the Tyneside vowel system. *Journal of Sociolinguistics* 6(1): 44–63.

Watt, Dominic and Lesley Milroy 1999 Patterns of variation and change in three Newcastle vowels: Is this 'dialect levelling?'. In: Paul Foulkes and Gerard J. Docherty (eds.), *Urban Voices: Accent Studies in the British Isles*, 25–47. London: Arnold.

Wells, John C. 1982 *Accents of English*, 3 Volumes. Cambridge: Cambridge University Press.

Wieling, Martijn and John Nerbonne 2010 Bipartite spectral graph partitioning for clustering dialect varieties and detecting their linguistic features. *Computer Speech and Language* 25: 700–715.

Wieling, Martijn, Robert Shackleton Jr., and John Nerbonne submitted Analyzing Phonetic Variation in the Traditional English Dialects: Simultaneously Clustering Dialect and Phonetic Features.

Xu, Rui and Donald C. Wunsch, II. 2009 *Clustering*. Hoboken NJ: John Wiley & Sons.

Websites

DECTE: The Diachronic Electronic Corpus of Tyneside English. <http://research.ncl.ac.uk/decte>

ENROLLER: An Enhanced Repository for Language and Literature Researchers. <http://www.gla.ac.uk/enroller>

NECTE: The Newcastle Electronic Corpus of Tyneside English. <http://research.ncl.ac.uk/necte>

TEI: Text Encoding Initiative. <http://www.tei-c.org/index.xml>