Identifying the Main Determinants of Phonetic Variation in the Newcastle Electronic Corpus of Tyneside English

Identifying the Main Determinants of Phonetic Variation in the *Newcastle Electronic Corpus of Tyneside English*

Hermann Moisl
University of Newcastle upon Tyne, UK[1]

Warren Maguire
University of Edinburgh, UK[2]

Abstract

The *Newcastle Electronic Corpus of Tyneside English* is a corpus of dialect speech from North-East England. It includes phonetic transcriptions of 63 interviews together with social data relating to each interviewee, and offers an opportunity to study the sociophonetics of Tyneside speech of the late 1960s. In a previous paper we began that study with an exploratory multivariate analysis of the transcriptions. The results were that speakers fell into clearly defined groups on the basis of their phonetic usage, and that these groups correlated well with social characteristics associated with the speakers. The present paper develops these results by trying to identify the main phonetic determinants of the speaker groups.

Short title: The Main Determinants of Phonetic Variation

Word length: 6441

[1] Telephone: +44 (0)191 222 7781
Fax: +44 (0)191 222 8708
Address: School of English Literature, Language, and Linguistics, Percy Building, University of Newcastle, Newcastle upon Tyne NE1 7RU, UK
Email: hermann.moisl@ncl.ac.uk
Web: http://www.ncl.ac.uk/elll/staff/profile/hermann.moisl

[2] Telephone: +44 (0)131 650 3947
Address: School of Philosophy, Psychology and Language Sciences, The University of Edinburgh, Teviot Place (Doorway 1), Edinburgh EH8 9AG, UK.
Email: w.maguire@ed.ac.uk

The *Newcastle Electronic Corpus of Tyneside English* (NECTE; Corrigan, Moisl & Beal, 2005) is a corpus of dialect speech from Tyneside in north-east England, which includes the cities of Gateshead on the south shore of the river Tyne, and Newcastle on the north shore (Figure 1).

Figure 1

It is based on two pre-existing corpora of audio-recorded speech, one of them gathered in the late 1960s by the *Tyneside Linguistic Survey* (TLS) (Strang,1968; Pellowe, Nixon Strang, & McNeany, 1972; Pellowe & Jones, 1978; Jones-Sargent, 1983), and the other between 1991 and 1994 by the *Phonological Variation and Change in Contemporary Spoken English* (PVC) project (Milroy, Milroy, Docherty, Foulkes, & Walshaw, 1994; Docherty & Foulkes, 1999). The TLS material includes detailed phonetic transcriptions of 63 interviews together with social data relating to each interviewee, and as such offers an opportunity to study the sociophonetics of Tyneside speech of the late 1960s in detail. In a previous paper (Moisl, Maguire, & Allen, 2006), we began that study with an exploratory multivariate analysis of the

transcriptions with the aim of generating hypotheses about phonetic variation among speakers and speaker groups in the corpus, and how such variation correlates with social factors. The results were that speakers fell into clearly defined groups on the basis of their phonetic usage, and that these groups correlated well with social factors associated with the speakers.

The present paper develops these results by trying to identify the main phonetic determinants of the speaker groups. The discussion is in three main parts. The first describes the NECTE phonetic data, the second outlines the results of our earlier study, and the third identifies and discusses the phonetic features that are most important in determining the speaker groups found in that study.

1. The NECTE phonetic data

1.1 *The TLS Phonetic Transcriptions*

One of the main aims of the TLS project was to see whether systematic phonetic variation among Tyneside speakers of the period could be significantly correlated with variation in their social characteristics. To this end

the TLS developed a methodology which was radical at the time and remains so today: in contrast to the then-universal and still-dominant theory driven approach, where social and linguistic factors are selected by the analyst, the TLS proposed a fundamentally empirical approach in which salient factors are extracted from the data itself and then serve as the basis for model construction.

To realize its research aim using strictly empirical methodology, the TLS had to compare the audio interviews it had collected at the phonetic level of representation. This required that the analog speech signal be discretized into phonetic segment sequences, or, in other words, to be phonetically transcribed. Details of the TLS transcription scheme are available in Jones-Sargent (1983) and Corrigan, Moisl & Beal (2005); for present purposes, it is sufficient to note that two levels of transcription were produced, a broad one that the TLS referred to as the PDV ("Putative Diasystemic Variable") level, and a highly detailed narrow one designated STATE. The PDV-level transcription was analyzed in the results reported here.

1.2 *Data Construction*

The analyses discussed below are based on comparison of phonetic profiles associated with each of the TLS speakers. A profile for any speaker S is the number of times S uses each of the phonetic segments defined by the PDV transcription scheme in his or her interview. For computational analysis, the speaker profiles have to be mathematically represented, and this is done using vectors. A vector is a sequence of slots or elements indexed by the positive integers 1, 2...*n*, where *n* can be any desired positive integer, and each element contains some --usually numerical-- value; in Figure 2, for example, the vector consists of four elements, that is, *n* =4, and the value in v_3 = 7.5:

Figure 2

In this representation, a speaker profile P is a vector having as many elements as there are phonetic segments in the PDV scheme such that each vector element P_j represents the *j*th segment, where *j* is in the range 1..number of segments in the PDV scheme, and the value stored at P_j is an

integer representing the number of times S uses the *j*th segment. There are 156 segments in the PDV scheme, and so a profile is a length-156 vector. For example:

Figure 3

There are 63 TLS speakers, and their profiles are represented in a matrix having 63 rows, one for each profile.

Figure 4

At the PDV level, therefore, the data used in this study comprises a 63 x 156 matrix $M_{63,156}$; the subscript serves to distinguish this matrix from others used in what follows.

1.3 *Data Preprocessing*

Prior to analysis, $M_{63,156}$ was transformed in two ways.

1.3.1 Compensation for variation in file length

The number of phonetic segments per speaker interview varies significantly, and this variation in length has to be taken into account when

conducting the analyses in order to avoid misleading results (Moisl, 2007). The segment frequency values in the matrix $M_{63,156}$ were adjusted in accordance with the following function:

Figure 5

where *freq'* is the adjusted frequency, M_{ij} is the value at the (i,j) coordinates of the data matrix $M_{63,156}$, *freq* is the existing frequency value at M_{ij}, μ is the mean number of codes per interview across all 63 interviews, and I is the total number of segments in interview *i*. This function adjusts the frequency profile of each speaker in relation to the mean number of segments per speaker across all interviews. More specifically, it increases the frequency values for relatively shorter interviews in proportion to the mean interview length, and decreases frequency values for relatively longer interviews relative to the mean.

1.3.2. Dimensionality reduction

A general problem in multivariate data analysis is sparsity: the number of data items required to give reliable analytical results increases

exponentially with the dimensionality of the data, that is, with the length of the vector representing each data item, so that, even for moderate vector lengths, getting enough data quickly becomes an insuperable problem. This is widely known as the "curse of dimensionality", and the way to reduce the effect of the curse is to keep the dimensionality as low as possible consistent with the need to describe the domain of inquiry adequately (Verleysen, 2003; Moisl, 2007). In the present instance the number of data items is fixed at 63 speakers and, since there are 156 PDV phonetic segments, the dimensionality of the vector representing each speaker profile is 156. The data is therefore very sparse, and any reduction in the dimensionality of the profile vectors would be beneficial. In fact, our earlier study and the summary which follows show that many and indeed most of the PDV segments are superfluous in the sense that they contribute little or nothing to distinguishing speakers from one another. As such, the dimensionality of the profile vectors can be substantially reduced with minimal loss of information using one or more of the range of available reduction methods (Moisl 2007).

The dimensionality reduction method used here was to eliminate low-variance segments. The variance of a variable x is a measure of how much the values that x takes in a data set deviate from the mean, and therefore how much variability there is in x:

Figure 6

Assuming a set of m values $\{x_1, x_2 ... x_m\}$, the mean μ is $(x_1 + x_2 + ... + x_m) / m$, the amount by which any given value x_i differs from μ is $x_i - \mu$, and the average difference from μ across all values is $\Sigma_{i=1..m} (x_i - \mu) / m$. In relation to our matrix $M_{63,156}$ each of the columns representing a segment is a variable. By calculating the variance of the 63 frequency values in each column, it is possible to identify the segments which are useful in distinguishing speakers from one another, and which are not: for any given segment, low variance indicates that the speakers differ little in that segment and that it is consequently not very useful in distinguishing them, and high variance indicates the obverse, with gradations of usefulness in between. The

variances for the 156 columns of $M_{63,156}$ were calculated, sorted in descending order of magnitude, and plotted:

Figure 7

There are a few high-variance segments, a moderate number of middling-variance segments, and a majority of low-variance ones. The segments to the right of – generously – the 80th have such low variance that they can be eliminated from consideration. They were therefore removed from $M_{63,156}$, resulting in a reduced-dimensionality 63 x 80 matrix $M_{63,80}$.

2. Outline of Hierarchical Cluster Analysis Results

$M_{63,80}$ was analyzed using hierarchical cluster analysis, an exploratory multivariate technique that shows interrelationships among speakers as binary trees familiar from phrase structure trees from natural language sentences (Everitt, Landau, & Leese, 2001; for application-specific details see Jones & Moisl (2005) and Moisl, Maguire, & Allen (2006)). The tree for $M_{63,80}$ is shown in Figure 8:

Figure 8

The lengths of the horizontal lines represent relativities of similarity between pairs of speaker profiles or speaker profile groups – the longer the line, the more dissimilar the profiles. Knowing this, it is clear that there are two main clusters, here labelled NG1 and NG2, that NG1 contains well-defined subclusters NG1a and NG1b, and that NG1a also contains well-defined subclusters NG1a(i) and NG1a(ii). Correlating these clusters with the social data such as gender, age, and socio-economic status available for the TLS speakers, it emerged that those in the NG1 cluster were all from Gateshead on the south side of the river Tyne and largely working-class, and those in NG2 were all middle-class speakers from Newcastle on the north side.

The much larger Gateshead cluster NG1 was then examined to see if its structure also correlated interestingly with social characteristics of speakers. We were primarily interested in vocalic segments and so looked only at vowel segments. The frequency matrix was recalculated using vowel-segment frequency data for the Gateshead speakers only, length-normalized,

and dimensionality reduced as above to a 56 x 40 matrix $M_{56,40}$. This matrix was then cluster analyzed, with the following result:

Figure 9

There are two main clusters, labelled G1 and G2, and G1 itself consists of two main subclusters G1a and G1b. Once again, there was a systematic correlation with the social data available for the speakers. The clearest correlation is between cluster structure and gender: G2 consists entirely of working class males, and G1 mainly though not exclusively of females. In G1 there is a clear split between a cluster consisting mainly of working-class females (G1a), and one consisting of males and females with a higher socioeconomic status (G1b). Finally, there is no obvious correlation between cluster structure and age.

3. The Main Determinants of Phonetic Variation

We know, then, *that* the NECTE speakers fall into clearly-demarcated groups on the basis of variation in their phonetic usage, and that these groups correlate well with social factors. We do not, however, know *why* that is, nor

what regularities in phonetic variation underlie the categorization of speakers. This section addresses that question.

3.1 *The NG1 (Gateshead) / NG2 (Newcastle) Groups*

The procedure for identifying the segments most important in distinguishing the speakers in NG1 from those in NG2 was as follows:

i. The rows of $M_{63,80}$ were rearranged so that the 56 vectors which constitute NG1 occupied rows 1..56 of the matrix, and the 7 vectors of NG2 occupied rows 57-63.

ii. The columns of $M_{63,80}$ were rearranged in order of descending variance, with the highest-variance segment in column 1.

iii. Centroid vectors for the NG1 and NG2 clusters were constructed by taking the means of the vectors in $M_{63,80}$ that constitute NG1 (rows 1..56) and NG2 (rows 57-63) in accordance with the function

Figure 10

where v_j is the jth element of the centroid vector (for j = 1..the number of columns in M), M is the data matrix $M_{63,80}$, and m is the number of row vectors in the cluster in question (56 for NG1, 7 for NG2).

iv. The resulting vectors $NG1_{centroid}$ and $NG2_{centroid}$ were co-plotted to show graphically how, on average, the two speaker groups differ on each of the 80 PDV segments, the aim being to identify those on which the groups differ most and are thereby the main determinants of phonetic variation for NG1 and NG2.

Figure 11

There is too much detail here for convenient interpretation. Attention can be restricted to a smaller number of higher-variance segments to the left of the plot, since these are more significant in terms of variability between NG1 and NG2. How many segments should be looked at? That depends on how detailed an idea of the pattern of variability is required; only the 6 highest-variance ones were selected for consideration here, and re-plotted:

Figure 12

The segments for which the vectors differ most are the most significant for differentiating NG1 and NG2 speakers. These differences are ranked in descending order of magnitude in the following table:

Table 1

- *Rank* sorts the selected 6 segments in descending order of numerical difference between $NG1_{centroid}$ and $NG2_{centroid}$.
- *Numerical difference between* $NG1_{centroid}$ and $NG2_{centroid}$ specifies the actual numerical difference between the two vectors for each of the 6 segments.
- *Variable index on x-axis* identifies the locations of the segments on the plot in Figure 12. The third-most-important segment [ɔ:], for example, is indexed 5 on the *x*-axis.
- *TLS variable code* and *TLS variable symbol* give the TLS code for the segment in question, together with the corresponding TLS phonetic symbol (Jones-Sargent 1983, pp. 295-302).

The most important segments for distinguishing the Newcastle from the Gateshead speakers can be read off from the table. Two varieties of

schwa (0194 and 0198) are characteristic of Newcastle speakers, and Gateshead speakers use them hardly at all. Next in importance is [ɔː], which is again characteristic of Newcastle, though it occurs also for Gateshead. [ə] (0208), [ɪ], and [eɪ] are characteristic of Gateshead, though they also occur to a small degree among Newcastle speakers.

3.2 *The Gateshead G1 / G2 Groups*

The procedure here is the same as for NG1 / NG2, and will not be described again. The vectors $G1_{centroid}$ and $G2_{centroid}$ were calculated and the 6 highest-variance segments plotted:

Figure 13

Table 2

Interpretation of Table 2 is as for NG1 / NG2. Three segments are significantly more important than the others in distinguishing G2 from G1. The G1 group uses [ɔː] much more often than G2, [ɑː] is characteristic of G2 and is hardly ever used by the G1 group, and [eɪ] is more often used by G2 than by G1.

3.3 *The Gateshead G1a / G1b Groups*

The procedure is again the same as for NG1 / NG2. The vectors $G1a_{centroid}$ and $G1b_{centroid}$ were calculated and the 6 highest-variance segments plotted:

Figure 14

Table 3

Again, interpretation of Table 3 is as for NG1 / NG2. The main segments that distinguish G1a from G1b are, in descending order, [aɪ], [eɪ], and once again [ɔː]. The first is characteristic of G1b and the second of G1a; [ɔː] is more mixed, but is more often used by G1a than G1b.

3.4 *Discussion*

Of all of the vocalic segments in Tyneside English, our current analysis of the TLS phonetic data suggests that three sets of vowels are of particular importance in determining the groups in Figures 8 and 9. Although all of these segments have been commented on before, their relative (and cumulative) sociolinguistic importance has hitherto escaped attention. These three sets are:

- Various types of [ə].

- [ɔː] (0118) and [ɑː] (0122), which correspond to RP [əʊ], and are found in words of the GOAT lexical set as defined by Wells (1982, pp. 146-7).
- [aɪ] (0128), [ɑː] (0130), and [eɪ] (0134), which correspond to RP [aɪ], and are found in words belonging to the PRICE lexical set as defined by Wells (1982, pp. 149-50).

Each of these sets of vowels is discussed in turn.

3.4.1 [ə]-type vowels

Figures 12-14 and Tables 1-3 above reveal that a number of schwa-like vowels are of particular importance in distinguishing some of the groups identified in Figures 8 and 9; for these and associated numerical codes see Jones-Sargent (1983, p.299):

- 0194, a "reduced" vowel in words such as *bak**er***.
- 0196, a "non-reduced" vowel in words such as *Chin**a***.
- 0198, a "reduced" vowel in words such as *stand**ard***.
- 0208, a "reduced" vowel in words such as *hous**es***.

0194 and 0198 are more characteristic of group NG2 (the middle-class Newcastle speakers) than of the group NG1 (the Gateshead speakers). 0208,

on the other hand, is more characteristic of NG1 (Gateshead) than NG2 (Newcastle), and, within Gateshead, is more characteristic of group G1b (the middle-class speakers). Lastly, the schwa-type vowel coded 0196 is more characteristic of group G1a (the working-class Gateshead females) than group G1b (the middle-class Gateshead speakers).

The quality of certain unstressed vowels is a well known marker of localized Tyneside English. Wells (1982, p.376) notes of "Geordie", a colloquial term for a resident of Tyneside and for its localized speech, that:

> The weak vowel of *lett*ER is particularly open in Geordie. Often it is very back: I write it as [ɑ] … The vowel is not necessarily as back as this; some speakers use a more or less front [ɛ]. Words of the *comm*A set also have this [ɑ ~ ɛ] … Tyneside has [ə], not the more usual [ɪ], as the weak vowel in words such as *voices*, *ended.*

The lexical sets *lett*ER and *comm*A are defined in Wells (1982, pp. 165-7). For the most part, the patterns revealed by our analysis of the TLS phonetic data accord closely with Wells' comments. The TLS phonetic codes corresponding to Wells' *lett*ER and *comma* lexical sets are 0194, 0196 and 0198. Of these,

only 0196, which is defined by Jones-Sargent (1983, p. 299) as phonetically [ɛ̝ ~ ɐ ~ ɑ], encodes the local pronunciations referred to by Wells, and it is hence no surprise to find that this vowel is preferred by the female working-class speakers over the middle-class speakers in Gateshead. Conversely, 0194 and 0198, which encode [ə] in *lett*ER, are both much more characteristic of the (exclusively middle-class) Newcastle group than the (largely working-class) Gateshead group.

Interpretation of the distribution of the remaining schwa-type segment 0208 is difficult since, according to Jones-Sargent (1983, p. 299), it encodes two different pronunciations [ə] and [ɪ] which, in light of the comments regarding the pronunciation of *voices* and *ended* in Wells (1982), we might expect to have different social distributions. This perhaps accounts for the behaviour of 0208, which is more characteristic of the (largely working-class) Gateshead group (NG1) on the one hand, but is more characteristic of the middle-class Gateshead speakers (G1b) than the working-class Gateshead females (G1a) on the other. It is possible that analysis of this vowel at the

STATE level referred to in Section 1.1 above might explain the apparently contradictory distribution of 0208.

It is noteworthy that Wells (1982, p. 376) refers to research by McNeany (1971) in his discussion of the pronunciation of unstressed vowels in Tyneside English. Since it was, in fact, McNeany who phonetically transcribed and encoded the TLS data and then used that data in his study, Wells' statement on the pronunciation of unstressed vowels in Tyneside English is based on exactly the same data as is analyzed in the current paper. As such, it is not surprising that a similar picture emerges in both. There is, however, a further consequence of McNeany's interest in unstressed vowels in Tyneside English which potentially has considerable impact on our interpretation of the data analyzed in this paper. Of all of the vocalic variation which occurs in the TLS data, only variation in the pronunciation of unstressed vowels was examined in detail by the original TLS team, as summarized in McNeany (1971). It might on the one hand be that, in analyzing the TLS phonetic data, McNeany was struck by the considerable variation which undoubtedly exists in unstressed vowels in Tyneside English and recognized

its central importance. Or, on the other, it might be that McNeany was interested in unstressed vowels in Tyneside English and hence paid particular attention to them in his transcription of the TLS recordings, with the result that variation in their pronunciation was overstated in relation to that in other vowels. It is consequently possible that, rather than unstressed vowels being of central importance in defining social groups in the TLS because they vary so much more than other vowels, they are important because they were analyzed in more detail by the TLS researchers. Without independent confirmation of the importance of variation in unstressed vowels in Tyneside English, we cannot be certain whether we are dealing with a real phenomenon or an artifact of the (necessary) human discretization of the acoustic signal referred to in Section 1.1 above.

For the other two vocalic segments to be discussed in this paper (the GOAT and PRICE vowels) we are fortunate in having independent studies to compare to the results of our analysis.

3.4.2 The GOAT vowels

Figures 12-14 and Tables 1-3 reveal that two variants of the GOAT lexical set, [ɑː] and (particularly) [ɔː], are of central importance in distinguishing the groups in Figures 8 and 9: [ɑː] is favoured by group G2 (the Gateshead working-class males) over group G1 (the Gateshead speakers other than working-class males), whilst [ɔː] is more characteristic

- of the Newcastle speakers (group NG2) than the Gateshead speakers (group NG1),
- of the Gateshead group G1 (speakers other than working-class males) than the Gateshead group G2 (the working-class males), and
- of the Gateshead group G1a (the working-class females) than the Gateshead group G1b (the middle-class speakers).

Variation in the GOAT vowel is a well known feature of Tyneside English. Watt & Allen (2003, 269) summarize the pronunciation of the GOAT lexical set as follows:

> It is perhaps misleading to state that the vowel of *boat* is [oː] in this accent, when in fact this is only the most frequent of several possible pronunciations of the vowel, some of which are markedly

divergent from this quality and which would perhaps stand as better exemplars of the vowel than this variety than [oː] does because they are more localised. The stereotyped T[ynside]E[nglish] pronunciations [ʊə] and [əː] are examples of this, as are the archaic [aː] and [aʊ], which among older speakers occur sporadically in words like *snow* [snaː] and *soldiers* [ˈsaʊldʒɐz]. Other pronunciations, such as [niː] *no* and [stɪən] *stone*, serve to cloud the picture further.

Furthermore, research by Watt & Milroy (1999) reveals four chief variants of the GOAT vowel in the PVC corpus mentioned at the outset, [oː], [əː], [ʊə] and [oʊ], each of which is associated with particular social profiles. Watt & Milroy (1999, p. 36) describe the variant [oː] as “the unmarked variant”, preferred by all groups except the working-class males, and it is clear from examination of Jones-Sargent (1983, p. 296) that this variant corresponds to the TLS segment [ɔː] (0118). The distribution of this vowel in the TLS data is further discussed below.

Watt & Milroy (1999) find that the variants [ɵː] and [ʊə] are almost exclusively the preserve of males, particularly from the working-class group. These “old fashioned” variants (Watt & Milroy, 1999, p. 37), correspond, despite the symbology, to the TLS segment [uː] (0120) which, although it does not appear in Figures 12-14 and Tables 1-3, is almost completely restricted to the working-class Gateshead male group (G2).

Lastly, Watt & Milroy (1999) find that the [oʊ] variant is almost completely restricted to the speech of middle-class females, old and young, and of young middle-class males. They describe it as characteristic of “high prestige supra-local speech patterns” (pp. 37-38). It is clear that Watt & Milroy’s [oʊ] variant is equivalent to the TLS segment [əʊ] (0116) which, although it does not appear in Figures 12-14 and Tables 1-3 above, is most characteristic of the middle-class groups NG2 (Newcastle) and G1b (Gateshead).

Examination of the TLS transcription and coding scheme (Jones-Sargent, 1983, p. 296) reveals that [ɑː], found in words such as *cold*, *know* and *old*, is equivalent to Watt & Allen’s [aː] variant. Watt & Allen’s description

of [aː] as “archaic” and characteristic of “older speakers” fits with the distribution of [ɑː] described above -- that it is most typical of the speech of working-class males. Interpretation of the distribution of [ɔː] (equivalent to Watt & Allen’s [oː]) is more complicated, however, since it is more characteristic of the (exclusively middle-class) Newcastle group NG2 than of the (mostly working-class) Gateshead group NG1, but within Gateshead is more characteristic of the working-class female group G1a than the middle-class group G1b. This apparently contradictory behaviour of [ɔː] can, however, be explained by taking into account other variants of GOAT in the TLS data. Firstly, the relatively high frequency of [ɑː] in GOAT in the working-class Gateshead male group (G2) means that the proportion of [ɔː] in GOAT in Gateshead is reduced, such that [ɔː] is more characteristic of the Newcastle group (NG2) than the Gateshead group (NG1) overall. That is, the frequency of [ɑː] (and perhaps other GOAT variants such as [uː] and [əʊ]) in Gateshead is higher than the frequency of GOAT variants other than [ɔː] (chiefly [əʊ]) in Newcastle.

Within Gateshead itself, the high frequency of [ɔː] in group G1a (the working-class females) is not surprising, given that other GOAT variants are either characteristic of working-class males ([ɑː] and [uː]) or of middle-class speakers ([əʊ]), and this is consistent with Watt & Milroy's definition of [oː] as "the unmarked variant" (1999, p. 36). That [ɔː] is more characteristic of the working-class females in this case than the middle-class speakers follows from this: the frequency of [ɔː] in the working-class female group is not diminished by "competition" from other variants whereas an alternative GOAT variant, [əʊ], is possible for the middle-class Gateshead speakers and consequently reduces the frequency of [ɔː] for that group.

Despite these complexities, then, the variants of GOAT in the TLS are distributed in a very similar manner to the variants of GOAT revealed in other, independent studies of Tyneside English. In addition to revealing these patterns of distribution, our analysis of the TLS phonetic data suggests that variation in the GOAT vowel is of central sociolinguistic importance in Tyneside English.

3.4.3 The PRICE vowels

Figures 12-14 and Tables 1-3 reveal that three variants of PRICE, [aɪ], [ɑː] and [eɪ], are particularly important for distinguishing the groups in Figures 8 and 9. Of these, [aɪ] is favoured by group G1 (the Gateshead speakers other than working-class males) over group G2 (the Gateshead working-class males), and by group G1b (the middle-class Gateshead group) over G1a (the working-class Gateshead females). That is, [aɪ] is most characteristic of middle-class and female speakers, and least characteristic of male and working-class speakers. [ɑː] is more characteristic of group G1a (the working-class Gateshead females) than of group G1b (the middle-class Gateshead speakers), a pattern which is shared by [eɪ]. In addition, [eɪ] is favoured by the Gateshead speakers (group NG1) over the Newcastle speakers (group NG2), and by the working-class male Gateshead group (G2) over the other Gateshead speakers (G1). That is, unlike [aɪ], [ɑː] and [eɪ] (in particular) are most characteristic of working-class and male speakers and least characteristic of middle-class speakers.

An explanation of the distribution of [ɑː] is relatively straightforward: since this variant is primarily associated with the first person singular pronoun

I, it appears to be the TLS symbol used for the traditional northern English pronunciation of this pronoun recorded in, for example, the Survey of English Dialects (Orton & Halliday, 1962; see the responses to questions IX.7.1, IX.7.4, IX.7.7 and IX.7.9). As such, it is not surprising that this traditional dialect feature is most characteristic of working-class speakers in the TLS.

Other variants of PRICE in Tyneside English, corresponding to TLS [eɪ] and [aɪ], are examined by Milroy (1995) in the PVC corpus . Although Milroy finds that the quality of the diphthong in PRICE words is partly determined by phonological environment, his research shows that, despite this, [ei] in PRICE is most characteristic of working-class males and least characteristic of middle-class females in the PVC. It is clear that this vowel corresponds to the TLS segment [eɪ] (0134), which is most characteristic of the G2 (working-class Gateshead male) and G1a (working-class Gateshead female) groups. Milroy (1985) also finds that two other variants of PRICE, [ai] and [ʌi], are most characteristic of middle-class and female speakers. Although there is no segment corresponding to [ʌi] in the TLS phonetic analysis, it is clear that Milroy's [ai] corresponds to the TLS segment [aɪ] (0128), which is most

characteristic of the G1 (Gateshead other than working-class male) and G1b (Gateshead middle-class) groups. That is, a similar pattern of distribution for variants of the PRICE vowel is revealed in two independent examinations of Tyneside English from different periods and using different methods. As with the variants of the GOAT vowel, our analysis of the TLS phonetic data not only confirms that there is significant variation in PRICE in Tyneside English, but also that it is of central sociolinguistic importance in the dialect.

4. Conclusion

In a previous study (Moisl, Maguire, & Allen, 2006) we found that the 63 speakers included in the Tyneside Linguistic Survey component of the Newcastle Electronic Corpus of Tyneside English could be partitioned into clearly defined groups on the basis of their phonetic usage, and that these groups correlated well with selected social characteristics of the speakers. The aim of the present study was to extend those results by trying to identify the main phonetic determinants of the groups. The discussion was in three main parts. The first part described the NECTE phonetic data, the second outlined the results of our earlier study, and the third identified and discussed

the main phonetic determinants of the speaker groups. By comparing the centroids of the speaker clusters generated by hierarchical cluster analysis in the earlier study, it was possible clearly to identify phonetic segments that distinguish middle-class Newcastle from mainly working-class Gateshead speakers, working-class male from working-class female Gateshead speakers, and Gateshead working-class female from middle-class Gateshead speakers, male and female.

It is hoped that these results will be of interest to sociolinguists and dialectologists in general, and to those concerned with Tyneside dialect in particular. More generally, we feel that this study and our previous one serve to demonstrate the usefulness of quantitative exploratory multivariate analysis in corpus-based linguistics, and thereby encourage its use in the relevant disciplines. Finally, two disclaimers:

- Exploratory multivariate analysis generates hypotheses, not statistically significant results, and such hypotheses have to be tested in the usual way. Our cluster analysis describes the phonetic similarity structure of a corpus, and the conclusions we drew from that

description constitute hypotheses about Tyneside speech; their validity for Tyneside speech in general can only be tested by evidence additional to that used here.

- The cluster trees on which the foregoing discussion is based were generated by a particular selection of vector proximity measure and hierarchical clustering algorithm: squared Euclidean distance and Ward's method respectively. It is a commonplace in the cluster analysis literature that different combinations of proximity measure and clustering algorithm can and do generate different results when applied to the same data. In addition, results from hierarchical and the wide variety of available nonhierarchical clustering methods do not always agree. This is partly because the various hierarchical and nonhierarchical methods make different explicit or implicit assumptions about what constitutes a cluster and how clusters so defined can be algorithmically identified, and partly because they depend to greater or lesser degrees on parameter values that are user-specified, often on a heuristic basis. It is not obvious which method or combination of

parameter values or both is to be preferred in any specific application, or why. This leads to an obvious question: what are these clustering methods really telling us about the structure of the data they describe -- how reliable, in other words, are they, and are they in fact of any use at all if they cannot be relied on to reveal the true structure of the data?

In the literature there are two main approaches to an answer. One is to attempt to establish the validity of cluster results using numerical measures (for example, Everitt, Landau, & Leese, 2001, chap. 8; Duda, Hart, & Stork, 2001, pp. 557-9). The other is to apply a variety of different clustering methods to the same data and to compare the results: a clear convergence on one particular cluster structure is held to support the validity of that structure with respect to the data. And, of course, the two approaches can be used in combination.

Applying these comments to the results of this study, our next step is to evaluate the validity of the cluster tree on which they are based.

5. References

Corrigan, K., Moisl, H., & Beal, J. (2005). *The Newcastle Electronic Corpus of Tyneside English*. Retrieved October 31, 2006, from http://www.ncl.ac.uk/necte/.

Docherty, G., & Foulkes, P. (1999). Sociophonetic variation in 'glottals' in Newcastle English. *Proceedings of the 14th International Congress of Phonetic Sciences*, 1037-1040, Berkeley: University of California.

Docherty, G., Foulkes, P., Milroy, J., Milroy, L., & Walshaw, D. (1997). Descriptive adequacy in phonology: a variationist perspective. *Journal of Linguistics*, 33, 275-310.

Duda, R., Hart, P., & Stork, D. (2001). *Pattern Classification* (2nd ed.). New York: Wiley Interscience.

Everitt, B., Landau, S., & Leese, M. (2001). *Cluster Analysis* (4th ed.). London: Arnold.

Jones, V., & Moisl, H. (2005). Cluster analysis of the Newcastle Electronic Corpus of Tyneside English: a comparison of methods. *Literary and Linguistic Computing*, 20, 1-22.

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

McNeany, V. (1971). Vowel-reduction in localized Tyneside and R.P. speech. Unpublished manuscript, Catherine Cookson Archive, University of Newcastle Upon Tyne.

Milroy, J. (1995). Investigating the Scottish Vowel Length Rule in a Northumbrian dialect. *Newcastle and Durham Working Papers in Linguistics*, 3, 187-196.

Milroy, J., Milroy, L., & Hartley, S. (1994). Local and supra-local change in British English: the case of glottalisation. *English World-Wide*, 15, 1-33.

Milroy, L., Milroy, J., Docherty, G., Foulkes, P. & Walshaw, D. (1997). Phonological variation and change in contemporary English: evidence from Newcastle-upon-Tyne and Derby. In J. Conde Silvestre & J. Hernandez-Campoy (Eds.). *Variation and Linguistic Change in English* (pp. 35-46). Cuadernos de Filologia Ingelsa.

Moisl, H. (2007). Exploratory multivariate analysis. In A. Lüdeling & M. Kytö (Eds.) *Corpus Linguistics. An International Handbook*. Berlin: Mouton de Gruyter.

Moisl, H., Maguire, W., & Allen, W. (2006). Phonetic variation in Tyneside: exploratory multivariate analysis of the Newcastle Electronic Corpus of Tyneside English. In F. Hinskens (Ed.) *Language Variation. European Perspectives*. Amsterdam: Meertens Institute.

Orton, H., & Halliday, W. (1962). *Survey of English Dialects (B). The Basic Material, Vol. 1: The Six Northern Counties and the Isle of Man*. London: Arnold.

Pellowe, J., Nixon, G., Strang, B., & McNeany, V (1972). A dynamic modelling of linguistic variation: the urban (Tyneside) Linguistic Survey. *Lingua*, 30, 1-30.

Pellowe, J., & Jones, V. (1978). On intonational variety in Tyneside speech. In P. Trudgill, (Ed.) *Sociolinguistic Patterns of British English* (pp. 101-121). London: Arnold.

Strang, B. (1968). The Tyneside Linguistic Survey. *Zeitschrift für Mundartforschung*, Neue Folge, 4 , 788-794.

Verleysen, M. (2003). Learning high-dimensional data. In S. Ablameyko, L. Goras, M. Gori, & V. Piuri, (Eds.) *Limitations and future trends in neural computation* (pp. 141-162) Amsterdam: IOS Press.

Watt, D., & Allen, W. (2003). Illustrations of the IPA: Tyneside English. *Journal of the International Phonetic Association*, 33, 267-271.

Watt, D., & Milroy, L. (1999). Patterns of variation and change in three Newcastle vowels: is this dialect levelling?. In P. Foulkes, & G. Docherty (Eds.) *Urban voices: accent studies in the British Isles* (pp. 25-47). London: Arnold.

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

Figure 1: Tyneside in North-East England

V =	1.6	2.4	7.5	0.6
	1	2	3	4

Figure 2: A vector

	iː	ɪ	ɛ	…	ŋ
Speaker	54	25	12	…	21

Figure 3: Vector representation of a speaker profile

Phonetic segments

Speakers		iː	ɪ	ɛ	…	ŋ
	Speaker 1	54	25	12	…	21
	Speaker 2	36	34	23	…	7
	Speaker 3	67	18	24	…	13
	…					
	Speaker 63	64	32	23	…	2

Figure 4: Matrix representation of the TLS speaker profiles

$$freq'(M_{ij}) = freq(M_{ij}) \times \left(\frac{\mu}{l} \right)$$

Figure 5: Interview length normalization function

$$v = \sum_{i=1..m} \frac{(x_i - \mu)^2}{m}$$

Figure 6: The variance function

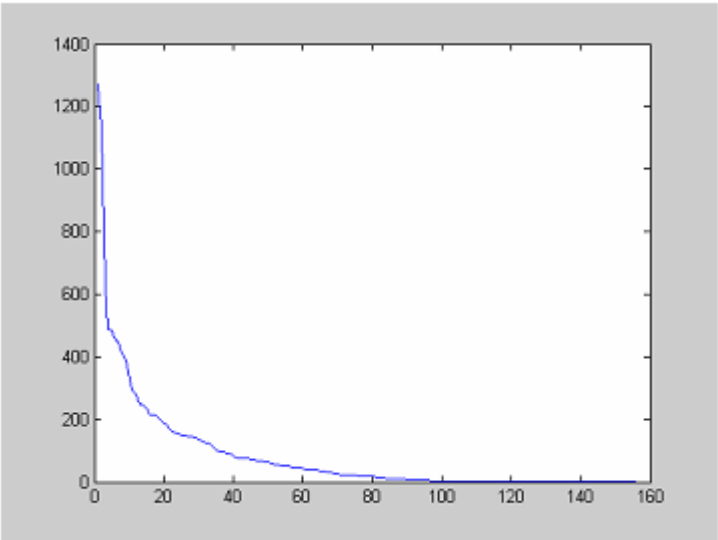

Figure 7: Variance plot of the 156 PDV segments in $M_{63,156}$

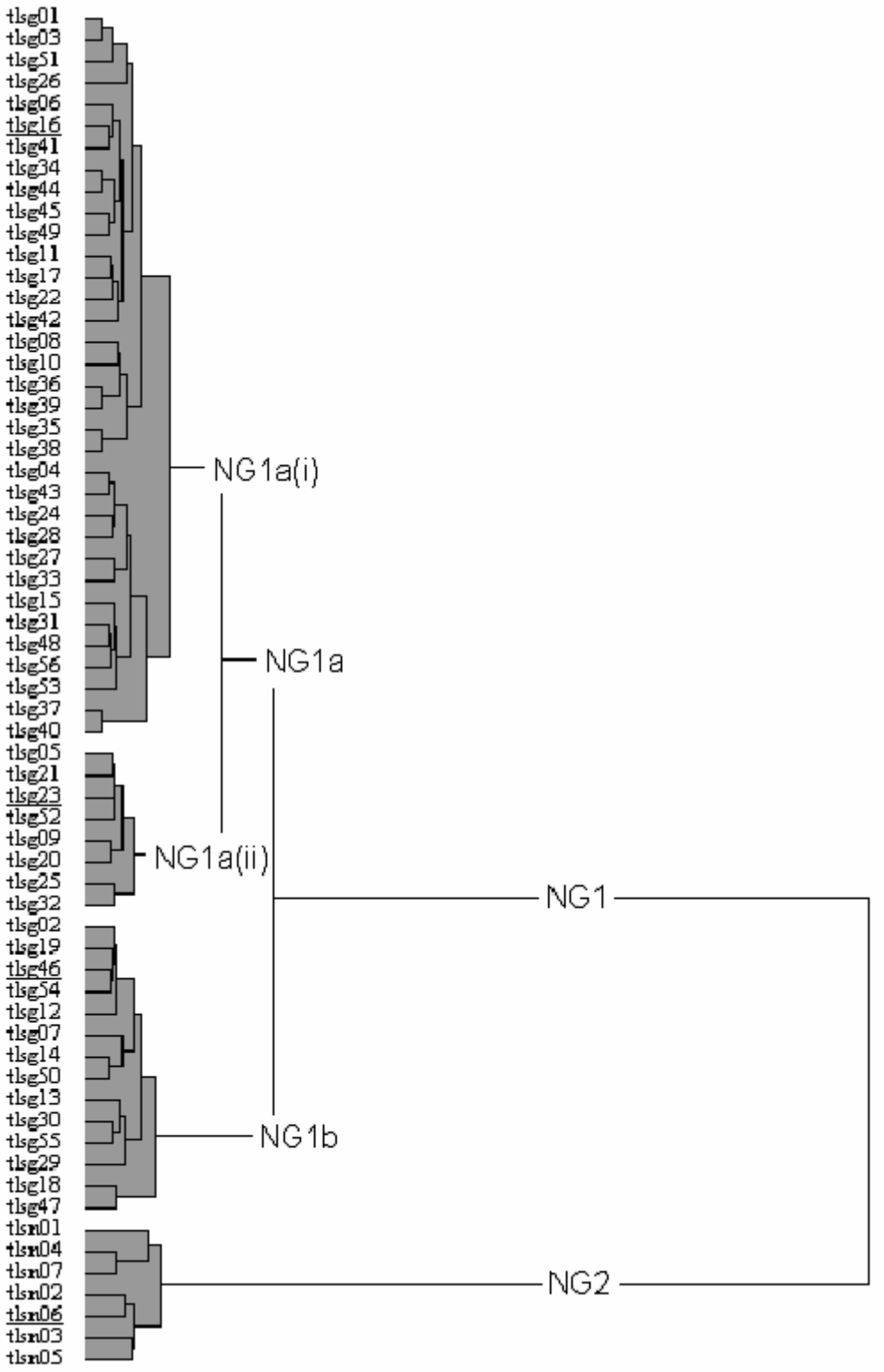

Figure 8: Hierarchical cluster analysis of $M_{63,80}$

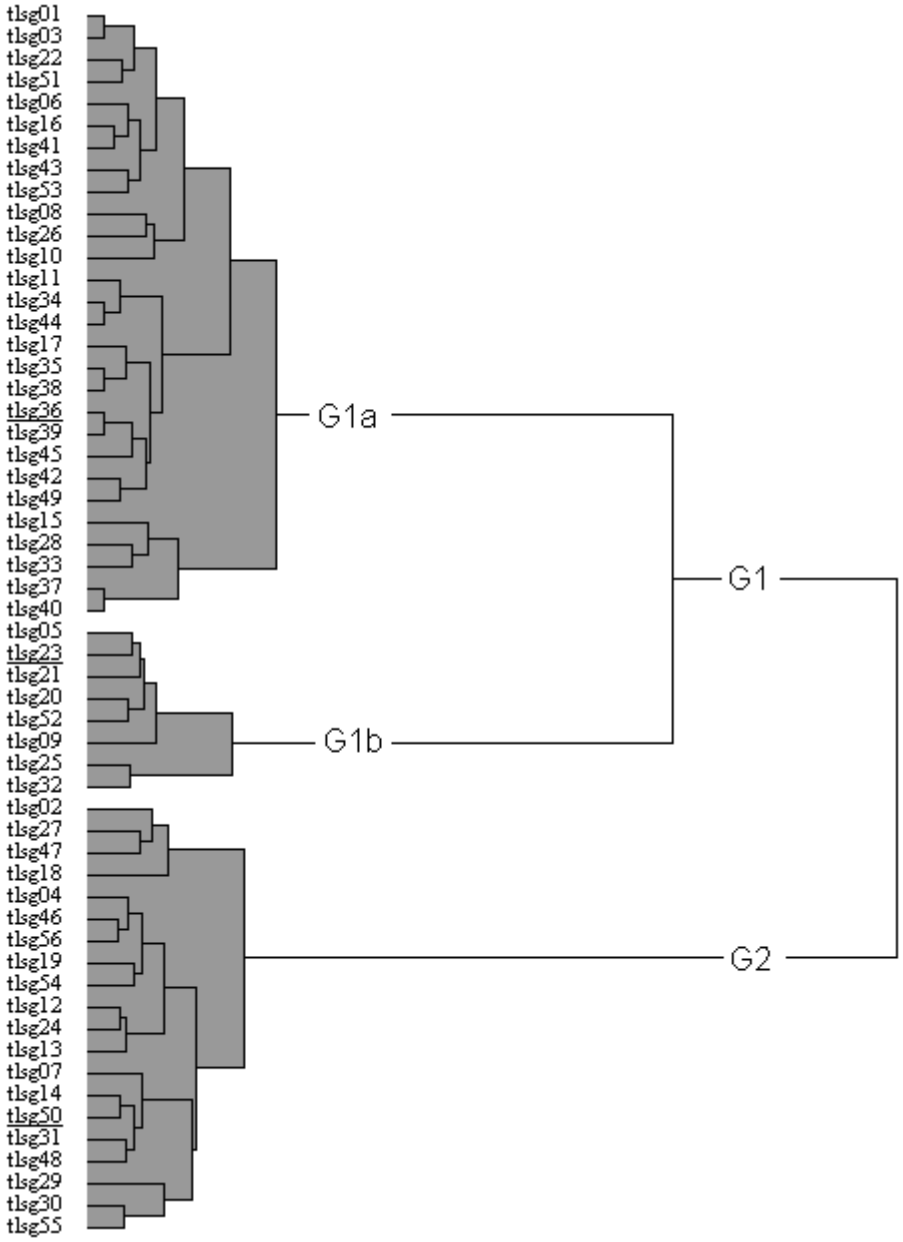

Figure 9: Hierarchical cluster analysis of $M_{56,40}$

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

Figure 10: The centroid function

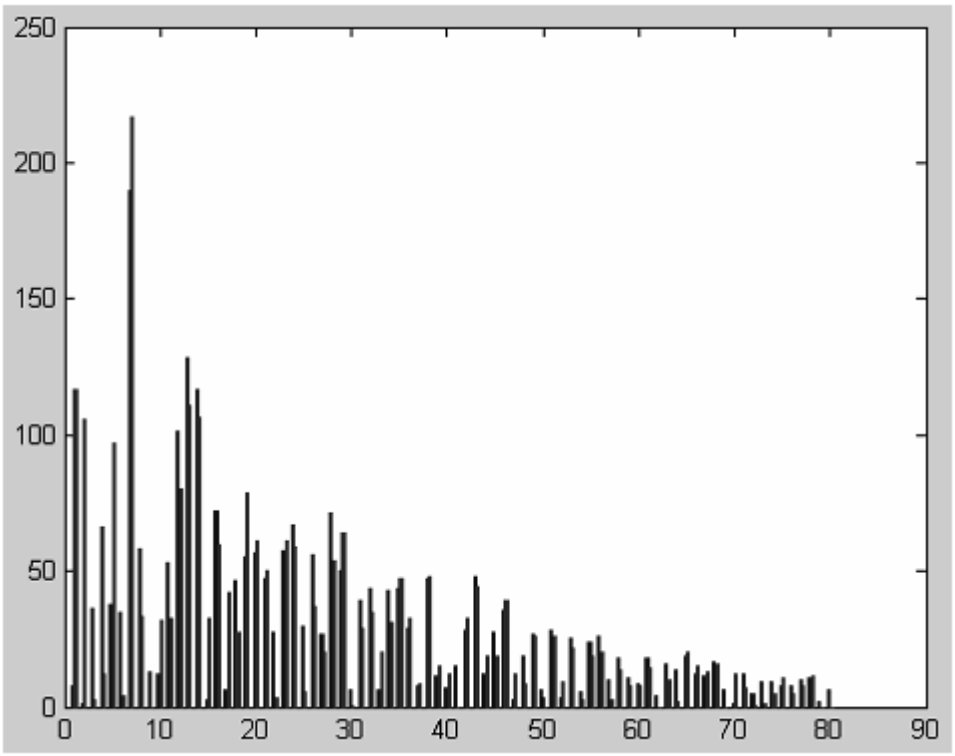

Figure 11: Co-plot of vectors $NG1_{centroid}$ and $NG2_{centroid}$

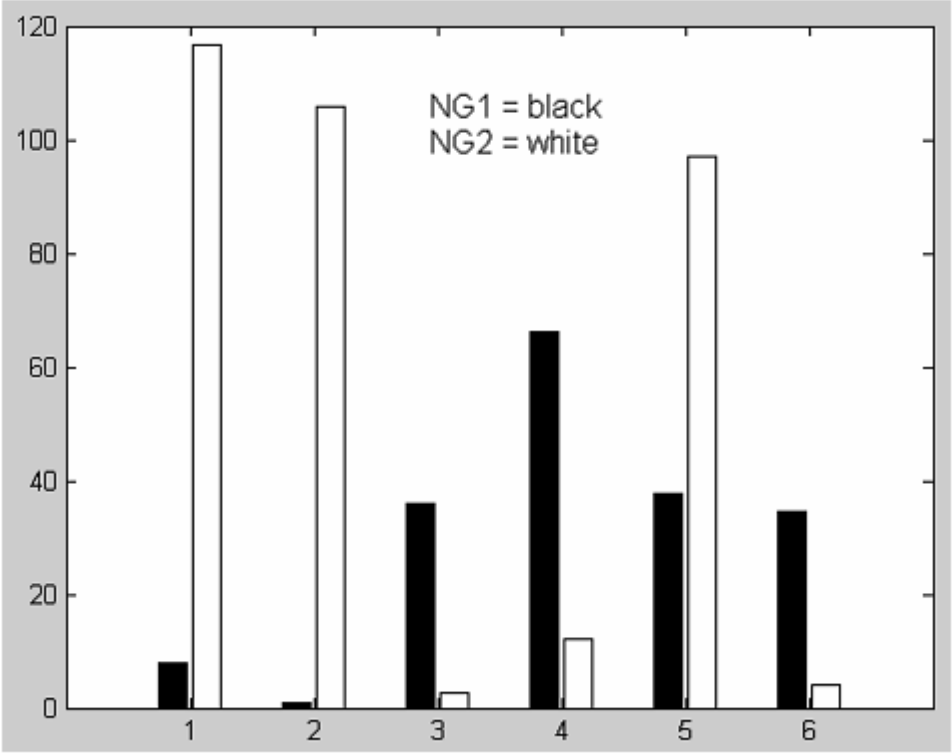

Figure 12: Co-plot of first six elements of vectors $NG1_{centroid}$ and $NG2_{centroid}$

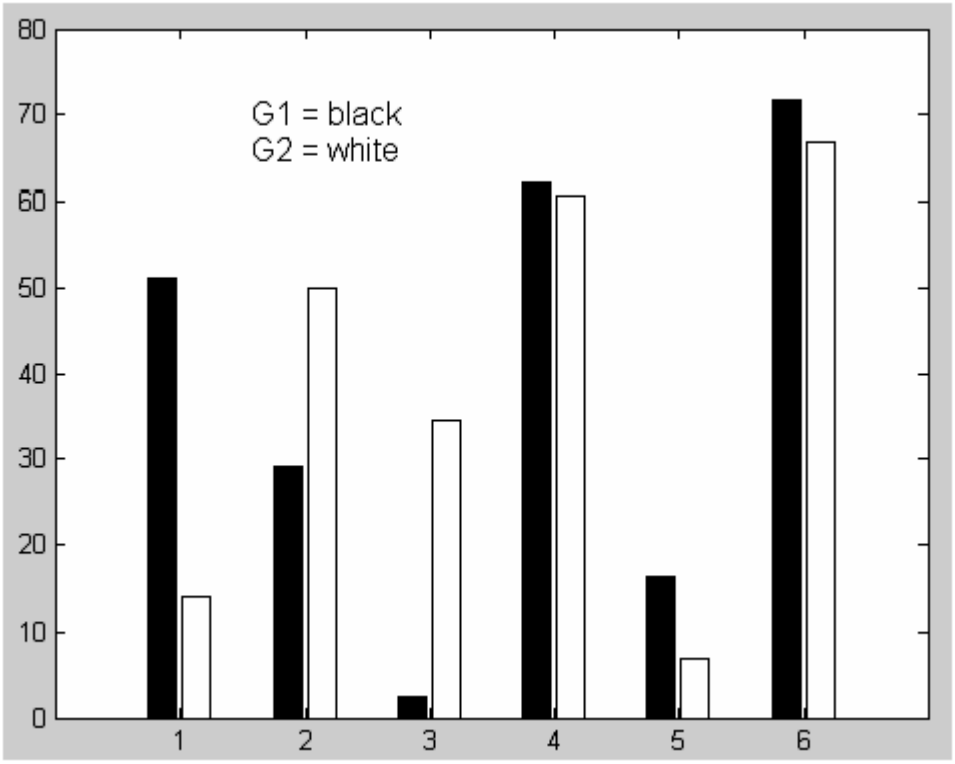

Figure 13: Co-plot of first six elements of vectors $G1_{centroid}$ and $G2_{centroid}$

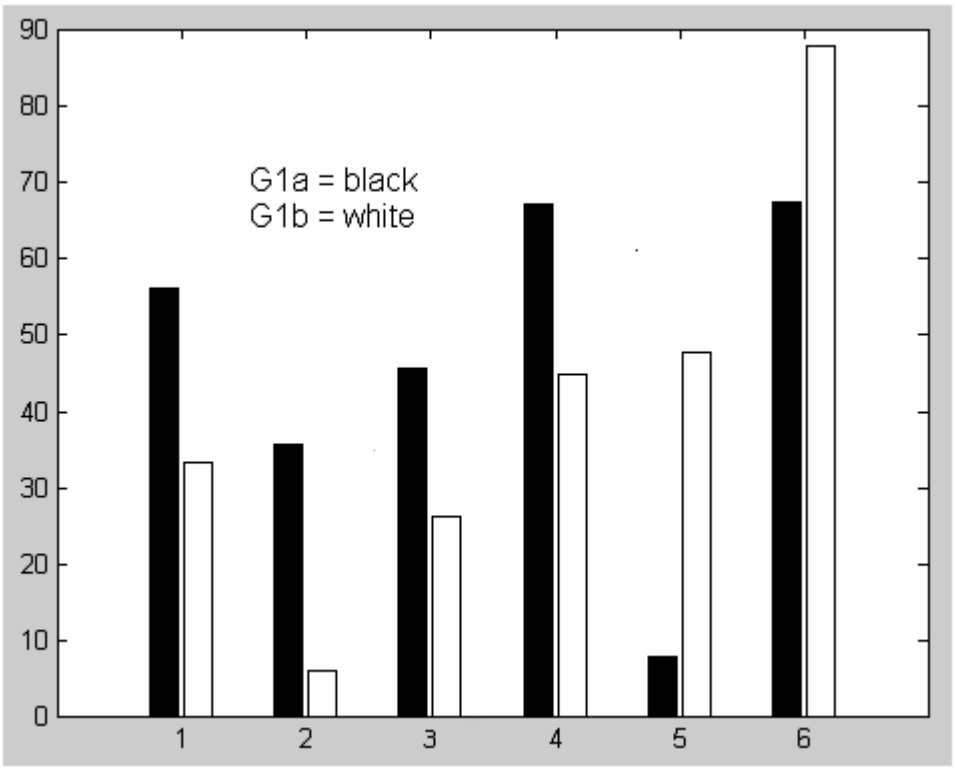

Figure 14: Co-plot of first six elements of vectors $G1a_{centroid}$ and $G1b_{centroid}$

Rank	Numerical difference between $NG1_{centroid}$ and $NG2_{centroid}$	Variable index on x-axis	TLS variable code	TLS variable symbol	Example
1	108.7	1	0194	ə (reduced)	bak**er**
2	105.1	2	0198	ə (reduced)	stand**ar**d
3	59.2	5	0118	ɔː	sm**o**ke
4	54.0	4	0208	ə (reduced)	hous**es**
5	33.2	3	0014	ɪ	b**i**g
6	30.7	6	0134	eɪ	kn**i**fe

Table 1: Key for interpretation for Figure 12

Rank	Numerical difference between $G1_{centroid}$ and $G2_{centroid}$	Variable index on x-axis	TLS variable code	TLS variable symbol	Example
1	36.8	1	0118	ɔː	smoke
2	31.8	3	0122	ɑː	know
3	19.9	2	0134	eɪ	knife
4	9.7	5	0128	aɪ	I
5	8.6	6	0074	ʊ	cup
6	7.7	4	0024	ɛ	well

Table 2 : Key for interpretation of Figure 13

Rank	Numerical difference between $G1a_{centroid}$ and $G1b_{centroid}$	Variable index on x-axis	Variable code	Variable symbol	Example
1	39.6	5	0128	aɪ	l
2	29.8	2	0134	eɪ	knife
3	22.7	1	0118	ɔː	smoke
4	22.4	4	0130	ɑː	l
5	20.7	6	0208	ə (reduced)	houses
6	19.1	3	0196	ə (non-reduced)	Sandra

Table 3: Key for interpretation of Figure 14