Cluster Analysis of the Newcastle Electronic Corpus of Tyneside English: A Comparison of Methods

Hermann Moisl
University of Newcastle, UK

Val Jones
University of Twente, The Netherlands

Abstract

This article examines the feasibility of an empirical approach to sociolinguistic analysis of the Newcastle Electronic Corpus of Tyneside English using exploratory multivariate methods. It addresses a known problem with one class of such methods, hierarchical cluster analysis—that different clustering algorithms can yield different analyses of the same data set, and that there is no obvious way of selecting the best one. The proposed solution is to analyze the data using hierarchical methods in conjunction with one or more fundamentally different types of clustering method, and then to select the analysis on which the hierarchical and the other method(s) agree most closely. A dimensionality reduction method, the self-organizing map (SOM), is used to exemplify this approach. The result is a close though not perfect match between the SOM and complete-link hierarchical analyses, but there is an important reservation—the SOM results vary with changes in user-defined training parameters, and are consequently also open to the criticism of inconsistency. The SOM cannot therefore be an objective arbiter for hierarchical clustering, but the analysis on which they agree gives a better basis for understanding the structure of the data than either method can provide on its own.

Correspondence:
Hermann Moisl,
School of English Literature,
Language, and Linguistics,
University of Newcastle,
Newcastle upon Tyne NE1 7RU, UK
E-mail:
hermann.moisl@ncl.ac.uk

1 Introduction

The Newcastle Electronic Corpus of Tyneside English (NECTE) (http://www.ncl.ac.uk/necte/) project is based on two separate corpora of recorded speech from the North-East of England, one of them collected in the late 1960s as part of the Tyneside Linguistic Survey (TLS) (Strang, 1968; Pellowe *et al.*, 1972), and the other in 1994 by the Phonological Variation and Change in Contemporary Spoken English

Literary and Linguistic Computing, Vol. 20, Suppl Issue, 2005.

doi:10.1093/llc/fqi026 Advance Access Published on 17 June 2005

(PVC) project (Milroy *et al.* 1997). It combines the TLS and the PVC collections into a single corpus and makes it available to the research community in a variety of formats—digitized sound, phonetic transcription, standard orthographic transcription, and part-of-speech tagged, all aligned and accessible on the Web.

We are currently developing a methodology for sociolinguistic analysis of the NECTE corpus, and have begun by looking at the one formulated by the TLS. This was radical at the time and remains so today—in contrast to the then-universal and even now dominant theory-driven approach, where social and linguistic factors are selected by the analyst on the basis of a predefined model, 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 this end, an electronic corpus was created from a subset of the data, and various cluster analyses were applied to it in order to derive social and linguistic classifications of the sample. These classifications were then examined with a view to deriving and relating to one another the most important linguistic and social determinants of linguistic variation in the Tyneside area (Jones, 1978; Jones-Sargent, 1983).

Preliminary results were promising, but their interpretation was necessarily limited by known theoretical and implementation factors—on the one hand by problems relating to interpretation of the cluster analyses, and on the other by limitations imposed by the computational technology available in the late 1970s. The technological limitations have resolved themselves—analyses that were difficult for the TLS investigators can now easily be done on a standard PC. However, the challenge of interpreting the cluster analysis results remains. This article addresses that challenge.

The remainder of the discussion is in three main parts. The first part describes the cluster analysis problem that the TLS faced, the second proposes an approach for resolving it, and the third tests the proposed solution by applying it to the NECTE data. Results indicate that the cluster analysis problem can be mitigated, and the conclusion is that, to the extent to which such mitigation is possible, the problem becomes less of an obstacle to implementation of the TLS's empirical approach to sociolinguistic analysis. Given the increasing availability of large electronic corpora, this conclusion has implications for the conduct of sociolinguistic and dialectological research in general.

2 The TLS cluster analysis problem

In this section we describe the TLS data and some of the cluster analyses originally performed on it together with associated problems. The analyses were performed at the University of Newcastle upon Tyne in the late 1970s by Val Jones, a co-author of this article, and the results were reported in Jones (1978) and Jones-Sargent (1983).

2.1 TLS data

The TLS began in the late 1960s with a series of taped interviews, each about 40 minutes long, with 200 informants sampled from the Tyneside conurbation in North-East England. The entire speech output from each interview was analyzed and transcribed at a number of levels of representation, including standard English orthography, segmental phonology, and syntagmatic, paralinguistic, prosodic, and grammatical features. In addition, a fairly extensive data set of coded social data was established for each informant. For cluster analysis, a subset of 52 of the original sample of 200 informants was used, and the encoded transcriptions of their interviews together with the associated social data was input and stored on the mainframe computer at Newcastle University. The cluster analysis package Clustan (Wishart, 1969) was used. For each of the fifty-two informants, the number of token occurrences of each of 542 'state' segment types S occurring in the analysis was counted, where a state represents a fairly narrow phonetic transcription within a phonemic frame of reference which permits both phonetic and phonemic representation and comparison across different varieties (diasystems); in the frame of reference the state variants were grouped into fifty-one Overall Units (OUs). Each informant's segmental phonological profile was thus represented as a 542-element integer-valued vector V, in which any element V_i contained the number of token occurrences of state S_i. The set of informant vectors was stored in a 52×542 matrix which, after normalization for variation in the number of segments per interview, served as input to the various clustering algorithms used in the analysis.

2.2 The TLS cluster analysis and its problems

Jones-Sargent (1983) performed cluster analyses based on the segmental phonological data and on the social data, and then attempted to relate the two in a sociolinguistically meaningful way. Since this discussion is about methodology, we can simplify matters by concentrating on the phonological analyses alone.

In order to derive social and linguistic classifications, Jones-Sargent (1983) used hierarchical agglomerative cluster analysis, a technique which aggregates data points into groups in accordance with their relative proximities in a multidimensional data space, assigning a constituency structure to the clusters in terms of order of fusion based on pairwise distance or similarity coefficients; this constituency structure can be and usually is represented by a tree diagram. There are numerous distance measures and clustering algorithms to choose from (Everitt, 2001), and therefore a large number of possible combinations. After experimentation, Jones selected the combination of Euclidean distance and Ward's method. Because of the aforementioned software limitations, that is, implementation limitations on the number of variables which could be processed, the segmental

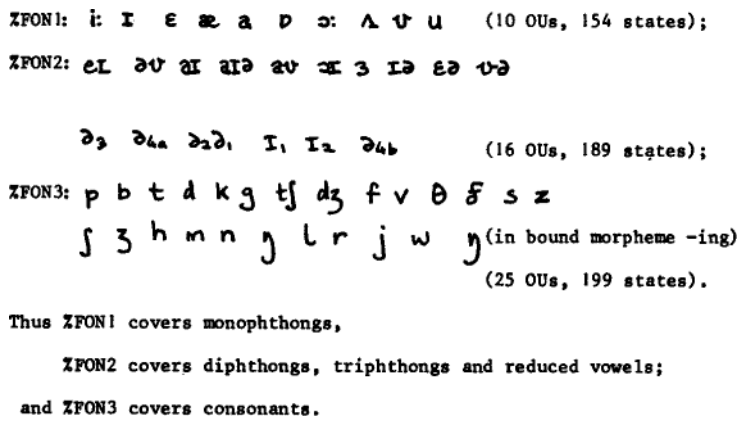

Fig. 1 Tripartite division of segmental phonological variables, from Jones-Sargent (1983, p.195)

phonological data had to be partitioned into three groups and analyzed separately. The partition of variables (OUs) is shown in Fig. 1, where %FON1 covers monophthongs, %FON2 diphthongs, triphthongs, and reduced vowels, and %FON3 consonants. The results are shown in Fig. 2, where the trees show the constituency structure of the informant-clusters at the leaves, and the length of the branches indicates relative distance in 542-dimensional space at which fusion of the pairs occurred.

In all the trees there are two very clear clusters—the lowest one and the remainder. A social correlation exists with respect to these two main clusters—the lowest cluster corresponds to a small group of well-educated middle-class Newcastle speakers, and the remainder to broadly working class, less well-educated speakers from Gateshead. Apart from this, relatively little clear correlation between phonological and social data was found.

Jones presented a number of interpretations, one of which was that the interrelationships between linguistic and social factors in the data were, in general, too complex to emerge as a set of simple correlations between linguistic and social clusters. Another interpretation was that there was something amiss with the cluster analysis—that its classifications of the data did not accurately represent its true structure, and that if the true structure were elicited, a systematic relationship between phonological and social factors might emerge. Jones had investigated the possibility of false structure being imposed by clustering methods with an experiment on artificial data sets (Sargent (nee Jones) 1979), and demonstrated that there was indeed a propensity of cluster analysis algorithms to impose structure on data by showing that absolutely unclassifiable data sets are forced into spurious cluster formations by certain combinations of distance/similarity coefficients and clustering algorithms. Specifically, the situation is that, given a data set D, different combinations of distance measure and clustering algorithm give structural analyses of D, which

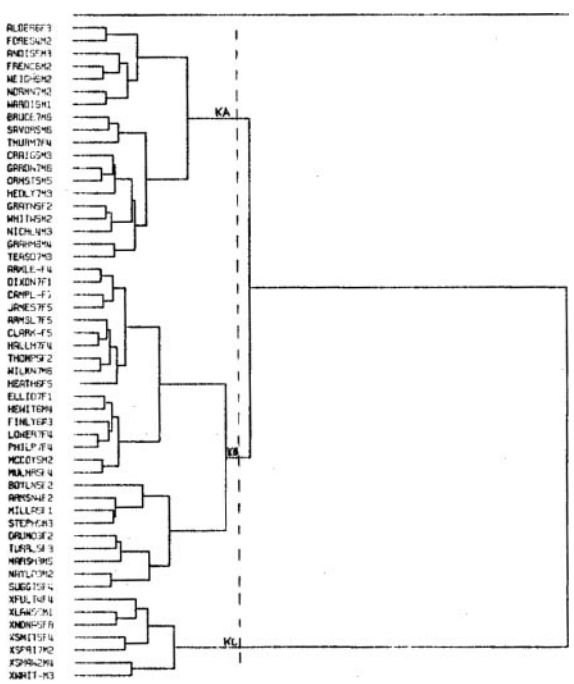

Fig. 2 Cluster trees for the segmental phonological groups of Fig. 1, from Jones-Sargent (1983, p.197–199)

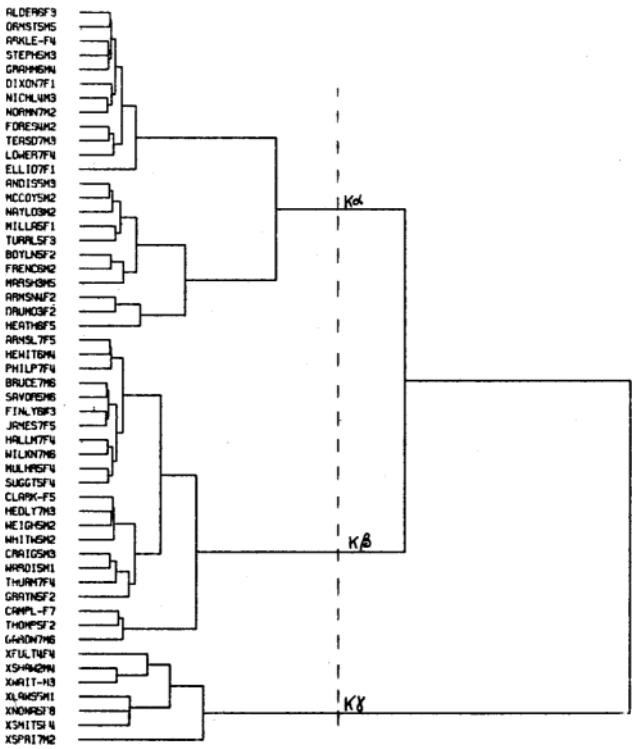

Fig. 2 Continued

differ from one another to greater or lesser degrees. This finding has since been repeatedly confirmed by both theoretical and applied research in cluster analysis, and is exemplified with respect to the NECTE data in our own cluster analyses using various combinations of distance measure and clustering algorithms, a selection of which is shown in Fig. 3 below.

Before having a look at these analyses, a few preliminary observations are necessary:

(1) The data that Jones originally worked with is no longer directly available. Little work was done on the TLS after Jones-Sargent (1983), and the original interviews and transcriptions, together with the electronic files used for analysis, were almost lost. One of the main aims of the NECTE project has been to preserve and reconstitute these materials for the sociolinguistics and dialectology community. In the course of doing this, 64 segmental phonological transcriptions and the corresponding computer files were recovered, and a slightly larger data set than the one Jones used is therefore available in principle. In practice, however, fifteen of these computer files have not yet been restored, so only forty-nine informant files are currently available.

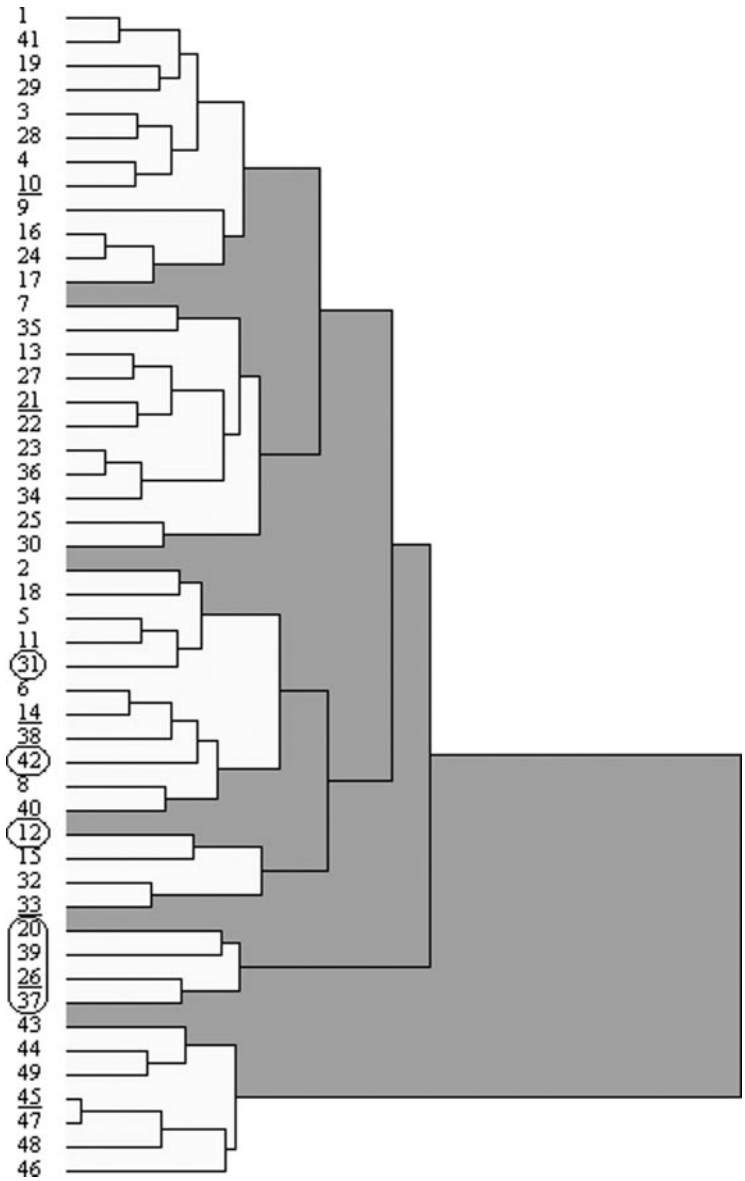

Fig. 3 Cluster analyses of the NECTE data using four different clustering algorithms

The number of occurrences of each of the possible state segment types for each informant were counted and the totals normalized for variation in the number of segments per informant. All the segmental variables for which there was zero or very low variance were then removed to avoid skewing the results by including unnecessary variables in the analysis. As a result, the data set in what follows is a 49×271 matrix, where each of the 49 rows constitutes an informant's segmental phonological profile.

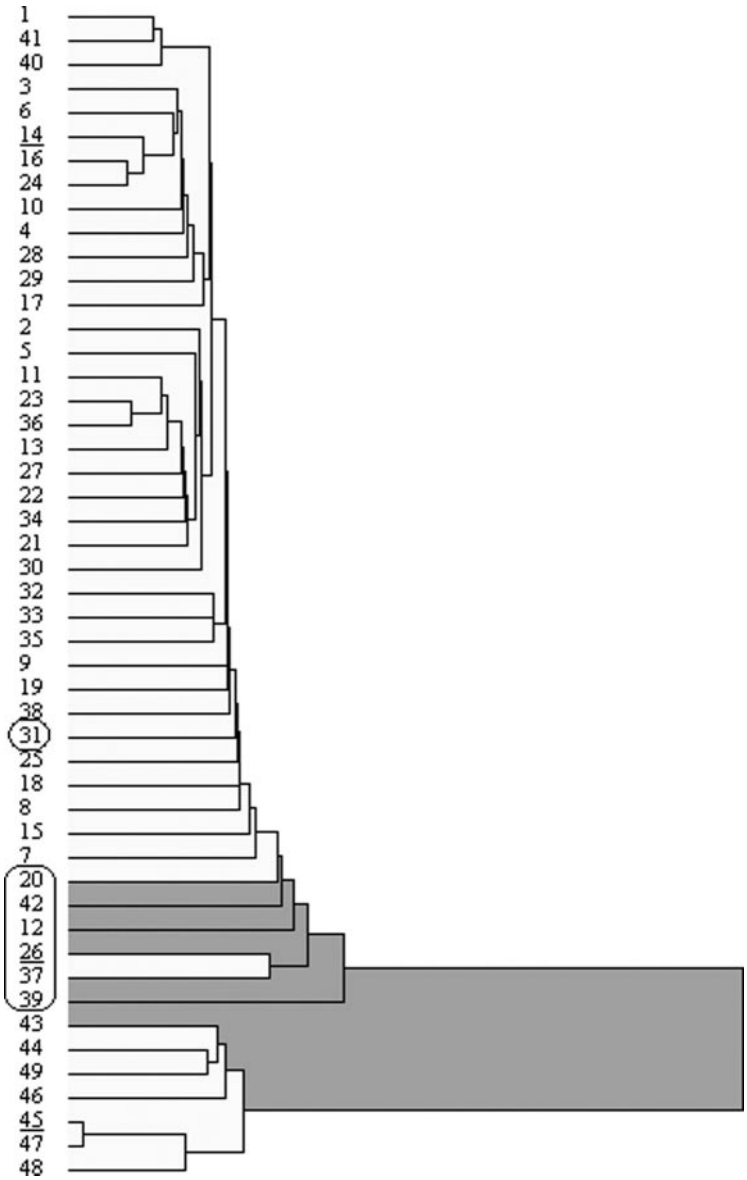

Fig. 3 Continued

(2) The original analyses were done in three parts because the clustering packages available at the time did not allow all the data to be processed at once. This limitation has long been superseded, and our own analyses were done on the entire data set.
(3) The NECTE labelling of informants differs from that of the TLS, which complicates comparison with the original results. This does not matter for present purposes, though, since the discussion is interested in structural variability of cluster trees, not in detailed examination of particular informants.

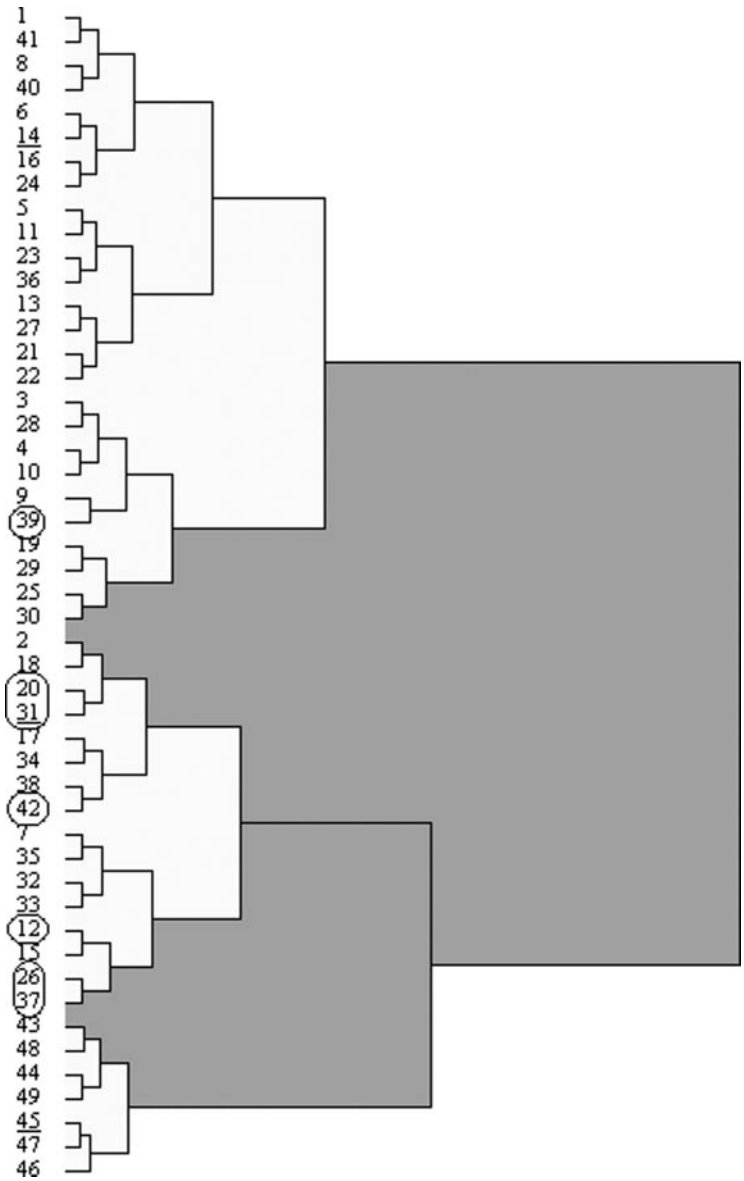

Fig. 3 Continued

The following are sample results from our own analyses of the NECTE data reconstructed from the TLS, using the current version of Clustan. To simplify matters, the squared Euclidean distance measure was used in all cases, and only the clustering algorithm was varied. Each tree is labelled according to the algorithm that generated it; details of the various algorithms are available in Everitt (2001).

In these trees, the numbers are informant labels. As in the earlier TLS analysis, there is a strong distinction (shown in the diagram by shading) between the Newcastle informants at the bottom of the trees

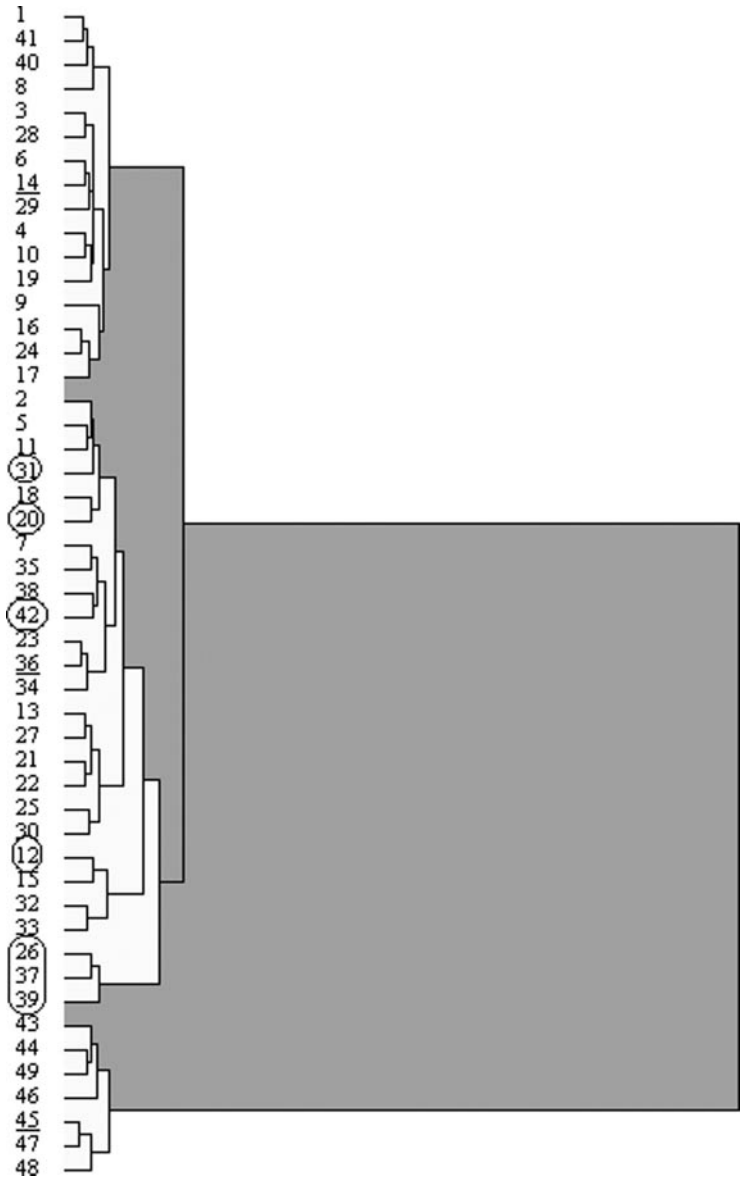

Fig. 3 Continued

and Gateshead informants above. The Gateshead subtrees look like they differ across clustering algorithms, however, and close inspection confirms it. To exemplify the variation, a well-defined subtree was chosen at random from the complete linkage tree in the upper left cell of Fig. 3, and circled for clarity. The corresponding cluster in the single-link tree contains two additional informant labels which, in the other trees, are fairly widely scattered around. The increase in sum of squares one is lacking one of the original labels, 20, which is now in a completely different part of the tree, and the sum of

squares tree has lost any sense of the original cluster. In summary, there is a family resemblance across trees, but they differ in detail to varying degrees.

The problem was for TLS, and remains for us, that different combinations of distance measure and clustering algorithm in general yield different analyses of the same data set, and that there is no obvious way of selecting the 'best' analysis. How reliable a tool, therefore, is hierarchical cluster analysis for sociolinguistic research?

3 Proposed solution to the cluster analysis problem

Ways of assessing the validity of hierarchical cluster solutions have been developed, but the general view is still that there is no single best combination of distance measure and clustering algorithm, that 'best' has to be seen relative to the characteristics of the data to be analyzed, and that hierarchical cluster analysis should be used with care and full awareness of its pitfalls (Everitt, 2001; Gore, 2000).

The approach proposed here develops the technique, employed by Jones and often used in engineering applications, of analyzing the data of interest using two or more fundamentally different types of clustering algorithm. The idea is to generate a range of analyses, and then to select the classification on which the various methods agree most closely. There is no guarantee that the selected classification is the optimal one, and not even any obvious way of estimating the probability that it is, but it seems intuitively clear that if a range of different methods converge on a similar grouping of data points, then one can have greater confidence in that classification than in one generated by a single method.

Among alternatives to hierarchical cluster analysis are non-hierarchical clustering techniques that project structure in high-dimensional data into low-dimensional space so that it can be graphically displayed. Only a brief account of such dimensionality reduction techniques can be given here (for details see Hair *et al.*, 1998 chs. 3, 10; Tinsley & Brown, 2000, chs. 10, 12; Tabachnik & Fidell, 2001, ch.13).

The fundamental ideas underlying dimensionality reduction techniques are

(1) that there is often redundancy in high-dimensional data sets in the sense that the variables overlap to greater or lesser degrees in the information they represent, or, in other words, that the variables are correlated,
(2) that most or all of the information in the original, high-dimensional correlated variables can be captured by removing the redundancy and representing what remains by a smaller number of uncorrelated variables, and

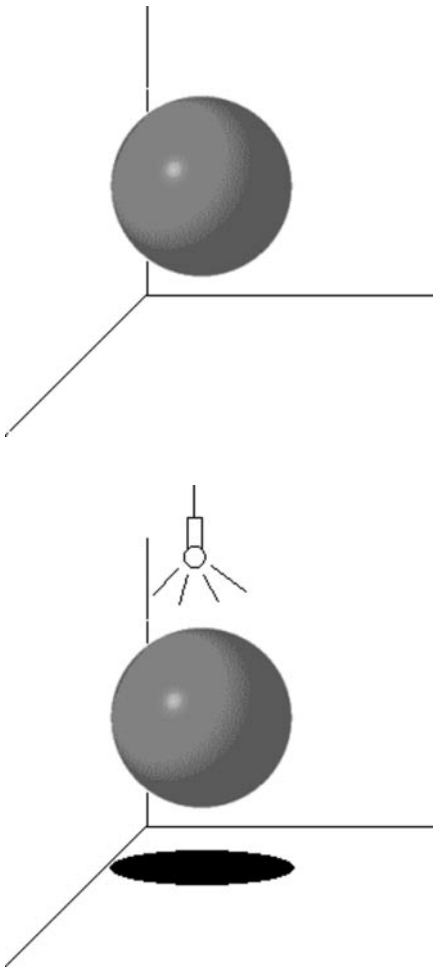

Fig. 4 projection of a three-dimensional object into two dimensions

(3) that if the information in the original variables can be represented by three or fewer variables without losing too much information, then the structure in the data can be graphically represented.

An intuition for what is involved here can be gained by picturing a three-dimensional shape in a three-dimensional space, as in Fig. 4a. If a light shines onto the shape from above, as in Fig. 4b, its shadow is projected onto the floor of the space. The shadow is a two-dimensional representation of the three-dimensional sphere, just as one's own shadow is a two-dimensional representation of oneself; these representations capture some of the essentials of the three-dimensional shapes, though at the cost of losing some information. In other words, dimensionality reduction can be achieved by projection. This idea, moreover, generalizes mathematically to any dimensionality—a shape of arbitrary dimensionality n can be projected into any m-dimensional space, where $m < n$.

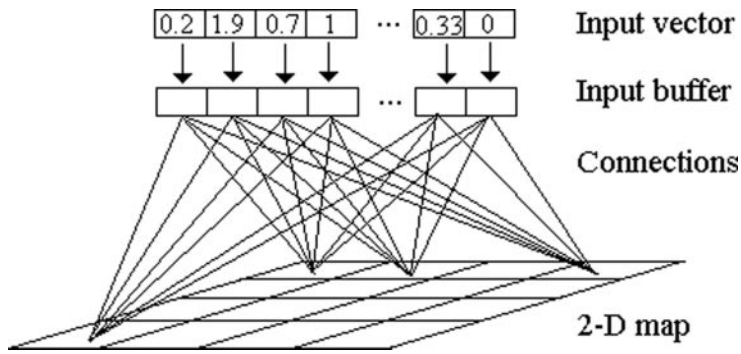

Fig. 5 A self-organizing map

Dimensionality reduction techniques include principal component analysis (Jolliffe, 2002), factor analysis (Gorsuch, 1983), multidimensional scaling (Borg & Groenen, 1997), Isomap (Tenenbaum *et al.* 2000), locally linear embedding (Roweis & Saul, 2000), and self-organizing maps (SOMs) (Kohonen, 2001). We have chosen to begin with the last of these because it has been successfully used for cluster and related types of analysis in a wide variety of disciplines (Kaski *et al.*, 1998; Oja *et al.*, 2003), and its properties are thus well documented. A SOM was used to analyze the NECTE data, and the results compared with a variety of hierarchical classifications.

The SOM is an artificial neural network that was originally invented to model a particular kind of biological brain organization. It can, however, be used as a data analysis tool without reference to biology. Used in this way, it is a method for projecting data of arbitrary dimensionality into two-dimensional space (Fig. 5).

A SOM has three components:

(1) An input buffer with as many cells as there are dimensions in the data—for, say, a data set D in which each item is a length-6 vector, the first component of the *i*th vector $d_i \in D$ is loaded into the first buffer cell, the second component of d_i is loaded into the second cell, and so on.
(2) A two-dimensional grid or lattice of processing units that respond selectively to inputs, as described below.
(3) Connections between the buffer and the lattice. Each connection has a specific transmission efficiency or strength, and, if the SOM is to behave in a useful way, connection strengths must vary systematically, again as described below. Each input buffer cell is connected to all the units in the lattice, but for clarity only a few connections are shown.

Assume a data set D consisting of k length-n vectors. Then dimensionality reduction of D is achieved by loading the k vectors d_i successively into the input buffer. For each d_i, the values in the buffer are propagated through all the connections in the SOM. Because

of the variation in connection strength, a given d_i activates one unit more strongly than any of the other units, thereby associating each d_i with a specific unit in the lattice. When all the d_i have been projected in this way, the result is a pattern of activation across the lattice. This pattern is the projection of the *n*-dimensional data into two-dimensional space, and the data's cluster structure can be seen in the lattice configuration.

Clearly, the above projection of D will only correctly represent D's structure in the lower-dimensional space if the configuration of connections is appropriate to the task—random connections will give a random result. Because a SOM is an artificial neural network, the connection configuration appropriate to projection of any given D is not usually specified explicitly, but is rather learned incrementally from the data itself. The details of SOM learning are too complex for presentation here (Kohonen, 2001), but in essence it involves a large number of successive presentations of data vectors, with modification to the connections at each presentation in such a way as to associate each d_i with a specific lattice unit; training stops when all the d_i are so associated and no more changes to the connections are required.

4 Application to TLS data

This section uses a SOM to cluster-analyze the same NECTE data as that used in the above hierarchical cluster analyses, and then compares the result to those analyses with the aim of determining the extent to which the SOM can aid in the selection of the 'true' data structure. The analysis was carried out using a SOM implementation which we developed in Matlab, and the results were confirmed using the SOM Toolbox for Matlab produced and distributed by the Neural Networks Research Centre at the Helsinki University of Technology (http://www.cis.hut.fi/research/som_lvq_pak.shtml). Figure 6 shows a 20×20 unit map of the NECTE data.

4.1 Problems

4.1.1 *Map partitioning*

The preceding hierarchical cluster analyses give a good indication of the structure of the data, and thus guide the partition of the map into clusters. But what if no such prior indications are available? Figure 7 shows two possible partitions out of (very) many. Which is correct relative to the data structure underlying the map representation? Everything depends on prior information and/or necessarily subjective visual intuitions. To be useful as an analytical tool, however, the SOM's representation of data structure has to be unambiguously interpretable on its own merits.

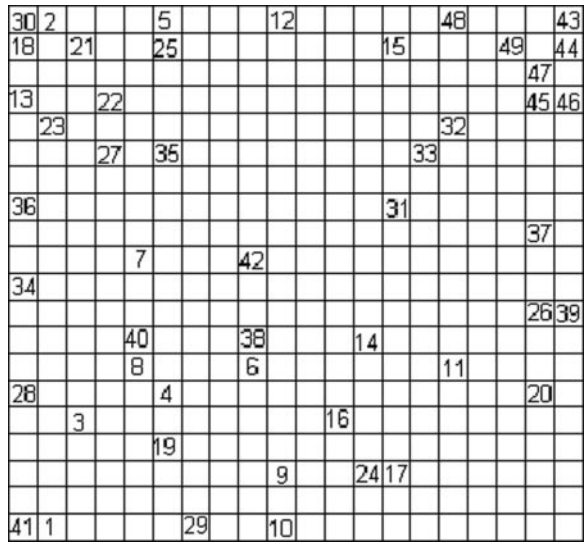

Fig. 6 A SOM map of the NECTE data

4.1.2 *Interpretation of map proximity*

Self-organizing maps map the topology (Munkres, 1999) of high-dimensional data to a two-dimensional representation of that data. This means that points which are close together in the high-dimensional space will be close together on the SOM map. However, the converse does not hold—just because points are close together on the map does not mean that they are close in the input space.

To see this, imagine a sheet of paper with three points A, B, and C marked on it, as in Fig. 8a. If the sheet is now folded as in Fig. 8b, A and B can be brought closer together than B and C, even though A and B are much further apart on the surface of the paper than B and C. If the folded paper is now rotated so that it is seen from above (Fig. 8c) the relative distance of B and C on the projection reflects their relative distance on the surface of the paper, but the relative distance of A and B on the projection does not.

A SOM preserves topological distance only, so spatial distance among points on the low-dimensional map bears no systematic relationship to the distance among points in the high-dimensional input space. When interpreting a map, this means that spatial proximity between and among points is not a reliable guide to visual identification of clusters.

4.1.3 *Parameter selection*

When setting out to use a SOM for data analysis, the researcher has to specify a variety of parameters such as the number of units in the lattice (Kohonen, 2001). There is a large and, in theory, unbounded number of parameter value combinations, and it is

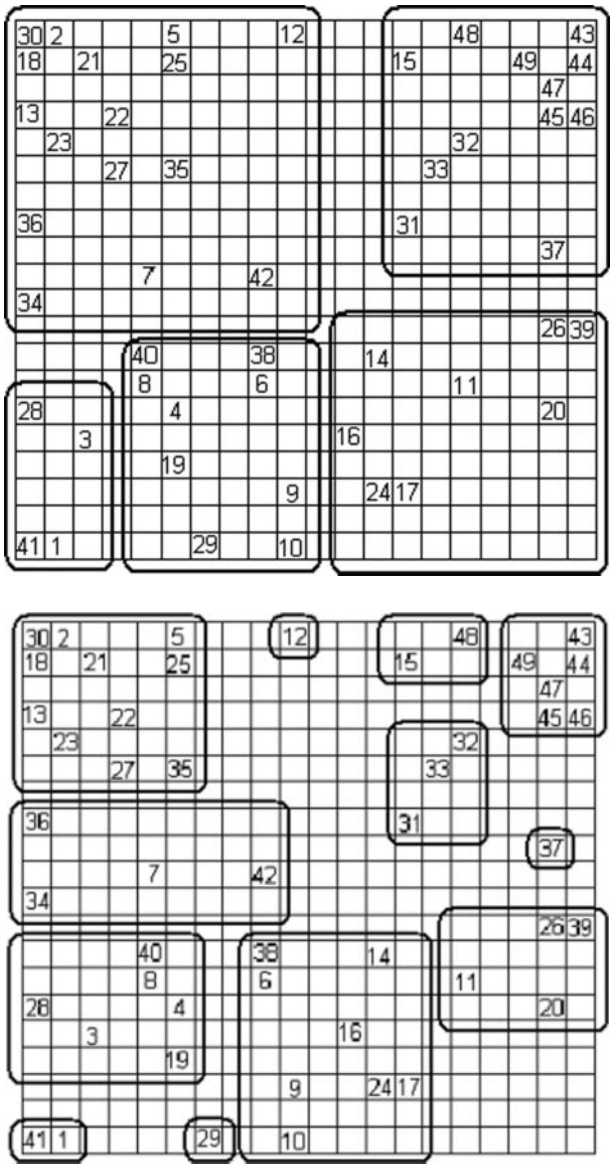

Fig. 7 Different possible partitions of Fig. 6

possible in principle that different choices will yield different maps relative to any given data set, creating a problem analogous to the one observed for different choices of hierarchical clustering algorithm. Numerous analyses were carried out using different parameter combinations, and apart from obviously unsuitable choices such as a very small number of units that cannot usefully display the data structure, the SOM was found to be fairly insensitive to parameter value choices—the various maps had a strong family resemblance similar to the one found for the hierarchical cluster analyses. But there was significant variation as well. This means

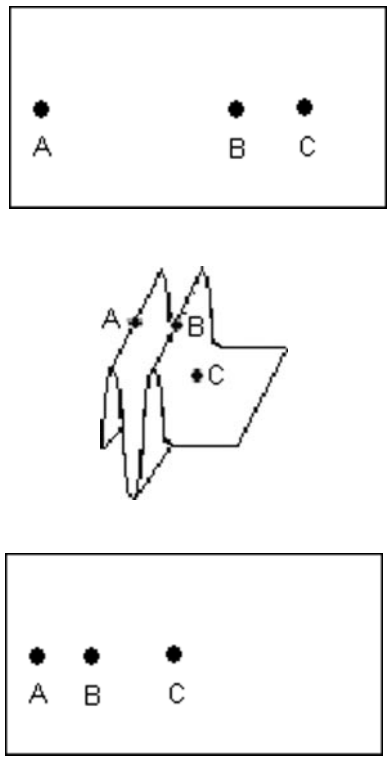

Fig. 8 Topological and spatial distance

that one cannot simply take a SOM with one specific selection of parameters as canonical for purposes of comparison with other clustering methods.

4.2 Solutions

The essential problem in both (i) map partitioning and (ii) interpretation of map proximity is that activation maps like the one in Fig. 6 do not contain enough information to allow for reliable interpretation in the general case. The solution lies in enhanced map visualization methods that calculate and show cluster boundaries explicitly rather than requiring the analyst to guess them. Various ways of doing this have been developed (Merkl and Rauber, 1997; Vesanto, 1999). A popular one is the unified distance matrix, or u-matrix, which represents the degrees of activation of map units as a landscape in which the 'valleys' represent clusters, and the 'mountains' represent boundaries between clusters; details of how a u-matrix is calculated from a SOM are given in Ultsch (1999, 2003).

A u-matrix representation of Fig. 6 shows quite clearly where the clusters and the cluster boundaries are; informant labels have been moved off the landscape for legibility at the cost of losing within-cluster relativities (see Fig. 9). Note, for example, how the

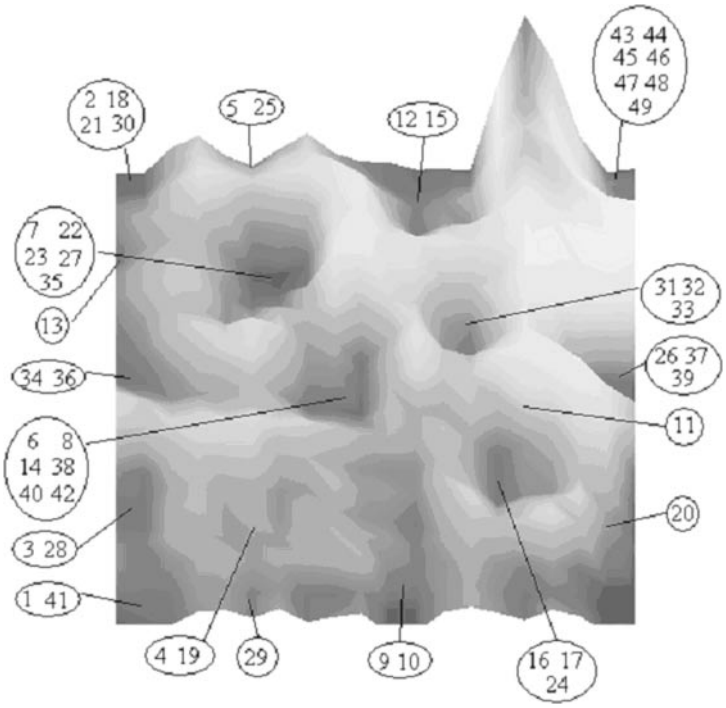

Fig. 9 A u-matrix representation of Fig. 6

Newcastle group of informants is separated from the rest by a particularly high mountain range, and that informants 15, 32, and 33, which one might be tempted to include in the Newcastle group on the basis of spatial proximity in Fig. 5, are in fact strongly distinct from it.

With regard to (iii), the consequences of parameter selection, we have no solution at present. The aim in future work is systematically to assess the degree of variability among maps in response to parameter selection, and if possible to determine optimal settings for them.

4.3 Comparison with hierarchical cluster analyses

Detailed comparison of the u-matrix representation of the SOM in Fig. 7 and the hierarchical cluster trees of Fig. 3 shows that the best match is with the complete link tree. This is shown in Fig. 10 below, where corresponding regions of the SOM and the tree are labelled A–H.

The most obvious agreement between SOM and complete link tree is C, the Newcastle group, where the relatively large distance between it and the rest of the tree corresponds to the highest peak in the SOM landscape separating it from the remainder of the map. There is also an exact agreement between regions G, A, D, B, and F on the SOM and the corresponding subtrees in the cluster tree. The only

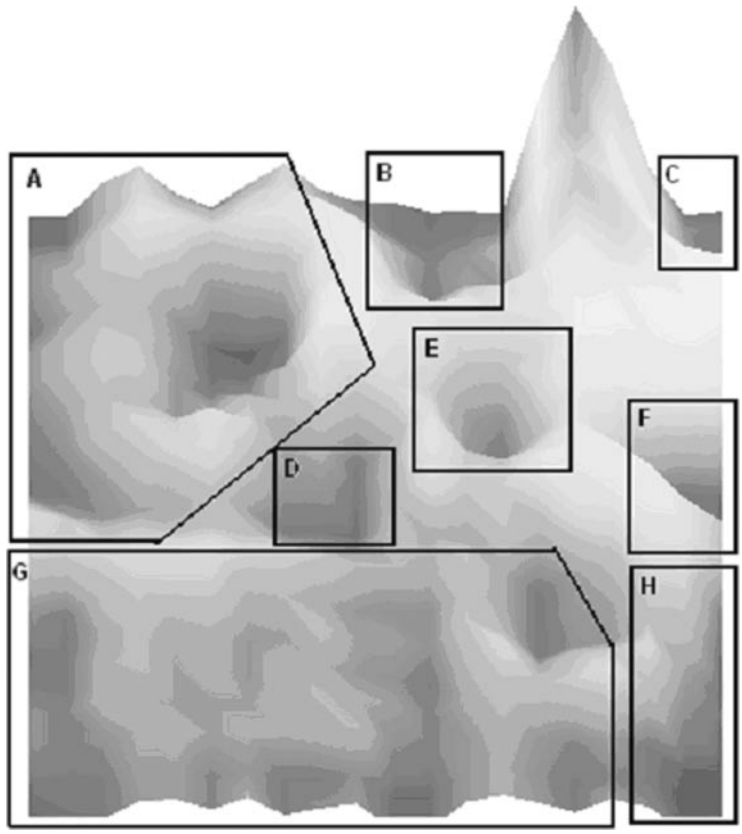

Fig. 10 Comparison of SOM and complete link tree clustering

inconsistencies are in E and H:

(1) E on the SOM includes 31, but in the cluster tree 31 is fairly distant, between A and D.
(2) H on the SOM includes 11, but in the cluster tree it is fairly distant, again between A and D.

Finally, the relativities of the groups on the SOM correspond to those of the main subtrees in the cluster tree—G and A are adjacent, D is next to them, and so on.

The remaining three cluster trees all correspond to the SOM in separating the Newcastle group C strongly from all the others. With respect to the other groups, however, the close agreement between the SOM and the complete link cluster tree breaks down to greater or lesser degrees. In the sum of squares tree, for example, G is largely as it is in the SOM and the complete link tree, but includes 6, 8, and 40 from D, and lacks 17, which is now near the bottom of the tree, just above C. Such breakdown is more extensive for the increase in sum of squares tree, and most in evidence in the single link tree.

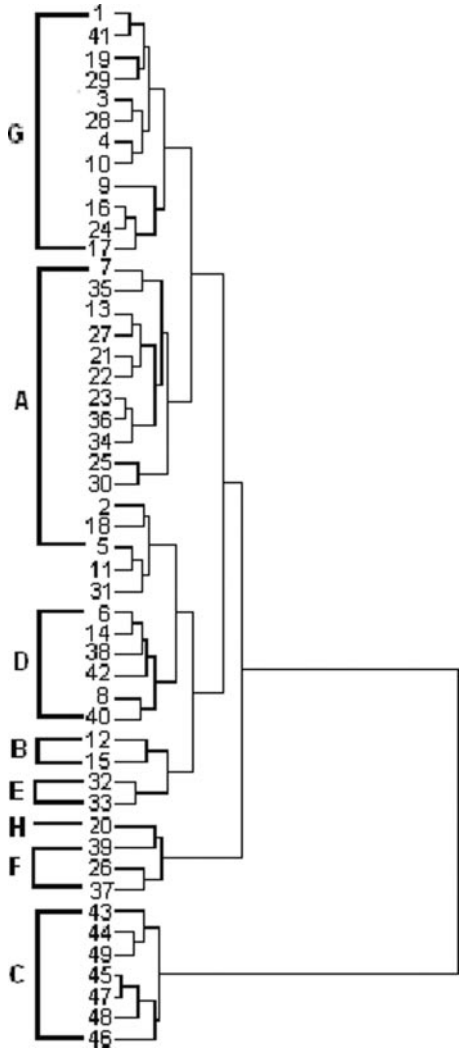

Fig. 10 Continued

5 Conclusions

The problem this article set out to address was the feasibility of an empirical approach to sociolinguistic analysis of the NECTE corpus in the light of the variation in results that different hierarchical cluster analysis methods generate for any given data set. The proposed solution was to analyze the NECTE data using hierarchical methods in conjunction with one or more fundamentally different types of clustering algorithm, and then to select the analysis on which the hierarchical and the other method(s) agree most closely. A non-hierarchical method, the SOM, was used to exemplify this approach. The result was a close though not perfect match between the SOM and the complete link analyses, and less good matches between, in

descending order of closeness, sum-of-squares, increase in sum-of-squares, and single link. The SOM can therefore be said to support the complete link analysis of the NECTE data.

There is, however, an important reservation. The SOM analyses of the NECTE data varied with changes in user-defined training parameters, and they are consequently open to the same criticism of inconsistency as the hierarchical methods. In other words, the SOM cannot be an objective arbiter for the results of hierarchical cluster analysis. It may be that variability in the SOM results can be reduced or eliminated, or that some other cluster analysis method will be found to be more consistent, but at this stage all we would wish to claim is that, because the SOM and at least one hierarchical clustering method give very similar results, that result appears to provide a good basis for understanding the structure of the NECTE data and for generating hypotheses about it, which is after all the point of exploratory multivariate analysis.

Finally, it needs to be stressed that the foregoing discussion has been methodological. The analyses were based on incomplete data, and should not be construed as an attempt at a definitive analysis of the NECTE corpus.

References

Borg, I. and Groenen, P. (1997). *Modern Multidimensional Scaling—Theory and Applications*. Berlin: Springer.

Everitt, B. (2001). *Cluster Analysis*, 4th edn. London: Arnold.

Gore, P. (2000). Cluster Analysis. In Tinsley, H. and Brown, S., *Handbook of Applied Multivariate Statistics and Mathematical Modeling*. New York: Academic Press.

Gorsuch, R. (1983). *Factor Analysis*, 2nd edn. Mahwah, NJ: Lawrence Erlbaum Associates.

Hair, J., Anderson, R., Tatham, R. and Black, W. (1998). *Multivariate Data Analysis*, 5th edn. Upper Saddle River, NJ: Prentice-Hall International.

Jolliffe, I.T. (2002). *Principal Component Analysis*, 2nd edn. Berlin: Springer.

Jones, V. (1978). *Some Problems in the Computation of Sociolinguistic Data*. Ph. D. thesis, University of Newcastle upon Tyne.

Jones-Sargent, V. (1983). *Tyne Bytes. A Computerized Sociolinguistic Study of Tyneside*. Frankfurt Am Main: Peter Lang.

Kaski, S., Kangas, J., and Kohonen, T. (1998). Bibliography of Self-Organizing Map (SOM) Papers: 1981–1997. *Neural Computing Surveys*, **1**: 102–350.

Kohonen, T. (2001). *Self-Organizing Maps*, 3rd edn. Berlin: Springer.

Merkl, D. and Rauber, A. (1997). Alternative ways of cluster visualization in self-organizing maps, *Proceedings of WSOM'97: Workshop on Self-Organizing Maps*, *Helsinki*, Finland.

Milroy, L., Milroy, J., Docherty, G. J., Foulkes, P., and Walshaw, D. (1997). Phonological Variation and Change in Contemporary English: Evidence from Newcastle-upon-Tyne and Derby. In Condé Silvestre, J. C., and

Hernández-Compoy, J. M. (eds) *Variation and Linguistic Change in English*, pp. 35–46. Cuadernos de Filología Ingelsa. Oxford: Blackwell.

Munkres, J. (1999). *Topology*, 2nd edn. Upper Saddle River, NJ: Prentice Hall.

Oja, M., Kaski, S., and Kohonen, T. (2003). Bibliography of Self-Organizing Map (SOM) Papers: 1998–2001 Addendum. *Neural Computing Surveys*, **3**: 1–156.

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

Roweis, S. T. and Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, **290**: 2323–26.

Sargent, V. (1979). Cycles and the equal society. *Classification Society Bulletin*, **4**(3): 31–45.

Strang, B. (1968). 'The Tyneside Linguistic Survey'. Paper presented at the International Congress on Dialectology, Marburg, 1965, *Zeitschrift für Mundartforschung, Neue Folge*, **4**: 788–94.

Tabachnik, B. and Fidell, L. (2001). *Using Multivariate Statistics*, 4th edn. Allyn and Bacon.

Tenenbaum, J.B., de Silva, V., and Langford J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, **290**: 2319–23.

Tinsley, H., and Brown, S. (2000). *Handbook of Applied Multivariate Statistics and Mathematical Modelling*, Academic Press.

Ultsch, A. (1999). Data Mining and Knowledge Discovery with Emergent Self-organizing Feature Maps for Multivariate Time Series. In Oja, E., and Kaski, S. (eds), *Kohonen Maps*, Elsevier.

Ultsch, A. (2003). *U^*-Matrix: a Tool to Visualize Clusters in High Dimensional Data*, University of Marburg, Department of Computer Science (Technical Report 36).

Vesanto, J. (1999). Som-based data visualization methods. *Intelligent Data Analysis*, **3**: 111–26.

Wishart, D. (1969). *FORTRAN II Programs for 8 Methods of Cluster Analysis (CLUSTAN 1)*, Computer Contribution 38, State Geological Survey, University of Kansas.