SOME PROBLEMS IN THE COMPUTATION OF SOCIOLINGUISTIC DATA

BY

VALERIE M. JONES

Thesis submitted for the degree of Doctor of Philosophy in the University of Newcastle upon Tyne.

AUGUST 1978.

CONTENTS

-oooOooo-

ACKNOWLEDGEMENTS

The research reported upon here was undertaken whilst the author was in receipt of a grant from the Department of Education and Science.

I am very grateful for the computing facilities made available by the Computing Laboratory of the University of Newcastle upon Tyne and for the help given me by its members of staff. Dr. John Leece was helpful in interpreting some of the early difficulties in using the CLUSTAN package.

Various parts of this work have been reported to Research Seminars in the School of English, Newcastle upon Tyne and I am grateful for the reactions of its members. Parts of Chapters 4, 5 and 6 have been reported to the International Summer School on Computational Linguistics (Pisa 1977) and the International Symposium on Literary and Linguistic Computing (Birmingham 1978).

I am indebted to Vince McNeany for the collection and basic linguistic analysis of the data which are discussed here.

Finally, I am grateful to my supervisor, Mr. John Pellowe, for his interest in, and support of, my research.

LIST OF FIGURES

Figs. cont.

Figs. cont.

LIST OF TABLES

Tables cont.

Table cont.

ABBREVIATIONS

BIN VAR	-	(CLUSTAN) binary variable
BIT	-	Binary Digit
BIT string	-	sequence of binary digits
BPFR	-	Binary Percentage Frequency Ratio
CLU(S) VAR	-	CLUSTAN Variable
D	-	Binary Euclidean Distance coefficient
D^2	-	Squared Euclidian Distance coefficient (numeric data)
dig.	-	digit
f.e.	-	further education
K	-	cluster (e.g. KA = cluster A)
L	-	localised
LK	-	linguistic cluster (at 2-K level)
LDP	-	linguistic data processing (Lamb: 1965)
l.m.a.	-	legal minimum (school leaving) age
NC	-	non-comparable (missing data)
NL	-	non-localised
NPL	-	New Programming Language (later called PL/1)
Num.	-	numeric (CLUSTAN term for quantitative rather than binary data)
NUMAC	-	Northumbrian Universities' Multiple Access Computer
MC	-	multiple coding
MTS	-	Michigan Terminal System
OM	-	ordered multistate
OU	-	overall unit
PDV	-	Putative Diasystemic Variant
PL/1	-	Programming Language 1
Q	-	question
SocSp	-	social space
SC	-	social class

SES	-	Socio-economic status
SocK	-	Social cluster
TLS	-	Tyneside Linguistic Survey
UM	-	unordered multistate
RRR	-	reading, writing and arithmetic
V,VAR,Var.	-	variable
x-linguistic	-	extra-linguistic (social)
%FON1	-	segmental phonological subspace 1. (Monophthong vowel OUs)
%FON2	-	segmental phonological subspace 2 (Dipthong, triphthong and reduced vowel OUs)
%FON3	-	segmental phonological subspace 3 (consonant OUs)

CHAPTER 1

INTRODUCTION

The research described in this thesis is concerned with some of the problems encountered in the processing of sociolinguistic data.

Different methodologies are seen as different sets of strategies for coping with the problems which arise from investigations of sociolinguistic variability within any speech community.

One early approach to the analysis of sociolinguistic variation (that of Labov: 1963, 1966) is discussed, and some of the difficulties raised by this approach are indicated. One investigation of sociolinguistic variability in a British urban setting (Trudgill: 1974) is also described (Trudgill's study is based on Labov's (1966) general methodology).

The Tyneside Linguistic Survey[FN] (T.L.S.) is offered as an alternative approach, which overcomes some of the problems inherent in Labov's methods.

FN. Department of English Language, University of Newcastle upon Tyne, U.K.

The T.L.S. methodology embodies a set of heuristic strategies whereby the social and linguistic configurations within the speech community are empirically determined. As such, the T.L.S. represents an attempt to model dynamically the sociolinguistic ecology of the community under study.

The model and methodology of the T.L.S. are described in detail.

Not surprisingly, the radical approach of the T.L.S. generates a new set of problems for sociolinguistics. These problems highlight issues of significance to linguists and sociolinguists alike, which, however, have been ignored, or glossed over in previous work.

Some of the problems are concerned with linguistic analysis, and with the design of an adequate linguistic coding frame. Some are concerned with adequate representation of social differentiation within a community. Some of the problems are specific to the methods employed within the T.L.S., (e.g. problems in classificatory theory, difficulties in interpretation of

the results produced by cluster analysis).

Linguistic and social data collected by means of tape-recorded interviews are processed by various computational techniques. The computer programs written to process this data are described, and the results of the sociolinguistic classification are presented. The findings have important theoretical and practical consequences for the discipline of sociolinguistics.

The prime focus of this thesis is the problems inherent in the analysis of sociolinguistic data. A complete survey of the history of sociolinguistics is not attempted: rather, two contrasting methodological approaches are described, and compared as alternative strategies for dealing with the problems identified.

Labov's approach to sociolinguistic modelling is now described briefly.

The Labovian sociolinguistic methodology

Labov (1966), in his study of "The social stratification of English in New York City", attempted to deal with the "structured heterogeneity" (Weinreich, Labov & Herzog: 1968) of language and society, by discovering correlations between linguistic and social factors, and determining the mechanisms underlying synchronic, and possibly diachronic, variation.

Labov describes his study as "an investigation of language within the social context of the community in which it is spoken." (Labov, 1966, p.3) which aims at discovering and defining "a consistent and coherent structure for the speech of this community." (Labov, 1966, p.9).

Labov's aims were to measure the linguistic behaviour of a sample of native New Yorkers living in the Lower East Side, mainly with respect to five phonological variables:

(r) (oh) (eh) (th) (dh) ,

to quantify the frequency of usage of stigmatised and prestige variants of each across a range of speech styles, and eventually to "deal with the New York vowel system as a whole." (Labov , 1966, p.5). Labov's interviews were structured to elicit samples of speech ranging from 'formal' to 'casual' speech styles, and included a subjective evaluation test designed to discover the opinion of each informant as to the 'correct' version of each variable, and also the variant which he claimed to realise most frequently.

Informants from 5 ethnic groups were classified socially according to a socio-economic class index based on information derived from a previous, independently motivated, survey (conducted by Mobilisation for Youth). This survey provided sociological information on the entire population of the area under investigation. This social information was reduced to a ten-point scale based on three factors: occupation, family income and education. Labov comments, (Labov, 1966, p.171), that this scale was "a useful device for dividing the population along the socio-economic scale into three units of approximately equal size."

The social classification corresponds to the groups designated working

class, lower middle class, and middle class, occupying the categories on the 10-point scale 0-2, 3-5 and 6-9 respectively. In examining the distribution of different realisations of the sound features across these groups, Labov found evidence of awareness of the prestige value of the middle class variants throughout all but the lowest of the working class strata. Generally the frequency of the prestigious alternant increased with formality of style, and with increase of socio-economic class. An unusual phenomenon was observed in the case of the variant presence of post-vocalic /r/:the usual distribution pattern was disrupted for lower middle class speakers. In careful speech this group tend to use this (prestige) feature with higher frequency than the middle class group, and this crossover effect is interpreted by Labov as evidence that this feature is undergoing diachronic change in New York speech, and therefore functions more overtly as a social marker.

The relationship between the incidence of prestige forms and the social stratification was found not to be linear: social mobility was concluded to be the most significant factor determining the social value attributed to variants of features by informants. Those informants showing upward social mobility (shown to be linguistically insecure[FN] as well as socially mobile), appeared to be more wary of using stigmatised forms.

FN. The lower middle class group (categories 6-8 on the Mobilisation for Youth 10-point SEC scale) have the highest scores on Labov's Index of Linguistic Insecurity. (Labov: 1966, p.477).

Trudgill, (1974), in his study of sociolinguistic variation in Norwich English, adopts Labov's general methodology. He quotes (p.33 fn.) Shuy et al. (1968), on the construction of models of social and linguistic diversification: "The correlation of social status and linguistic performance first requires a careful delineation of each" and points out, (p.33), "for a linguistic study of any large community within a class society to be

in any way significant the class continuum must be objectively measured against the linguistic continuum, and vice versa."

Trudgill uses a more comprehensive social class index than Labov's (1966) one, taking into account six factors: occupation, income, education, housing, locality and father's occupation. Informants are stratified along this index into five social groups, based ultimately on the Registrar General's Classification of Occupations (1966 Sample Census).

These 5 social classes are established on the basis of a test variable, presence/absence of the localised grammatical feature, non-marked third person singular, (e.g. 'he come').

When the percentage frequency of this localised feature (non-marked verb form) was plotted against the class continuum for formal and casual styles, there was a quite clear-cut break point (a difference of 67% in frequency of usage of localised verb form) between the working-class groups and the middle-class groups, and other less well defined break points which Trudgill used in the division of the social continuum into five sections. These five classes Trudgill characterises in terms of typical occupational status, and in terms of rank in the familiar class hierarchy, ranging from middle middle class, (Class I) to Class V, lower working class, although, as he points out, occupation is only one of the six features contributing to the index.

Like Labov, Trudgill carefully structures the interview questions to elicit realisations of the variables under investigation. Trudgill takes sixteen phonological features into account, (three consonants, and thirteen vowels[FN]), through a series of speech styles ranging from reading styles (word lists, and continuous passages), and formal conversational style, to casual conversational style.

FN. The consonants are (h) (ng) (t), and the vowels are (a) (ā) (a:) (e) (er) (ɛr) (I) (ir) (o) (ou) (ō) (ū) (yu).

The last of these is claimed to be the closest approximation to 'spontaneous speech' (Labov, 1966, p.100), which it is possible to elicit in the interview situation. Trudgill splits the range of continuous phonetic variation for each vocalic phoneme into a practical number of discrete states along a spectrum ranging from standard to non-standard localised (Norwich) speech, and each informant's score for each variable represents a conflation of two parameters, degree of localisation, and frequency of localised variants.

Realisations of the variable (a), for example, as in 'bad', 'cap' etc., are divided into five easily perceptible phonetic 'types', which cover the total range of differentiation in Norwich English, thus:

(a) - 1 [æ]
(a) - 2 [æ:]
(a) - 3 [æ:ɛ]
(a) - 4 [ɛ:]
(a) - 5 [ɛ:e] (Trudgill: 1974, p.85f)

Each realisation of (a) - 1 scores 1, (a) - 2, scores 2, etc., and the total is divided by the number of instances of the variable to give an average value per informant per style. Thus exclusive realisation with variant (a) - 5 gives the maximum score 400, and exclusive use of the standard form, (a) - 1, gives the minimum score, zero. However, scores in between these extreme values represent a <u>combination</u> of two measures. The 2 measures which have been combined are the <u>degree</u> of localisation, and the <u>relative frequency of use</u> of localised variants. Thus it is impossible to distinguish, on the basis of their scores, e.g. an informant who consistently uses a moderately localised variant, (100% incidence of (a) - 3), and another informant who is not self consistent, but uses all variants in equal proportions. Both score 200 overall on this variable, (in the particular style under scrutiny). Whether such extreme individual variation within one style is likely or not, this hypothetical example demonstrates the blurring effect such a composite index can have on the intrinsic variability within the data. Because

individual variation within one speech style is levelled to an average score, which is itself a composite measure, it is impossible to abstract information on an individual's operational self-consistency within styles. Therefore the assumption that style is the single significant control factor in variation cannot be tested. (I.e. the derived scores are non-homologous across the population). By the same argument, it is also impossible to extract information on the incidence of variants at a given degree of localisation, without going back to the raw data from which the within-style scores were derived.

By conflating these two measures, then, the initial hypothesis that variation occurs at significant levels between, but not within, styles cannot be contradiced by the data, structured as it is. Thus a certain class of important outcomes are precluded from emerging from the data.

Selectivity and Atomism

Given that both social structure and linguistic variability within a community are complex, and involve many parameters of variation, any sociolinguistic model must be built on some simplifying assumptions. Ideally the least damaging assumptions (in terms of distortion of the data) should be found. It is just as important, however, that those assumptions should be borne in mind when conclusions are drawn from the results. Two strategies for reducing the complexity of social and linguistic classifications used in the investigations described above should be pointed out. The first is the policy of restrictive selection of variables, (linguistic, and social), the second is that approach which I call the 'atomistic' approach. Restrictive selection involves operating with a policy of studying only a small sub-set of available, (and potentially relevant) variables. Atomism involves looking at each of those variables in isolation, i.e. treating variables singly as if they behave independently, and their distributions do not interact. These kinds of experimental assumptions are perfectly valid, providing they are

made explicit, and are taken into account when interpretations are assigned to results.

Problems in Social Classification

Assessing the adequacy of a social classification is somehow more difficult than assessing a linguistic classification. Social class as a notion, and in reality, is amorphous. The sociolinguist is hampered by the lack of definition, (or rather the multiplicity of definitions) of the concept in the sociological literature, which itself results from the 'fuzzy'[FN] delineation of social and cultural sub-groups.

FN. Zadeh's theory of fuzzy sets (Zadeh, 1972, 1973a, 1973b) is an interesting attempt at providing a mathematic framework for representing 'humanistic'systems such as linguistic and social variability, which are characteristically "impervious to mathematical analysis and computer simulation" (Zadeh: 1973b, p.2).

The principle of restrictive selection of variables lies behind the abstract social class indices employed by most sociolinguistic researchers. Trudgill, stratifying his population on the basis of a derivative of the Registrar General's classification, and Labov, with his tripartite linear social index, make the assumption that the set of social variables incorporated, (six, and three respectively) are sufficient, and relevant social indicators to categorise their sample populations in a way suitable to their purposes. Obviously, an exhaustive social classification is a theoretical, as well as practical impossibility. Some classification theorists claim, with good reason, that no classification can exhaust the range of relevant[FN] dimensions of variation, since we sample from an infinite universe of 'potentially relevant' variables.

FN. Criteria for 'relevance', of course presuppose specific purposes.

Loevinger (1957) makes the stronger claim that 'content validity' of a

measurement space must be arbitrary, as our definition of 'universes' of variables, as well as selected sub-domains of these, must be ad hoc. She concludes that theory should be the prime generator of items for inclusion in any measurement space. For sociolinguistic classifications, the social and linguistic theories on which the model is based must determine the definition of dimensions of measurement. These theories must be adequate for the classification to be satisfactory.

The important issue is not whether all the social information has been included, but whether enough factors have been taken into account, and whether they are the most useful ones. It is impossible to know, in advance, which are the most useful ones, but we can formulate an operational definition of 'useful' in this context, as those social measures which, singly, or in groups, divide the sample population in a way which bears some at least partially systematic relationship to the linguistic diversity evidenced in that sample population. We do not know, a priori, which social variables we can afford to exclude. The non-relevance of all the social variables which are effectively (and tacitly) excluded by Labov and Trudgill has not been established.

Trudgill's social index includes the three factors Labov uses: occupation, education and income, and adds three more: housing, locality, and father's occupation. Esling, (1976), uses a similar index to that of Trudgill, but omits incomes. Reid, (1976), classifies Edinburgh schoolboys on one factor, father's occupation, and finds the attribution of social status on the strength of this single dimension unsatisfactory. Amongst those who use the methods developed by Labov, there exists no agreement on what the composition of a useful social index should be. Reid finds the set of methods which he calls the Labovian "research paradigm" inadequate in several ways, e.g. in "the minor place given to the study of motivation for variation," and the lack of attention given to "ways in which individuals "break" sociolinguistic rules to create social meanings " (Reid: 1976, p.16)[FN].

FN. See also Pellowe & Jones: 1978 (p.101ff.), on the intentional manipulation by speakers of their linguistic patterns in order to convey specific extra-linguistic information.

Douglas (1976), in her study of a Northern Irish rural community, finds that "linguistic variation and switching in Articlave are most easily explained in terms of informants' social aspirations," as in the case of "two informants, who are shop assistants but have the highest social ambition," who "show some of the most standard linguistic behaviour." (p.9) It seems that scalar measures of social class obscure an important part of the picture, viz. the effect an informant's self-image (as he thinks he is, or wishes he were), has on his language behaviour. Central to Labov's notion of 'style' is the hypothesised cause/effect relation between one kind of self-awareness, (monitoring of one's own speech, the Attention Principle, Labov (1972))[FN], and its effect on realisations of segments.

FN. "Styles can be ordered along a single dimension measured by the amount of attention paid to speech." (Labov, 1972, p.112).

in speech style towards the more prestigious variant realisation of a segment, (interpreted as the informant increasing his efforts to realise the target of the socially favoured variant), are claimed to correspond to increases in the level of attention which the speaker pays to his speech.

Several assumptions are being made here; they are questionable on the following grounds:

1. that all informants are aware of the same (unique) target;
2. that all informants are socially ambitious, and in the same direction; and that this (social ambition) is the only social psychological factor affecting speech behaviour in the interviews;
3. that style can be realistically treated uni-dimensionally;
4. that attention level is the single factor controlling style.

Informants may have different linguistic targets, whether on a prestige-stigmatised dimension, or otherwise. Data from the Tyneside Linguistic Survey shows that it is usually the case that one informant uses several qualitatively different phonetic realisations of the same speech sound (see below, Chap. 6). It is difficult to see how to assign relative prestige values to the range of localised variants used by one speaker. Moreover, there is no reason in principle to expect that all members of an urban population have the same linguistic variant as their prestige target, (since prestige itself is not a single simple dimension).

A feature which functions as a marker of social prestige for one group or subculture may be stigmatised by other groups. (I have heard the phrase "he wears a suit and tie" used by one speaker with approval, by another with extreme contempt).

Monod (1967) cites an example from his study of adolescent gangs in Paris, where 2 different sub-culture groups who, to an out-group member, are practically indistinguishable, are actually differentiated among themselves (amongst other things) by whether their hair is parted on the left or on the right. (Both groups have long hair). This example demonstrates at least two important facts about markers of social prestige:

i) features which are highly significant social markers for some sub-groups do not carry any signification for members of other groups. (Features which are prestigious for some, are neutral for others).

ii) social markers (in this case hairstyle) are discriminated to differential degrees of fineness by in-group, and out-group members. (To outgroup members, these two groups are long-haired: to in-group members, the side of the parting is a significant discriminator).[FN]

FN. I am indebted to Joan Beal for drawing my attention to this reference.

There is no reason to suppose that linguistic forms do not also have

differential prestige values for different sub-groups within a community.

Labov (1972, p.113) states, in his Principle of Formality, that: "any systematical observation of a speaker defines a formal context in which more than the minimum of attention is paid to speech."

Theoretically, by manipulating the degree of formality of the interaction, the interviewer creates differential degrees of linguistic self-awareness, and thereby elicits samples of speech ranging through contexts from casual to formal.

However, interaction phenomena other than this one may engender changes in speech style. Giles (1973b), indicates that such style changes may be "person based rather than context based" (p.88), and enlarges on this possibility with reference to Interpersonal Accommodation Theory (Giles, 1973a, 1973b). Changes in the interviewer's manner, designed to contral the formality of context, might cue the informant to respond by accommodating to what he interprets as the new rules of the interaction which the interviewer is signalling. (E.g. the informant may be responding to his own impression that he has offended the interviewer, or that the interviewer is 'warming' to him). A register switch may be triggered, which does not necessarily involve either a change in level of attention, or a consequent (according to Labov) movement along some more-vernacular to less-vernacular scale. (According to Labov, increase in vernacularness corresponds to decreasing attention being paid to speech).

In the Vernacular Principle, Labov (1972, p.112) asserts that "the style which is most regular in its structure and in its relation to the evolution of the language is the vernacular." This principle discounts the possibility of speakers having a range of equally natural styles, each of which is (normally) produced in the appropriate interactional circumstances. (Cf. Giles' (1973b) notion of "accent repertoire" (p.89)).

Labov, however, treats different styles as differential degrees of deviation from the norm of the individual's vernacular, along a linear scale

from natural to less natural styles.

The linguistic effects of having different social intentions (apart from 'correcting' one's speech) can then be safely ignored, and individual variability can be called "linguistic insecurity".

Smith (1976) argues for a sounder social psychological basis for sociolinguistic theory than has been evident to date. He cites Giles' Interpersonal Accommodation Theory as providing a more satisfactory framework to account for some strategies used by speakers to realise social intentions. This theory embodies the notions of "convergence" and "divergence", which refer respectively to style changes towards, or away from, the interlocutor's speech patterns. Interpersonal Accommodation Theory also accounts for the effectiveness of strategies used by a speaker, in terms of the hearer's evaluation of the extra-linguistic information signalled by the speaker, and how far those signals fulfil the hearer's expectations. In this approach it is the interaction, and not the speaker's behaviour in isolation, that is the prime focus.

Certainly many more intentions than can be subsumed under the heading of social ambitiousness are involved. Smith, referring to Giles' (1973b) notion of divergence, notes one social function fulfilled by maintaining, or increasing one's linguistic distance: "since speech style is for many groups an important clue to group membership, we can argue that divergence may be an important strategy for maintaining positive distictness in many circumstances." (p.31).

This may well be the case with the lowest working class strata in Labov's (1966) New York study. Labov, interprets this group's reluctance to shift towards middle class norms as lack of awareness of the prestige variants. Another possibility is that members of this group are aware of the prestige form, but fail to use it, either because they are not socially ambitious in the way Labov suggests other groups are, or they shun these forms as characteristic of a social class to which they are hostile, (or, at least, wish

to maintain their distinctness from).

An event which cannot be represented on a scale which has the vernacular at one extreme is that of hypercorrection towards, and beyond, the vernacular, and away from the prestige form. Divergence away from an interlocutor's 'superior' speech variety can be used to signal class (or personal) hostility. This shift may produce an exaggeratedly localised variety, which would presumably co-occur with increased attention level. By Labov's model, increase in attention co-varies with movement towards prestige forms; in this hypothetical but not unlikely case the reverse is true, demonstrating that attention to speech and prestige value are not simple covariates.

That such a uni-dimensional scale is an unsatisfactory artifact is highlighted by Smith's (1976, p.31) reference to: "convergence and divergence simultaneously along two or more descriptive dimensions that are differentially recognised by the ingroup and the outgroup." Two points are significant here; firstly that certain social and linguistic functions may be realised with, and recognised by, different token forms by different groups, (cf. my comments on the non-ubiquity of linguistic targets). Secondly, speakers do not move towards, or away from each other linguistically in any simple fashion at all. I have observed a Liverpudlian in conversation with a localised Tynesider shift towards a more localised Liverpudlian variety, which was as different from Tyneside speech as the less-localised variety used by the Liverpudlian at the start of the interaction. This could have been an instance of convergence along some abstract RP-to-undefined dimension, or divergence signalling identification with another region, or both at once. Moreover, while segmental phonology shifted towards Liverpool forms, non-segmental phonological patterns moved towards Tyneside systems (possibly favouring the interpretation of interactive convergence); thus in the same speaker we have divergent (in one sense) segmental phonology, and, simultaneously, convergent prosody. Nothing like this is accounted for in Labov's model.

The Labovian model does not incorporate any social attitudinal data.

Only one motivation factor, social ambition, is discussed, and it is not measured; it is merely assumed to be causal. Labov has not demonstrated social ambitiousness to be the central cause of variation, yet his methodology is based on assuming that it is. In addition, the factors on which the linear scale of styles is built are non-measurable. Labov (1972) states; "At present we can control some of the factors which cause attention to be paid to speech, ... but we have not yet quantified the actual behavioral feature: attention to or monitoring of speech." (p.112) It is, of course, possible to measure levels of physiological arousal in a subject, however, the application of a dynamometer and ECG machine to the informant might lessen the chances of eliciting his most relaxed vernacular style. If, however, by 'attention' Labov refers to some form of cognitive feedback loop, it is difficult to foresee how this feature could ever be measured. The notion of self-monitoring seems intuitively satisfactory: the problem is that it has not been demonstrated to co-vary with speech changes. Moreover, factors other than conscious, and variably successful, attempts to reach a self-imposed goal are operating.

Unfortunately, formality of context is also difficult to quantify absolutely. Herein lies a methodological paradox: stylistic variation is projected onto a linear scale calibrated by an unmeasurable parameter, attention, which is claimed to vary with formality of context. Labov has asserted earlier in the article dealing with principles of linguistic methodology, (Labov 1972), that hypotheses should be formulated in such a way as to be easily disproved. He also indicates that many empirical red herrings are consequent on "an initial misapprehension of the data." (p.104) It seems that Labov's theoretical standpoint has not adequately determined his methodological choices.

Trudgill acknowledges the complexity of social diversity in a community, and the fluidity and flexibility of "class boundaries and barriers" (Trudgill: 1974, p.32f). I have already mentioned that he establishes the validity of

his social stratifying principle by the use of a test variable, presence/absence of non-marked third person singular. We cannot assume that different variables in the same systems show linear correlation in their distribution patterns across a population, still less that informants will be ranked in the same way across variables from different systems, (here syntactic and segmental).

We must be suspicious, then of the practice of applying a social stratification which applied to one grammatical feature to distributions of phonological features.

Wolfram (1969) has shown that grammatical variables tend to show more sharp stratification along a social status index, whereas phonological variables tend to show "gradient stratification". "By sharp stratification is meant a quite definite break in the frequency at particular variants between contigious social classes in the sample; by gradient stratification is meant a progressive difference in the frequency of particular variants between the different social classes in the sample." (Wolfram: 1969, pp.120f.). For example, post-vocalic /r/ shows gradient stratification in Detroit Negro speech, whilst multiple negation shows sharp stratification. (This is a function of the type of variation, the former being in some sense discrete, the latter being continuous). Grammatical features, because of the discrete nature of their variants, may be easier to 'correct' in monitored speech. Garvey and Dickstein (1972) produced experimental results which indicate amongst other things, that a corpus of speech analysed at one linguistic level, (incidence of standard versus non-standard verb forms) produces different frequency distributions across a social index than when analysed at a different level, (e.g. lexical choice, and choice of predication type). Grammatical form (standard/non-standard verb-form), was found to vary significantly with the social measures sex, race, and socio-economic status (henceforth S.E.S.), whilst lexical choice varied only with SES. At the 'referential'level, (incidence of choice of predication type), variation was attributable only to the type of task performed in the experimental situation,

and failed to distinguish informants according to any social parameters.

Garvey and Dickstein point out: "In none of the studies cited ... has there been an examination of the effect of the linguistic level of analysis on the (potential) socially diagnostic significance of the findings" (p.376) and warn of "the need for distinguishing among levels of linguistic differences which interact with *situational factors on the one hand, and* with social status differences on the other. The findings demonstrate that SES and race differences in speech behaviour discovered at one level of linguistic analysis cannot be directly adduced as evidence that similar status differences exist at another level." (p.384)

It follows, then, that the social groupings established by Trudgill on the basis of the distribution of a grammatical feature do not necessarily comprise a valid division of the social continuum with which to correlate the distributions of phonological variables. Though he attempted to establish the validity of his social class index empirically, the level of analysis at which the selected variable operates may be irrelevant to the material of the survey, as the behaviour of his test variable is not likely to be paralleled by that of the experimental variables.

Garvey and Dickstein's thesis has wider general import for sociolinguistic survey design. In order to define "a coherent and systematic structure for the speech pattern of this neighbourhood" (Labov (1966) p.177), all levels of linguistic structure should be taken into consideration, not just the segmental phonological one. In practice a completely exhaustive linguistic analysis of the material may not be possible, but certain classes of variation other than segmental phonology can be incorporated to give a more comprehensive profile of speakers' varieties. (The TLS includes analysis at several levels of linguistic structure - see below, ch. 2).

The development of a more adequate social classification than has hitherto been in evidence in the discipline is a difficult task. Given that many more social factors than those used by the Labovian school may be relevant, e.g. attitudinal measures, and that it is impossible to reduce the

social profiles of speakers to a unidimensional scale without gross distortion of the data, then some kind of multi-dimensional model must be constructed. There are many problems here; which variables should be included, how many, and how should they be structured with respect to each other in a multidimensional scheme? The TLS methodology embodies a set of heuristic strategies for answering some of these questions empirically, and thereby converging on an optimal model of social variability in relation to language variation. A multi-dimensional model avoids the shortcomings of the selective and atomistic approaches, by initial inclusion of all potential diagnostic dimensions, (or as many as is practically possible), and by permitting scores for separate variables to be represented without being lost in an overall composite index; thus the effects of individual variables may be assessed, and non-co-variate ones eliminated. Also, dependency effects between variables in the social classification, (and in the linguistic one) can be empirically quantified. These matters are dealt with in full below chs. 5, 6, 7).

Problems in linguistic classification.

The Labovian model invokes the principle of restrictive selection in variable sampling in the linguistic as well as the social classification; firstly by selecting one sub-domain, (segmental phonological), secondly by taking into account a small sub-set of variables from this sub-domain. In the Martha's Vineyard study, Labov (1963) examines the distribution of one sound feature, and asserts he has found it "possible to assign a single social meaning to the linguistic feature in question " (Labov: 1963, pp.3f). This study is not designed, like the New York survey, to discover "a coherent and systematic structure for the speech pattern of this neighborhood" (Labov : 1966, p.1-7). In the Martha's Vineyard study, Labov is investigating one sound change in its social context. However, the principle of atomism operates in both studies, where linguistic and social features are studied

in isolation, and one-to-one correlations between linguistic and social factors are sought. Five phonological features are examined in the New York survey: (r) (oh) (eh) (th) (dh). Trudgill acknowledges that "the majority of segmental phonological elements in Norwich English are involved in variation of some social significance", (Trudgill: 1974, p.79), but claims that his selection of sixteen variables (three consonants and thirteen vowels), is valid, firstly because of a greater "amount of apparent social significance in the pronunciation of the segment or segments involved", and, secondly, on the grounds of "the amount of phonetic differentiation involved." (p.80).

Regarding the first criterion, the selection of variables rests on a subjective assessment by the investigator, of the relative degrees of social significance attached to all the variable elements in the segmental sound system. (Admittedly, this is the assessment of a linguistically trained native speaker). Moreover, the possibility of excluding many relevant parameters of linguistic variation remains. If, as Ringaard (1965) claims, "the transcriptions of phoneticians do not tell us so much about the speech of the areas they are studying as about the phoneticians themselves", it is also conceivable that the evaluation of relative social significance upon which a restricted selection of variables rests tells us more about the investigator's personal social frame than about the social meanings of variables for the population at large. This is statistically equivalent to assigning infinite weight to the results of a hearer judgment test applied to one informant, and zero weight to scores for the remainder of the sample. (Cf. also my remarks above (p.10) concerning variable targets and variable group affiliation).

Concerning the principle of restrictive selection; it is difficult to envisage how "the class continuum" can be "objectively measured against the linguistic continuum" (Trudgill: 1974, p.33) when large sections of the latter are excluded from the study, and the former is measured on an artificial linear scale of dubious validity.

The strategy of atomism is also used in the representation of the linguistic material. In the Labovian methodology informants are assigned to social strata, and the incidence of each variable is plotted for each stratum, independently of the other variables. Sociolinguistic generalisations are made from the distribution patterns of single features, on the basis of trends found to be shared by several variables. One consequence of the atomistic approach is that, by treating variables independently, a certain class of coutcomes is precluded from emerging, for instance dependencies between variables are not accounted for. Moreover, language features interact between, as well as within systems. Consider, for example, the differing phonetic consequences of word-stress position in lexemes which have alternative possibilities for placement of inherent stress, (e.g. Car'ibbean/Cari'bbean). The investigator may not be centrally concerned with these intra-and inter-systematic dependences, but the fact that they exist, and are not accounted for in the model may result in undetected interference effects.

Certain errors of interpretation may arise from combining the strategies of restrictive selection and atomism, if their consequences are not borne in mind. When several features show similar patterning the conclusion may be drawn that a general rule has been discovered, which applied to most members of the variable paradigm. The chances are that a sub-set of variables selected on intuitive grounds represent a more homogeneous class than the whole paradigm, (e.g. the similarity between Labov's (1966) variables (dh)/(th); and between (oh)/(eh).) Rarer patterns may be regarded as 'deviant', rather than providing evidence that different variables have non-homologous distribution patterns. Trudgill, out of sixteen sound features, finds only three which display what he calls the 'typical' pattern. Trudgill (1974, p.95) states: "The pattern of class, sex and style differentiation illustrated in the case of (ng) is the typical pattern associated with a normal linguistic variable " (my underlining). The use of 'typical' here is puzzling.

If indeed the cultural patterning within a speech community can be modelled meaningfully by use of a social stratifying principle, then an atomistic, and selective, approach can reveal information concerning the behaviour of single variables in relation to social class. However, general sociolinguistic deductions cannot be made from partial data. Trudgill, like Labov, formulates his aims (and claims), in broad, general terms: "to investigate the nature and extent of the correlation between and co-variation of linguistic and social parameters in the city of Norwich " (Trudgill: 1974, p.31). (Cf. Labov's aim of defining "a coherent and systematic structure for the speech pattern of this neighborhood." Labov , (1966), p.177).

In the light of these stated aims, we must consider the discrepancy between objectives and methods. Methodology delimits the range, and type, of discoveries which can be made; the scope of the data, and the analytic frame applied, place corresponding constraints on the sorts of deductions which it is possible to make. Assumptions which are implicit in the strategies which have been used must be taken into account at the stage of interpreting results. Both Labov and Trudgill seek to draw conclusions concerning the overall sociolinguistic configurations of their respective sample populations on the basis of a small number of variables. This can only validly be done if it can be demonstrated that these select few are statistically representative of the whole population of variables. Results from the TLS, presented later, show that this is an improbable hope where so few variables are concerned. Trudgill's own results demonstrate this; as thirteen of his variables show divergent distribution trends.

The initial dramatic findings and claims of sociolinguistic surveys of urban speech now need to undergo critical methodological assessment if the theoretical contribution of the subject is not to be seriously vitiated.

Methods are now available for the implementation of more refined, and theoretically adequate models of sociolinguistic variation, which avoid the shortcomings outlined above.

Recent advances by theorists of numerical taxonomy and of classification, stimulated by the availability of high speed electronic processors, have resulted in the development of numerical techniques for the handling of large copora of multivariate data. Multivariate methods (see Sokal & Sneath 1963) have been adapted for use in many disciplines, including: linguistics (e.g. Kroeber 1960; Needham 1967; Ross 1950); ecology (e.g. Anderson 1971); anthropology (e.g. Ihm 1965); archaeology (e.g. Hodson, Sneath & Doran 1966); social sciences (e.g. Ball 1965); biology (e.g. Williams & Lambert 1959; Williams & Dale 1965); psychology (e.g. Cattell & Coulter 1966) and psychiatry (e.g. Strauss, Bartko & Carpenter 1972).

Multivariate techniques are appropriate to apply to sociolinguistic data on two important counts.

Firstly, both linguistic, and social, differentiation involve variability along very many parameters. By using multivariate techniques, a sample can be analysed with respect to many social, and linguistic variables simultaneously. Thus the constraints imposed by both the approaches of selectivity and atomism can be avoided.

Secondly, sociolinguistic groups can be more usefully treated as "polythetic" classes than as "analytic" classes. The sociolinguistic investigations cited above (Labov: 1963; Labov : 1966; Trudgill: 1974) treated sociolinguistic groups as "analytic" classes (in the Aristotelian sense). Analytic classes are those which are defined by 'essences' or key features. These features are necessary and sufficient criteria for group membership. (This definition is also a reasonable gloss for "monothetic" groups).

Neither organisms, nor language, nor social behaviour can be successfully classified according to an 'analytic' scheme. If this kind of frame is adopted, many anomalies arise. For example, in biological taxonomy, if the presence of red blood corpuscles is set up as a definitive characteristic feature of vertebrates, then the classification founders

when cases such as that cited by Ruud (1954) are encountered. Ruud gives the example of certain species of fish, which do not have red blood corpuscles, yet we would want to classify them as vertebrates. Yet presence of red blood corpuscles is certainly a characteristic of most vertebrates, and generally uncharacteristic of non-vertebrates. Syllogistic proofs do not apply: the strongest prediction which can be made is a probabilistic one. The same situation holds for many characteristics of biologically related groups (See Sokal & Sneath: 1963).

Such characteristics are paralleled in the patternings of linguistic behaviour.[FN]

FN. Zadeh (1973b) demonstrates the probabilistic nature of linguistic classes in terms of his fuzzy set theory.

This is shown by the need to provide a theory which accounts for a degree of non-determinacy in usage of variant forms, such a system is formulated by Labov in terms of variable rules (Labov: 1969) and by Bickerton in terms of implicational scales (Bickerton: 1975).

A more useful classificatory strategy than the Aristotelian one is due to Wittgenstein. The concept of 'taxonomic affinity' developed by biological taxonomists (see Sokal & Sneath: 1963) has strong affinities with Wittgenstein's notion of 'family resemblance." His thesis is that certain classes of entities are related by shared features, but that none of those shared features is necessarily universally possessed by all members of the group. Thus there are no necessary attributes for testing for group membership. Moreover, characteristic features of a group may not be exclusive to that group. Therefore typical features of a group are neither necessary nor sufficient criteria for group membership. Such groups are termed 'polythetic' classes.

The list of shared features might vary quite radically across different pairs within the same group. The internal structure of a group is

determined by the number, and range, of attributes shared by its members.

By the use of multivariate techniques we can discover polythetic classes based on a large range of social and linguistic variables. These techniques, then, provide a useful set of strategies for eliciting the complex (and very probably) polythetic relationships within the fluid and flexible social structure of the community (see quote from Trudgill above, p.15 and within linguistic variability.

The T.L.S. exploits multivariate strategies in order to search for natural polythetic social and linguistic groupings within the sample population. Thus, no variables are predicted by the model as being key, or defining characteristics of groups. Rather, the natural groups emerge from the classification process. (Cf. e.g. Labov's approach, where definitive characteristics, such as salient linguistic markers, and social indicators are selected in advance, so that the groupings which are discovered are determined by this selection).

The T.L.S. methods do not preclude the possibility of monothetic classes emerging; if sociolinguistic groups are discriminated absolutely by a short-list of definitive social and linguistic features, then this will emerge from the classification.

By avoiding certain pre-empting assumptions which have characterised previous sociolinguistic research, we can a) test those assumptions; and b) approach more closely the goal of coping with the "structured heterogeneity" of social interaction and linguistic variability.

CHAPTER 2

THE METHODOLOGY AND MODEL OF THE TYNESIDE LINGUISTIC SURVEY

To date there has been no adequate attempt to model the overall linguistic variability of an urban community, and given the total ignorance of the mathematical properties underlying language variation, the Tyneside Linguistic Survey was conceived as a set of methods whereby the salient features of linguistic and social diversity are empirically determined by, rather than presupposed in, the model. The model is designed to generate different sets of possible results, by processing the data in different ways, and the model itself is subject to modification in the light of these results. To borrow a term from mathematics, the model is a 'machine' for generating hypotheses, in that it is capable of generating a number of possible solutions to the problems under investigation, all of which must be regarded as valid representations of the data, but which will vary in their degree of utility with respect to the purposes of the Survey. The 'usefulness' of the different sets of results so generated cannot be predicted in advance, for the process must be a cyclic one of self-evaluation: the design of the model must be constrained by its own products as these progressively illuminate the nature and structure of the raw data. In this sense (and in other ways, see below, pp.18ff), the model is 'dynamic'.

The aims, in general, of the Survey are as follows:

1. to identify, and exhaustively characterise, the varieties of speech which co-occur in the area under consideration, (initially the Tyneside conurbation),

2. to determine the distribution of both the speech varieties and their constituent elements across the relevant social sub-groups (which must also be empirically discovered by the model),

3. to extend the model to cope with a wider geographical compass;

   a. by successive inclusion of more of the conurbation ,

   b. by including neighbouring conurbations,

c. by eventually adapting the model to account for other urban varieties of English,

4. to extend the investigation onto a diachronic basis, so that changes through time in these distributions, (see 2.), may be measured.[FN]

FN. The discussion in this chapter is based on the following T.L.S. publications:
Pellowe: 1970a; 1970b; 1970c; 1970d; 1973; 1976;
Pellowe, Nixon & McNeany: 1972b; 1972a;
Pellowe, Nixon, Strang & McNeany: 1972
(henceforth Pellowe et al: 1972);
in particular, the latter two .

The T.L.S. represents an effort to avoid several methodological pitfalls mentioned in chapter 1. These include unrepresentativeness in selection of variables, and the problems inherent in an atomistic approach. The T.L.S. methodology is also designed to avoid the kind of reduction and misrepresentation of the social fabric of an urban community which results from the attempt to express social differentiation by means of a pre-conceived linear social index.

Sampling of informants

The sample must be large enough to adequately represent all speech varieties occurring on Tyneside, in other words, for the entire spectrum of (local) linguistic differentiation to be captured. For each type to be represented proportionately, the sample must be random. However, it was anticipated that a sociolinguistically influential and interesting group of speakers would be very sparsely represented, if at all, in a random sample, due to their relative rarity across the population, namely speakers of 'non-localised' (NL) varieties, that is those speakers, who are typically middle-class, well-educated and in high income groups, whose speech gives no indication of their geographic origin. With this fact in mind, the original sample was drawn as follows:

Phase 1 sample:

a. a handpicked sub-sample of 40 speakers known to have NL varieties;

b. 60 speakers, resident in a street intuitively judged to be 'middle-class';

c. 150 speakers chosen from the Electoral Register, by wards and polling districts. The selection was normalised according to size of ward or polling district, thus giving every member of the population an equal and calculable probability of selection.

Phase 2 sample.

To test the reliability of this sampling programme a second sample was drawn from the base population. By this means it is possible to ascertain whether the original sample succeeded in exhausting the number of speech types occurring in the total population, and if they are represented proportionally. Also, the estimates, based on the first sample, of the population frequencies for each type, and the estimate of the number of new types likely to be revealed by a new sample, can be verified (Using Good's (1953) technique). Moreover, the randomness of the source list, (Electoral Register) with respect to the study was tested by drawing a sample by a different method, namely by imposing on the area "a grid of intersections whose scale at any point is an inverse function of the population density there" (Pellowe et al: 1972).

Phase 3 sample.

As a further test of reliability of informant sampling methods, a stratified sample of 150 informants was drawn from the conurbation south of the Tyne, (Gateshead), the stratifying factor being 'rateable value per dwelling by polling district'. Furthermore, this sample provides information on the exhaustiveness or otherwise of the contents of the classificatory scheme in terms of inclusiveness of variables and of types of speech varieties. This sample also represents a latitudinal extension of the survey, whereby linguistic distinctions as a function of areal differences, may be measured.

Representativeness in the selection of linguistic variables.

As Trudgill (1974, p.79) points out, most segmental features displayed variability across the speech community which he is studying:

"the majority of segmental phonological elements in Norwich English are involved in variation of some social significance."

This comment has general relevance to any study of sociolinguistic variation. Trudgill, however, restricts his selection of linguistic variables to sixteen phonemes, whereas the T.L.S. operates on the principle of exhaustive inclusion of variables. Thus, the biassing effects of an approach based on a restrictive selection of variables (see above, pp.7ff.), is avoided. No assumptions are made, a priori, concerning the relative sociolinguistic significance of different variables.

One of the most distinctive characteristics of the T.L.S., then, is that the model is designed to exhaustively characterise all measurable dimensions of language variation in an urban community. Although this is an ambitious goal, when it is achieved there will exist a corpus of data which comprehensively characterises the patterns of linguistic variability in the Tyneside speech community. By minimising the loss of information, the model allows for a wide range of hypotheses to be tested, and by avoiding pre-empting assumptions, such as which variables are relevant, it also allows for non-predetermined results to emerge.

For example, Pellowe et al. (1972, p.9) claim:

"our model will not only generate Labov's variables analytically, but will indicate the extent to which, and reasons why, other parts of the linguistic structure are candidates from this role," (i.e. the role of sociolinguistically salient variables).

It is of vital importance that the data is not represented in such a way that the results are pre-determined by the strictures imposed by the model.[FN]

FN. An over-simplified model at best limits the range of possible outcomes. At worst, it predetermines the outcome by generating results which are a direct projection of the simplifying assumptions made. For example, Pellowe 1967; attempted, in a pilot run, to quantify the correlation between non-working class status (defined by membership of the Registrar General's social classes I-III) and NL speech. Pellowe et al (1972, p.3f) report that this framework was "insufficiently sensitive", and, moreover, that "such samples could not provide an adequate classificatory base upon which to identify the V's," (varieties).
"...the varieties which one identified were rather directly dependent upon the variables which one had chosen in order to identify these varieties (in other words the Vs which were classified were direct projections of the gross sociolinguistic perceptions of the analyst as an ordinary hearer)."

We cannot predict, in advance, which linguistic, and social, features discriminate sociolinguistic sub-groups.

In the urban situation population density is high, and social role structure is complex and multidimensional, and characterised by loose, symbolically mediated social bonding (Goffman: 1961, 1963; Pahl: 1968). A linear social index cannot adequately represent this social differentiation, and it is unlikely that the relationships between configurations of linguistic and social variables are capable of expression in terms of simple functions or one-to-one relationships. (See my comments on selectivity and atomism, above, ch.1, p.7 ff.)

Hence the need for a model which exhaustively characterises social and linguistic differentiation, and which generates groupings on the basis of complexes of features from both domains.

Thus we can determine empirically which variables, or variable complexes (both linguistic and social) characterise sub-groups of Tyneside speakers.

The data, which is collected in the form of tape recorded interviews with members of the sample, is analysed, and coded exhaustively in terms of all measurable variables from the linguistic systems of segmental phonology, syntax, intonation and pitch range. Also, various paralinguistic and collocational variables are coded. (For a description of the linguistic coding frame, see below, pp. 37ff., and Appendix A).

A wide range of x-linguistic (extra-linguistic or social) information was requested in the interviews, which provides a comprehensive social profile of each informant with respect to a variety of sociological variables (analytic information), also attitudinal and other kinds of data are elicited. (For a full description of the social coding frame see below, P. 45ff. and Appendix B).

Speech varieties, and their defining linguistic characteristics, are then identified by multi-variate methods. If each linguistic variable is conceived as an axis in a multi-dimensionsal linguistic variety space, then each informant's scores on all linguistic variables place him in a unique multi-coordinate locus in that space. This locus defines that informant's linguistic profile with respect to the sum total of dimensions of the space. Groups of speakers with similar linguistic profiles (with respect to the definition of the space) are evidenced by swarms (or clusters) of informants (sample points) occupying certain areas of the space. (See ch. 6 for a description of the derivations of linguistic variety clusters).

The same informants are then dispersed through the social space (SocSp), the dimensions of which are defined by the sum total of social variables which are included in the social coding frame. Each informant's social profile (with respect to all the social variables) determines his position in SocSp. Clusters of sample points in SocSp represent social groupings of the sample. (See ch. 5 for a descritpion of the derivation of social clusters).

From both the linguistic, and social, spaces, we can then extract those dimensions (variables) which participate most actively in binding these groupings together. Those dimensions, if found, will indicate variables which are <u>key diagnostics</u> for the groupings obtained. These are variables, then, whose <u>salience</u> is <u>empirically</u> demonstrated.

The model, then, generates linguistic variety clusters whose internal structure is dependent not on a selected sub-set of variables, but on an

exhaustive linguistic analysis of the speech output of members of the sample. It also generates social groupings which are based on a large number of social variables (which include those sociological measures used in previous sociolinguistic surveys - see citations in ch. 1.

Therefore the T.L.S. aim of exhaustive characterisation of varieties is fulfilled, and so is the aim of discovering the social sub-groups relevant to the population under investigation.

Moreover, the adequacy, or otherwise, of a restrictively selected sub-set of variables (linguistic and social) can be tested.

Firstly, if a sub-set of key-diagnostic variables does not emerge, then we can say that a classification based on any sub-set alone would not recover the relevant groupings of the sample.

Secondly, we can analyse the clusters obtained with respect to, e.g. the classic social indices (socio-economic status, (SES), social class (SC), occupation groups) and determine whether clusters are discriminated by these indices.

(This is done in ch. 5 for social clusters, and ch. 7 for linguistic clusters).

The second aim of the T.L.S. (to determine the distribution of both the speech varieties and their constituent elements across the relevant social sub-groups) is achieved by examining the relationships between linguistic clusters (and their key diagnostics), and the social clusters. Providing reliable linguistic diagnostics emerge, it will be possible to quantify a function to express the mapping from linguistic diagnostics to social clusters. Thus the role played by single linguistic features (or complexes of features) in marking social groups can be empirically determined, and their predictive power (i.e. the level of probability attached to these variables in allocating individuals to social groups) can be measured.

The results may or may not support, e.g. Labov's selection of salient sociolinguistic discriminators, or, it may be demonstrated that single

linguistic features cannot be reliably correlated with social groups. As both social and linguistic diversity are treated by the T.L.S. model as complex and multidimensional, it is probable that simple correspondences between linguistic varieties and social groups will not emerge. Only in the unlikely case of variety cluster membership being identical with social cluster membership will linguistic diagnostics bear a straightforward relationship to social groupings, and have simple predictive values. It is probable that the sample will be distributed differently across the two spaces, and members of each social cluster will be dispersed across several variety clusters, and vice versa. The mapping function of one linguistic diagnostic to one social cluster, then, will be a complex function of the number of variety clusters represented in the social cluster, the fraction of the social cluster represented by each variety cluster, and the proportion of each variety cluster which appears in the social cluster, as well as the linguistic variable's diagnostic value (if it is less than absolute) in predicting membership of the variety cluster of which it is a diagnostic.

The multi-variate technique which is used to classify informants into groups is known as 'cluster analysis'. The term covers a number of techniques for analysing multi-variate data

"which attempt to solve the following problem:

Given a sample of N objects or individuals, each of which is measured on each of p variables, devise a classification scheme for grouping the objects into g classes. The number of classes and the characteristics of the classes to be determined." (Everitt: 1974, p.1).

We want to determine the number of social, and linguistic groups in the sample, and determine the characteristics of these groupings.

"The techniques of cluster analysis...are useful tools for data analysis in several different situations. They may be used to search for natural groupings in the data, to simplify the discription of a large set of multivariate data, to generate hypotheses to be tested on future samples..."

(Everitt: 1974, p.5).

Each of these applications is of interest: we want to discover the natural sociolinguistic groupings within the sample, to simplify the description of a large corpus of sociolinguistic data, and to generate hypotheses concerning interactions between social, and linguistic, variables.

The clustering process used here (hierarchic fusion) involves the following steps:

1. each informant's scores for all variables in the coding frame (social and linguistic) are input to the clustering program;
2. all possible pairs in the sample are compared, with respect to each variable;
3. a measure of mutual similarity (or distance) is computed between each pair, taking into account scores for all variables;
4. the construction of a similarity matrix from 3;
5. on the basis of the similarity matrix, clusters of similar individuals are built. (See fuller description, ch.4).

5. is achieved by a series of scans through the similarity matrix. On each scan, the pair with the highest level of similarity i.e. with the most similar profiles, are fused, and thereafter treated as one individual. After each fusion, the similarity matrix is shrunk by one row and column, and all the paired similarity measures are recomputed.

The process of scanning and fusion is repeated until a pre-set similarity threshold (or number of clusters) is reached, otherwise the process continues until there is one cluster containing all the individuals. The classification can be examined at different stages (e.g. when there are 4, 3 and 2 clusters respectively) and cluster diagnostics can be printed for all variables, across the clusters which exist at those selected points in the fusion process. Thus, the sample can be examined at any level of internal structure.

Fig. 1 shows a (hypothetical) dendrogram (a diagram which summarises the steps in the fusion process). This diagram shows that, in a hierarchical

clustering of 10 informants there are different numbers of groupings at different stages in the process, which can be conceived of as a classificatory hierarchy, (groups contain sub-groups). The crosses represent individuals, the tree shows the fusion steps which occur at decreasing levels of similarity (from bottom to top). The broken lines show three levels at which the classification could be examined. The number of branches of the tree crossed by the broken line indicates the number of clusters present at that stage of the process. Thus at level a, there are 2 clusters, at b, 3, and at c, 5. Level c represents a finer classification into sub-groups than b, and b than a.

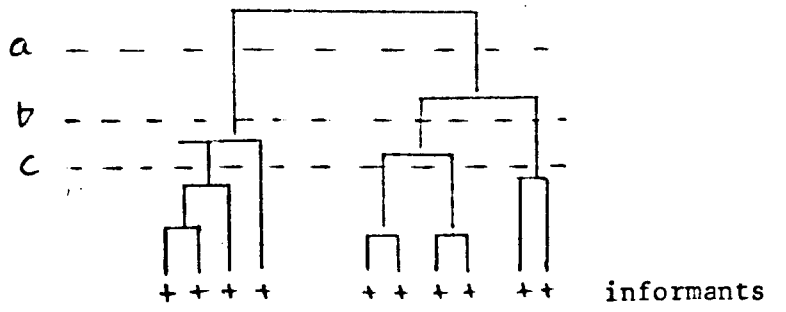


Fig. 1 Hypothetical dendrogram.

So, we can obtain information on linguistic and social groups at any desired level of fineness of the classification.

A dynamic model.

This is one way in which the model generates multiple solutions. Alternative solutions can be viewed as alternative realities: they can also be seen as competing realities, in the sense that the preferred solutions will be those which optimally structure the data in respect of the purposes of the classification. This is one form of dynamism. A related form of dynamism is a consequence of the range of clustering methods available. There are many different similarity coefficients, and clustering algorithms, which have different mathematical properties. Each elicits a different set

of structural properties from the data. By systematically varying the mathematical measures applied, a series of different classifications can be obtained from the same data. Thus we can converge on an optimal structuring of the data in terms of the naturalness of the classification. To return to the mechanistic metaphor, if the T.L.S. model is a machine for generating hypotheses, it is also a servo-mechanism,incorporating, as it does, self-evaluation procedures in terms of internal and external feedback. On the basis of this feedback, ongoing modifications to the model are implemented. External evaluation relies on hearer judgments concerning the groupability of speakers on the basis of naive perceptions of overall output by speakers.

Internal evaluation is achieved by comparing the classifications obtained under different circumstances:

e.g. by

1. partitioning the data according to level of linguistic analysis;
2. masking out (effectively obliterating) different groups of informants;
3. using different similarity measures;
4. using different clustering algorithms;
5. varying the method of extracting cluster diagnostics;
6. examining different stages of the fusion process for a given classification.

6. has been discussed above (pp.33-34).

1. We can section off the analyses pertaining to different levels of linguistic structure (e.g. segmental phonology, syntax, prosodics), and generate independent classifications at these different levels. An analysis of the differential distributional patterns of the sample across clusterings thus produced will complement and extend Garvey and Dickstein's (1972) study. (This paper demonstrates the dependency of sociological correlates of language variation on the linguistic level of analysis). These findings will be tested over a wider range of variables. (Garvey and Dickstein examined the behaviour of only one grammatical construction analysed at the levels of

grammatical form, lexical choice and choice of predication type).

2. We can mask out different groups of speakers. For example, by masking NL speakers, we can achieve a finer reticulation of the L speakers in the sample.

3, 4, 5. By varying similarity coefficients, clustering algorithms, and the method of extracting cluster diagnostics, we can discover an optimal set of mathematical measures. These will a) produce an optimal classification, and b) tell us something about the mathematical patterning underlying sociolinguistic variation.

Evaluation of these alternative classifications will take into consideration

a) the stability of clusters (when the classification reaches a state of equilibrium, i.e. when additional data does not radically effect the cluster patterning, it is reasonable to assume that the classification is tending towards an optimal representation;

b) investigator intuition based on knowledge of the sociolinguistic makeup of the community.

A different sense in which the model is dynamic involves the incorporation of 'new' variants. As, and when, a variant realisation occurs in the data for which no coding category exists, a new coding category can be created for that variant, and incorporated into the coding frame by an additive process. Brennan(1972p.31) refers to this procedure as "'sequential' development of attribute sets."

Several new linguistic variants have, in fact, been identified during analysis of the tape recordings. (Two examplesare given in ch.4). The inclusion of these 'new' variants caused some (trivial) computational problems, which are deal with in ch. 4.

When optimal social, and linguistic classifications of the sample are obtained, and the correspondences between the two are determined, Pellowe et al (1972) claim that the process whereby hearers derive x-linguistic

information (i.e. information extraneous to the overt message content) from speech will be modelled. Hearers draw many kinds of inferences from linguistic input concerning, for instance, the speaker's geographical origins, social class and/or social class allegiances, personality, attitude to the situation of the utterance, etc.

The hearer is likely to be more aware of features which differ greatly from his own realisation norms, and frame his perceptions in terms of 'distance' from himself. He may perceive speaker (a) as more different (further away) from himself, than speaker (b).

Hence the utility of the spatial metaphor incorporated into the model[FN], which accounts for differences between speakers in terms of distance and orientation.

FN. And the appropriateness of the Euclidean distance metric - see ch. 4 p.94 for an account of this distance coefficient.

The mapping function from linguistic to social space provides an analytic analogue to the process of derivation of x-linguistic information by hearers.

Coding of linguistic variables.

There are problems in formulating a framework to represent grammatical variation: this is an area of linguistic variability which is less well developed in the T.L.S. coding frame at present.

Lexical variability presents problems too: in order to ensure comparability between informants, it is essential that all variables have a good chance of being elicited during the interview. Unless word lists or reading passages[FN] are included in the interview regime, the lexical content of the interview cannot be controlled, and therefore we cannot be certain that all lexical items of interest will occur in all interviews. Lexical variables included, then, are few in number, and frequent in occurrence in interactive situations, e.g. 'yes'.

FN. The T.L.S. interviews are as informal as possible: the use of reading passages and word lists might introduce constraints on informants' speech behaviour according to global parameters (context, formality, register) which are not of central interest here.

The area of lexical variability, then is underepresented in the T.L.S. coding frame.

The T.L.S. coding frame.

1. The Linguistic coding frame.

A full account of the linguistic variables is given in Pellowe, Nixon & McNeany (1972a). As originally defined, the variables fall under the following 9 categories: (Pellowe et al: 1972a, p.21).

"(a) paralinguistic & prosodic (sensu Crystal & Quirk 1964) : 58

(b) vowel - stressed : 68

- environment /Vr/ : 22

- weak forms of stressed syllables : 8

- forms always unstressed : 7

(c) consonant : 45

(d) miscellaneous properties of syllable and word in continuous speech : 33

(e) grammatical complexity : 36

(f) fluency (hesitation phenomena etc.) : 9

(g) localised lexis (recognition & usage) : 2

(h) localised syntax (acceptability & usage) : 14

(i) lexical 'resource' : 1".

(b), (c) and (d) are hierarchical qualitative multistate variables. The others are quantitative variables.

An example of a quantitative variable (Pellowe, Nixon & McNeany: 1972a, p.21), is:

"% tone units wholly or partially marked by the paralinguistic feature of huskiness (Crystal & Quirk 1964; Crystal 1969)"

The nature of hierarchical qualitative multistate variables is explained below, under segmental phonological variables.

Segmental phonological variables.

The segmental phonological variables (categories (b) and (c)) are arranged as 3, or 2, level hierarchies. These hierarchies require some explanation.

The superordinate level is the overall unit (OU), defined as "an abstract phonological symbol which encapsulates the complete lexical set in which it occurs." (Pellowe, Nixon & McNeany: 1972a, p.2).

That is, there is a lexical set subsumed, for example, by the phonological entity i:, (week, treat, seed, eel...), and another by the phonological entity eI, (pay, raid, delay, make...).

All members of the lexical set subsumed by i:, then, will have the realisation of the segment 'i:' coded under OU i:, whatever phonetic form that realisation takes. The form that the realisation takes is coded according to the subordinate levels of coding structure. These are the PDV (putative diasystemic variant) and the state levels.

A state is a "symbol representing a phonetic realisation which is auditorily discriminable from all other states" "and a PDV is defined as "a class of states which is sociolinguistically discriminable as a class from all other such classes... within a particular overall unit" (Pellowe, Nixon & McNeany, (1972a), p.2).

Fig. 2 shows the organisation of OU i: in terms of PDVs and states.

By this scheme, the vowel in 'treat', for example, will always be coded under OU i: regardless of its particular phonetic realisation. If that vowel is realised as [ɛ] , (/trɛt/ is a fairly common localised realisation for this lexeme), the vowel will be coded under the third PDV of this OU, PDV/ɛ/ , and the second state of that PDV, state [ɛ]

The OU is set up "to normalise the distribution of systemic variants

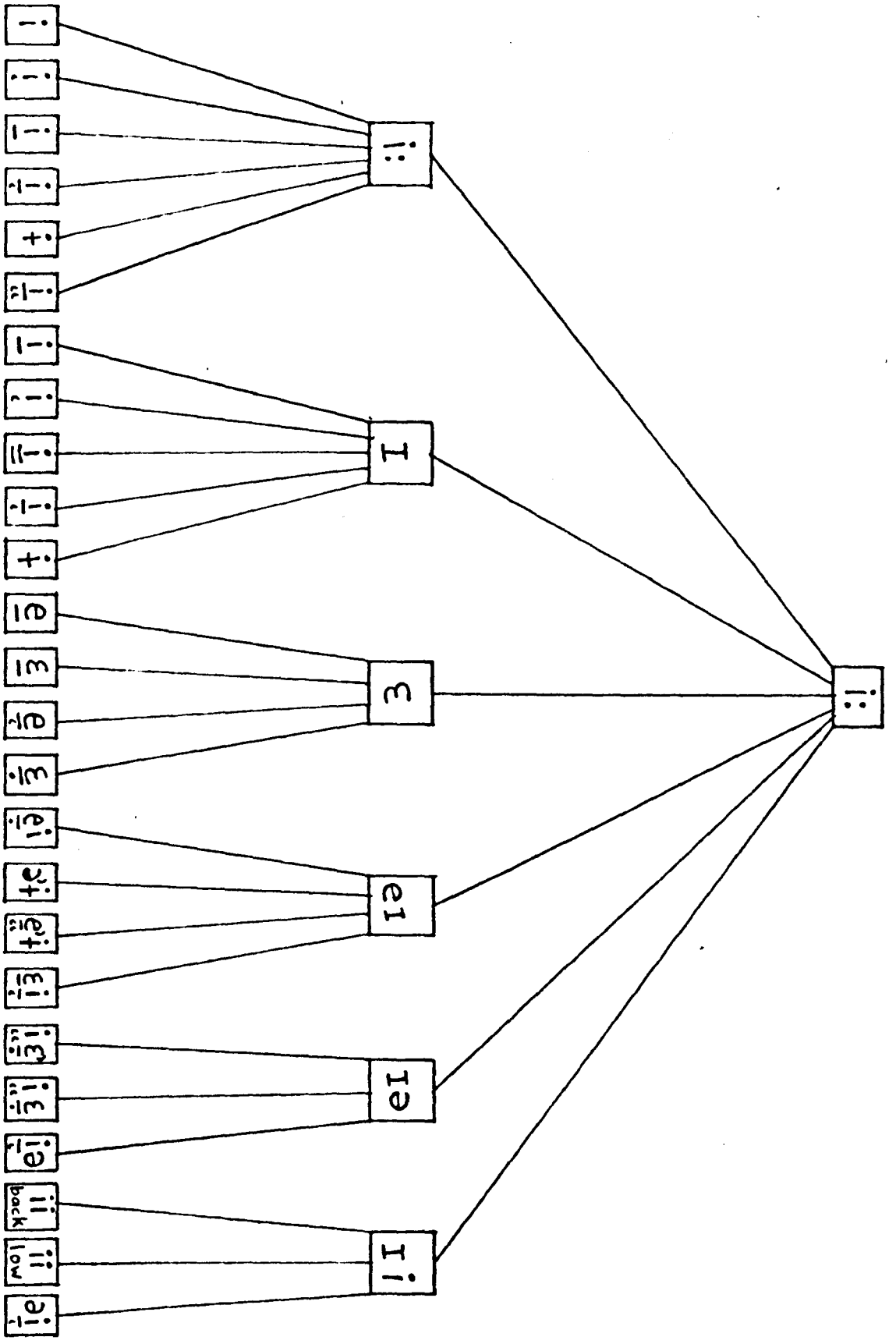


Fig. 2.
Breakdown of OU1 into PDVs and states.

without characterising them (in the VSp[FN]) as 'deviants' from centrality... Regardless of what variant value a segment in a word (or syllable) carries, we know the lexical set membership of the word as a word, in terms of the set of OUs. It is important to stress that the OU is a categorial label whose function is to ensure an undistorted comparability: it contributes no values to the calssification." (Pellowe et al: 1972, pp.23f.)

FN. VSp = Variety Space.

Fig. 3 shows an example of a consonantal OU, r. This OU does not have the PDV level: all variants are coded at state level only. This is true of some of the consonantal OUs. OU i: is an example of a 3-level coding structure, OU r is an example of a 2-level coding structure.

Appendix (A) gives a full specification of the 51 OUs, with their subordinate PDVs and states (derived from Pellowe, Nixon & McNeany: 1972a).

Category (b) variables cover the following OUs:

i: I ɛ æ ɑ ɒ ɔ: ʌ ʊ u eɪ

əʊ aɪ aɪə aʊ ɔɪ ɜ ɪə ɛə ʊə

and four types of the reduced vowel schwa (in different phonetic environments - see Pellowe, Nixon and McNeany: 1972, p.18ff), and two types of a reduced form of the vowel I.

Category (c) variables (consonants) are coded under an OU scheme, but the PDV level is not always applicable. When it is, it is differently defined than for vowels.

Category (c) covers the following OUs:

p b t d k g tʃ dʒ f v θ ð s z ʃ ʒ h m n ŋ l

r j w, also realisations of ŋ in bound morphemes (-ing). OU's p, b, t, d, k and g have, at PDV level, the distinctions syllable initial/ medial/final. (See Appendix (A)).

Each speech segment elicited in the interviews is coded according to the OU/PDV/state scheme. Numeric codes are used, each PDV is designated by

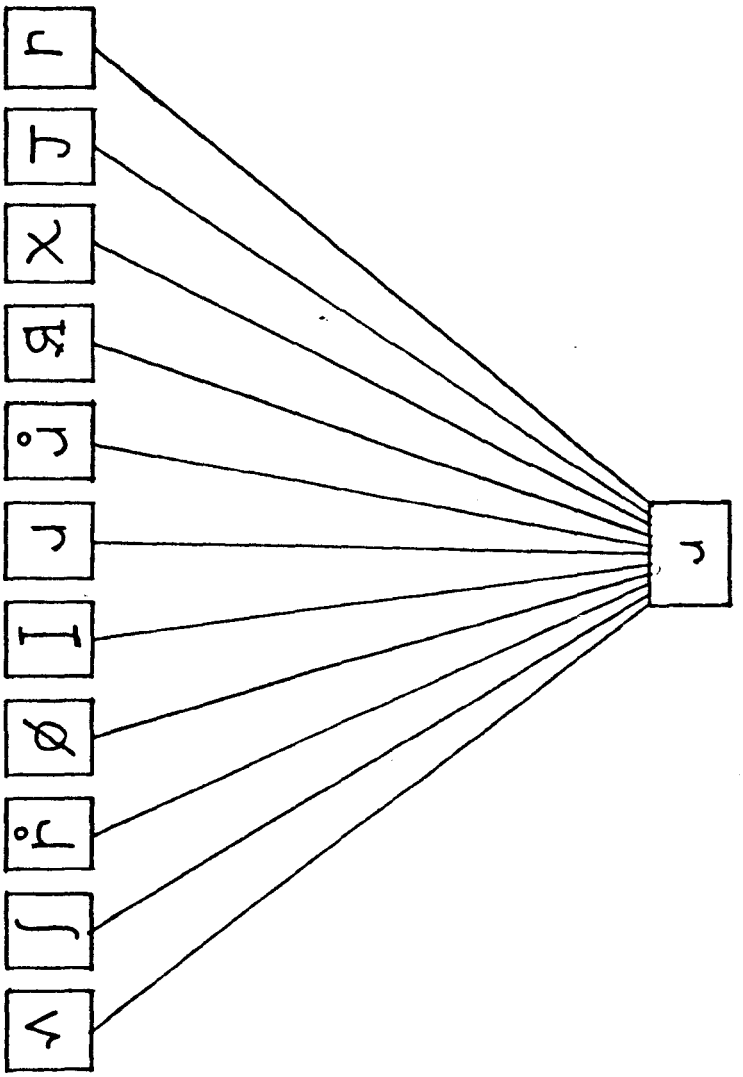


Fig. 3

A consonantal OU, (with no PDV level of structure.)

a 4-digit code, (PDV codes are even-numbered from 0002 upwards) and the state with which the segment is realised is designated by a fifth digit. Thus all instances from the lexical set subsumed by i: (the first OU), which are realised by the state i (the first state of the first PDV of OU i:) are coded 00021. (See first page of Appendix (A).)

Co-occurence phenomena.

Category (a), prosodic data (intonation and pitch range features) are coded by 3-place alphabetic codes. The first character represents the tone type, the third represents the pitch range feature, and the second represents the grammatical form class on which the tone and pitch range feature fall. (Zero can be coded for each of these 3 classes of feature). This data is known as the 3-alpha data. Any combination of features from these three systems can be recovered (e.g. falling tone, preceded by booster, on common noun). (See Pellowe and Jones: 1978).

Other linguistic variables.

Categories (d) through (i) are coded by 2-place alphabetic codes. These variables are not dealt with here. (See Pellowe, Nixon & McNeany: 1972a for a full description of these variables).

Fig. 4 illustrates the structure of the T.L.S. coding frame. Each informant's sociolinguistic profile can be thought of as his scores across all variables as depicted in this structure. The 2-alpha data is a list of values: the 3-alpha data consists of a 3-way co-occurrence table, and the segmental phonological data is structured according to the OU hierarchy. In addition, the social profile of the informants is coded according to the social coding frame. This coding frame is now described.

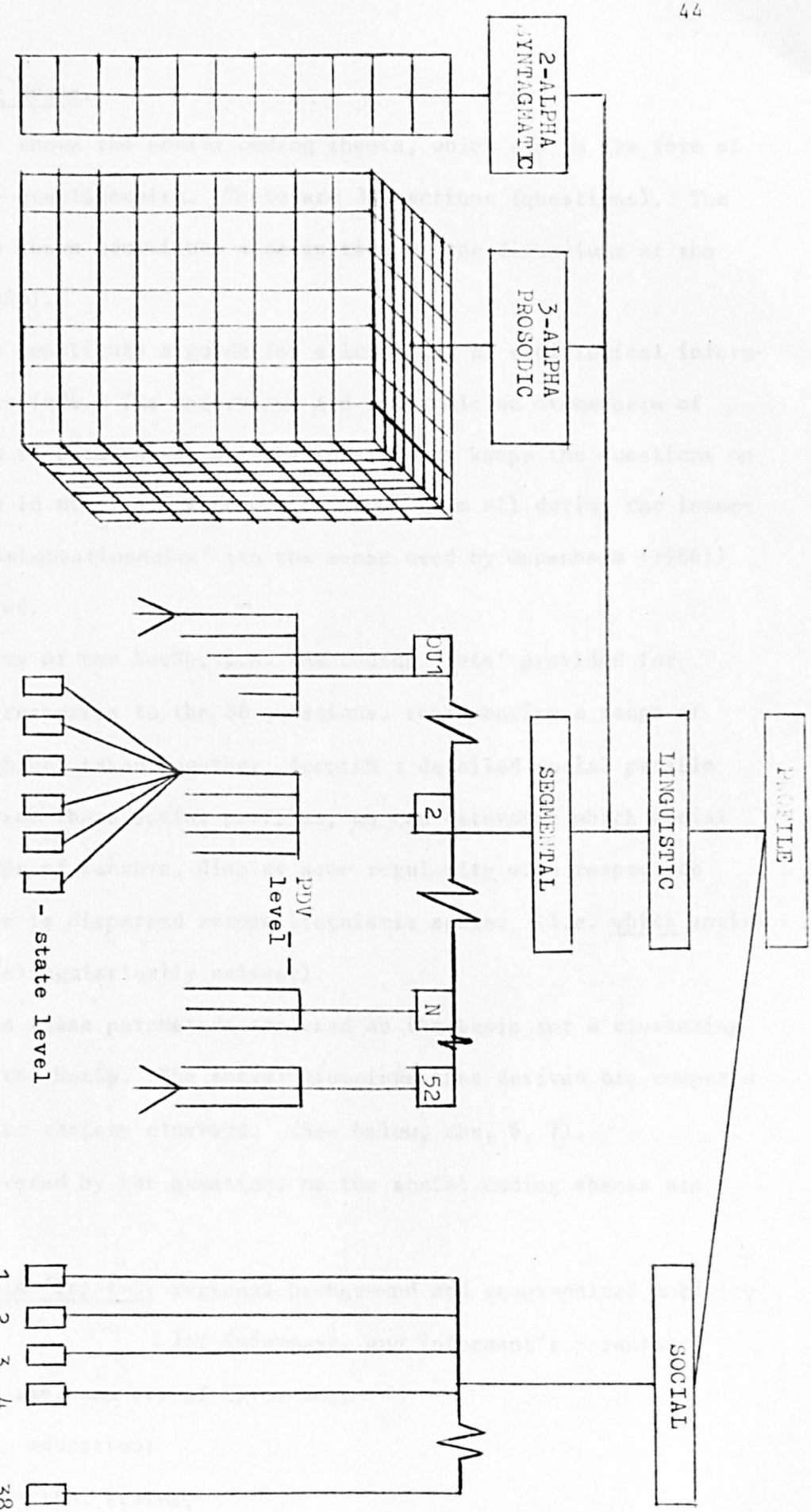


Fig. 4.

Structure of sociolinguistic profile of one informant - TLS coding frame. (Only one PDV is shown in detail.)

The social coding frame.

Appendix (B) shows the social coding sheets, which are in the form of a multiple choice questionnaire. There are 38 sections (questions). The coding categories shown constitute a definition of the dimensions of the social space (SocSp).

These sheets constitute a guide for elicitation of sociological information in the interviews. The interviews are informal: an atmosphere of free conversation is encouraged, but the interviewer keeps the questions on the coding sheets in mind, and attempts to cover them all during the interview. So a "formal questionnaire" (in the sense used by Oppenheim (1966)) is not administered.

The components of the SocSp, i.e. the coding 'slots' provided for informants' responses to the 38 questions, representing a range of social features which, taken together, furnish a detailed social profile of informants. From these social profiles, we can determine which social factors, and groups of factors, display some regularity with respect to the way the sample is dispersed across linguistic space. (I.e. which social features are sociolinguistically salient).

Scores across these parameters are used as the basis for a clustering of informants in the SocSp. The social groupings thus derived are compared with the linguistic variety clusters. (See below, chs. 6, 7).

The areas covered by the questions on the social coding sheets are as follows:

Questions (Qs) 1-5: regional background and geographical mobility (of informant, and informant's parents);

Qs 6-7: age, and sex of informant;

Qs 8-14: education;

Q. 15: marital status;

Q. 16: religion;

Qs 17-20: characteristics of informant's nuclear family;

Qs 23-24: the physical environment (the home);

Qs 21-22, 25-27: informant's attitude to the micro- and macro-environments (local community/the region);

Qs 28-32: occupational status (of informant, and informant's father), also job preferences, job satisfaction;

Qs 33-36: leisure activities, also degree of leisure satisfaction;

Qs 37-38: political allegiance.

There are two kinds of information here: a) analytic, and b) attitudinal.

a) Analytic information is either quantitative. e.g. age, or size of nuclear family, or qualitative, e.g. sex, or marital status.

b) Attitudinal data is coded either by an ordered rating scale (quasi-quantitative), or is qualitative. An example of an ordered rating scale is found in Q35, 'leisure satisfaction', where there is an ordinal relationship between the coding categories:

satisfied/partially satisfied/disgruntled.

Q10, however, 'attitude to education', has response categories which are qualitatively different, and bear no ordinal relationship to one another:

negative/basic skills (RRR)/liberal/job oriented/job oriented and liberal.

This variable is treated as a qualitative multistate variable.

Chapter 4 describes the categories on the social coding sheets in greater detail, and indicates the way in which this coding frame was applied to the sub-sample of 52 Tyneside informants dealt with here.

CHAPTER 3

SOME NEW PROBLEMS RAISED

In the previous chapter I described the T.L.S. model. The design of this model reflects a rejection of some of the assumptions made in previous sociolinguistic studies, and represents an attempt to overcome some of the problems outlined in ch. 1.

The T.L.S. approach has, however, generated a new set of problems in sociolinguistic research, which are of theoretical and practical interest.

The chapter deals with several kinds of problems which have arisen. They are concerned with:

a) the possibility of full implementation of (specifically) the T.L.S model;

b) general problems associated with linguistic analysis;

c) general problems associated with classificatory procedures.

It is demonstrated here that firstly, given the way in which the T.L.S. coding frame was applied to the data, and also for reasons under b) above, it is not possible to entirely fulfil all of the original objectives of the T.L.S. as set out by Pellowe et al (1972, pp.1f). Secondly, it is shown that ad hoc decisions still have to be taken at various stages of the process, and, thirdly, that the use of multivariate analysis created its own problems, and introduces new sources of bias. The specifically computational problems which have arisen are dealt with below (ch. 4).

I have said that the T.L.S. aims to generate a classification of the Tyneside speech community which satisfies the empirical objectives of exhaustive representation of linguistic varieties, repeatability, and statistical adequacy. For a comprehensive picture of the mathematical characteristics of linguistic and social variability to be built up, the speech of a representative sample of the population must be plotted along all relevant dimensions of variation. Statistical techniques, and the capacity of the electronic computer to process large bodies of data provide the linguist with opportunities to attack problems of this magnitude and complexity. Thus a classification of 200 informants on the basis of 690 linguistic, and 38 social variables is feasible. However, multivariate techniques applied to this kind of data pose methodological problems of a different order to those faced in traditional dialectological surveys, or the statistically simpler Labovian sociolinguistic classifications, which deal with variables singly at the analytical stage. Some, but not all, of these problems have been encountered in other disciplines, e.g. by biologists using multivariate analysis techniques to generate taxonomies of animal species.

In general, the reliability and usefulness of a classification depends on:

1. inclusion of as many variables relevant to the classification as possible, (absence of relevant variables effectively attributes zero weight to them, thereby skewing the results);

2. exclusion of irrelevant variables, (whose presence would artificially depress, or inflate, similarity between individuals);

3. exclusion of logically dependent variables, (their inclusion would boost similarity levels);

4. differential weighting of variables, where appropriate (i.e. according to their relative importance in assigning individuals to classes).

Given only these general constraints (there are others too), to what

extent can the objective of an exhaustive and unbiassed classification be realised in practical terms?

Clearly, ad hoc decisions must be taken by the investigator at an early stage concerning which of all the possible variables are relevant to the purposes of the classification, which variables are logically inter-dependent, and which variables, if any, are to be given a more significant role in assigning entities to classes. So the investigator must, by observation, try to discover all the features which have variant realisations across the sample under study, (i.e. determine which features are variable). He must then attempt to isolate, and exclude, any variables which are logically dependent on other variables. He must then evolve a system of scoring for each variable. Similarity between individuals will be derived from a function of comparative scores over all these features.

The first problem, then, concerns the selection of variables. This is limited by the extent of the investigator's knowledge derived from observation of the entities under study. In biological taxonomy, for instance, certain facts will be known about the anatomical structure of specimens, and some of the characteristics of their biochemical and neurophysiological processes, and these may all be incorporated. However, there is always the possibility that new facts will come to light in future studies, and perhaps a completely new level at which to study the subject, just as the discovery of the atom opened up a whole new perspective from which to view the physical world. The range of potentially discoverable facts, and levels at which analysis can occur, is open-ended for social and linguistic characteristics, and no survey can claim to finally exhaust all possible analytical distinctions and perspectives. So any classification will reflect, and be a projection of, the extent of current knowledge in the field of study.

Out of the data currently available, the investigator will select only those variables which he considers relevant to his classification. For example, a classification of psychological disorders amongst middle aged

men may not take into account eye colour or father's birth place as relevant variables. In the case of such specialised classifications, with a restricted utility, (i.e. the range of questions which the investigator wishes to pose is limited), grounds for exclusion of certain variables are often fairly obvious. However, in a more general classification, where it is desirable that the data base should be capable of interrogation in many different ways, and the type of covariances or correlations to be sought cannot be predicted in advance, then as many variables as possible should be included, as each is potentially relevant to some question of distribution or co-occurrence. As Everitt (1974, p.48) points out,

"the initial choice of variables is itself a categorisation of the data which has no mathematical or statistical guidelines, and which reflects the investigator's judgement of relevance for the purpose of the classification. (This of course could also be said of the entities chosen for study)."

With a general classification in mind, the T.L.S. investigators consider all discriminable linguistic variables as potentially relevant, and aim to be as exhaustive as possible in the selection of variables in order to minimise the possibility of an a priori categorisation of the data.

Nevertheless, there are practical constraints on the range of variables included. Firstly, variables must have a high probability of occurring, or being encouraged to occur (elicitation) in an interview. Non-incidence of a variable in one interview precludes comparisons between that informant and every other informant with respect to that criterion. Thus, for each variable which is not elicited in an interview, the VSp is effectively reduced by one dimension for that individual. If many variables with a low probability of elicitation were included, the VSp, and mutual similarity indices between pairs of individuals, hence their locations in the space, would be grossly distorted. So certain lexical items, for example, though intuitively known to be characteristic features of the speech of some Tynesiders, are not

included as they or their alternants are not likely to be produced by all, or nearly all, informants in an interview. Thus lexical variables are underrepresented in the model, as compared to phonological ones, which have a greater probability of occurrence. This of course means that the clusters generated will be determined largely by phonological variables as these are more numerous than syntactic and lexical variables. It has been demonstrated, (Garvey & Dickstein: 1972) that linguistic features co-vary with social criteria differently, depending on the level at which analysis is implemented: in view of this, the disproportion between the number of variables included at the different levels of analysis, (phonological, grammatical, lexical) may not only make the classification more remote from the hearer's scheme of sociolinguistic variation, but also skew it towards the behaviour of phonological features at the expense of the other levels of variability.

As I have remarked above, variables must initially be identified by observation of speech. This raises a special problem for linguistic classifications. The analysts are themselves hearers and speakers, and as such their perceptions of realisations of different features will be variably skewed, relative to their own place in the VSp. Each analyst will discriminate a slightly different sub-set of the total set of (in theory) discriminable variables. Pellowe et al (1972, p.19f) acknowledge this, but express their expectation that as more analysts joint the team, their diverse linguistic backgrounds and perceptions will result in the recognition and inclusion of more variables: hence the number of excluded variables will tend to decrease. This is described as a progressive reduction of distortion in the VSp:

"as the number of investigators increases ... we find that an increasing number of topological deformations is contributed. However, because of the different types and directions of deformation, we find that conflation of different selections of criteria tends to a regular (i.e. undeformed) VSp "

(Pellowe et al: 1972, p.20).

As far as exhaustiveness of criteria is concerned, obviously the number of possibly relevant criteria included will tend to increase. In this way, the VSp itself will be filled out with added dimensions, and will therefore be less distorted by omissions. But, although the model is described as dynamic in the sense that 'new' variables can be incorporated as they are discovered during theprocess of analysis, (and this is the stage at which 'new' variables are likely to be discovered)the addition of new variables whilst analysis is ongoing means that, although the space is improved, informants already analysed have not been coded on the new criteria, and are therefore treated as NC (non-comparable) on these variables. (I.e. in effect, they score zero on these features).

The fact that a variable has not been observed before does not necessarily mean that it has not occurred in the data already examined, only that it has not been perceived by an analyst before. So an NC score on new criteria for informants already analysed may in some cases be erroneous. Once a variable has been established and incorporated into the coding frame, instances are more likely to be perceived in subsequent analysis. It is a well attested fact that sensory perception is channelled according to the constructs which exist for the perceiver. The existence of terminological labels aids perceptual discriminations. For instance, a knowledge of the terminology of wine tasting not only provides a descriptive vocabulary, but a set of reference points from which finer distinctions can be perceived. Therefore, as analysis proceeds, and new variables are accumulated, informants may be coded to a greater degree of fineness.

So, although the VSp itself is progressively improved, its contents are retrogressively distorted.

It would be possible, if very time consuming, to reanalyse all the data a second time, incorporating the new variables, (although it would be impossible to say finally that no more criteria could be discovered).

This would optimise both the VSp, and the scatter of informants through it.

Alternatively the new variables could be added as they crop up during analysis, and their effect on the composition of clusters could be determined by computing the similarity matrix and implementing the clustering program twice, masking the added variables on the second run. If the two sets of results differed only negligibly, then the added variables could be discounted as significant discriminators of sociolinguistic sub-groups.

As far as the dimensions of the space are concerned, then, the addition of new analysts will, as Pellowe et al. claim, reduce 'topological deformation.' However, this is not the case at the stage of analysing and coding informants according to these dimensions. Here analysts' perceptual differences will still operate despite the existence of an improved VSp. There is evidence to suggest that analysts will hear, and therefore code, realisations differently. As Ringaard (1965) unhappily notes:

"We must come to the sad conclusion that the transcriptions of phoneticians do not tell us so much about the speech of the area they are studying as about the phoneticians themselves."

Ladefoged (1960) has demonstrated the difficulty of establishing a standardised phonetic system whereby different analysts can be certain of reaching a reasonable degree of agreement in transcribing the same stretch of taped speech, and warns us that

" a phonetic statement can be considered to be adequate only if it has the same meaning for all who use it." Similarly, the coding of phonetic criteria is only of use insofar as all analysts agree closely in their perceptions and coding habits, which is unlikely. The problem extends to prosodic criteria also: analyses of level tone, for example, vary significantly with the tone patterns of the analyst. Even syntactic analysis is influenced by the analyst's own norms. (E.g. complementizing quantifiers and post modifiers are coded differently by analysts who themselve use non-standard forms of these features.)

So if we accept that the coding of informants will be biassed according to which investigator analyses which informants' data, then each analyst's section of the VSp will be distorted in a different direction and to a different degree. The notion that a multiplicity of personal deformations will have a mutually compensatory effect in this sense is a little unrealistic.

However, with several sources of bias, these can be quantified with respect to each other in a way that is not possible if only one analyst is concerned. The contributions of different analysts can be masked out in turn, and the clusters recomputed, or different analysts can code the same sub-set of data (duplication), and the effects of analyst bias can be calibrated. It will then be possible not only to measure the degree of deformation attributable to each analyst, but also, the linguistic features which are relatively more susceptible to differential distortion by analysts can be identified.

Once the variables have been selected, it then remains to work out a system of coding the variant forms of each. (For a description of the coding of variables see chapters 2, 4 and appendix x). It will be seen that in the OU scheme, the number of states (variant realisations), of P.D.V.s varies between one and eleven. The probability of occurrence,then, of any one state differs greatly from one P.D.V. to another. This means that the occurrence of one state is automatically weighted by the range of possibilities within its superordinate unit (its PDV). Is this a source of bias? Providing the variables are real and natural, (i.e. not artificially structured by the coding frame) and the paradigm of states associated with each PDV genuinely reflects the range and number of discriminable realisations of that segment, then this may be a reasonable source of inherent weighting of variables.

One shortcoming of the T.L.S. coding frame is that only certain kinds of co-occurrences can be automatically retrieved (viz. those features coded

in the 3-alpha data). Features analysed at one level (e.g. segmentals) cannot be related to those operating simultaneously at another level (e.g. lexis); so, for example, it is impossible to discover the lexical item in which a certain segment appears, or to identify, in every case, the syntactic function, or sentence position of a word carrying a certain tone. This kind of information is crucial, as the different levels interact hierarchically and simultaneously in a speech situation, and it is the combined effect of context and intra-systemic contrasts (e.g. changes in pitch range relative to immediately preceding pitch patterns) that endow features of these systems with signification. For instance, the abstraction of a total of rising tones from a stretch of utterance tells us something, but additional information about e.g. sentence position of tones would be useful. Rising tone in final position in many varieties of English indicates the interrogative mood, but on Tyneside this is often found to be a feature of indicative clauses. We know that this particular interaction between features of different linguistic systems (intonation/syntax) is a marked localised characteristic, but there are many cases where a nested coding system permitting complete retrievability would facilitate the discovery of hitherto unnoticed co-occurrences between systems. In this respect the T.L.S. model oversimplifies the hearer's competence in simultaneously perceiving and interpreting bundles of features operating at different levels.

Non-retrievability of context is a particularly serious handicap with prosodics because they operate at word and syllable level, (e.g. tonicity) and clause level (tonality etc.) and they affect the meaning of whole utterances as well as the lexical item carrying them. And neither lexical nor whole utterance context is retrievable from machine store, so the semantic function of prosodics cannot be deduced except by manual means i.e. going back to the raw data and checking through for the locations of particular features. This obviously defeats one of the objects of a mechanised information store, and precludes some of the possibilities for recognising

new interactions between elements of different linguistic systems.

In the initial survey design stage, certain a priori decisions must be taken, concerning relevance of variables to the classification, concerning relative probabilities of elicitation, and concerning the degree of fineness to which variables should be coded. All these decisions will affect central statistical parameters, and hence will have a deterministic effect on the structuring of the data. The best that can be hoped is that these decisions will be taken in such a way as to minimise bias.

Once the coding frame is worked out, there are still more decisions to be taken which will affect the classification. The investigator must select from all the statistical techniques available those he considers most appropriate for producing the kind of classification he desires. At each stage of the statistical and computational processes there are several options open, each of which treats the data differently, and will elicit a different structure from the data. There may be clear criteria for assigning more weight to certain variables, (i.e. assignation of individuals to classes is to rely more heavily on some features than others). Such a weighting scheme can be easily incorporated into the computation of similarity coefficients, but this reflects either a) an assumption that those features will discriminate existing groups more efficiently, or b) that it is the investigator's intention to base the arrangement of the data with respect to those features especially.

There are no such reasons for assigning differential weights to criteria on the T.L.S., firstly because it is not known in advance which features will turn out to be diagnostic of linguistic and social clusters, and secondly because the desired classification is to be a general one. So, all variables are treated as homologous, and carry equal weight in the calculation of similarity between pairs of individuals. A problem arises with negative matches - shared absence of a feature must be significant, but perhaps not as significant as shared presence of features, and a decision to assign half

weight to negative matches may be taken.[FN]

FN. Or, negative matches may be excluded altogether. See below (ch.4) for a description of a similarity coefficient which excludes the effects of negative matches.

Such a weighting is of course not known to accurately reflect (numerically) the degree of significance which hearers attach to this kind of negative similarity, nor can it be known in advance whether the contribution made by negative matches to the similarity matrix will be a helpful one in generating clusters which represent real groups of related varieties across the population. But this is an instance of the sort of arbitrary mathematical strictures which classificatory techniques must impose on the data.

With a sociolinguistic survey, it is impossible to devise, a priori, a system of weighting for variables: even in biological taxonomy, where the degree of phenetic relationship between specimens can be estimated by observation with reference to already existent taxonomic schemes, it is not advisable to adopt a system of a priori weighting (Sneath: 1957). Sneath recommends a posteriori weighting if "certain characters[FN] have a proven discriminatory function for desired groupings." When variety clusters, and their diagnostic features have been established it may then be possible to apply a differential weighting system to data collected (or reprocessed) subsequently on the strength of this.

FN. 'character' in biological taxonomy is the equivalent term to 'criterion' on the T.L.S.

Sneath discusses hypermultivariate and oligo-variate strategies, (the latter being multivariate, but with only a few variates) and points out that a reduction in the number of characters increases the effects of sampling error. Moreover, oligo-variate techniques are only powerful enough to discriminate between already established groups. Hence the need for inclusion of as many variables as possible in the Survey. However, Sneath considers

it essential that individuals missing any of the quantitative variables, (and all linguistic variables on the T.L.S. are treated as quantitative at the analysis stage as comparisons between individuals' profiles are made on the basis of continuous numeric quantities) should be eliminated from the classification. If this was done, probably all informants would be eliminated, or if not all, the elimination of some would disrupt the randomness of the sample. So although it is statistically important to include as many variables as possible, it is also essential (according to Sneath) to ensure that all criteria, (not most, as claimed by the T.L.S. researchers) are going to appear in the speech sample of every informant, if Sneath's argument applies to sociolinguistic data.

The next problem is the choice of the most appropriate measure of resemblance between individuals according to their scores across all criteria. Here again there are many different statistics available each of which represents the data in a different way, and the one which produces the optimal representation with respect to the type of data and the required classification must be sought. The most appropriate measure to use on the T.L.S. project is not immediately obvious. Using, for instance, Euclidean distance measures, different clustering algorithms produce different results depending on whether the data is raw or standardised [FN] (Everitt: 1974, p.48).

FN. Scores on states, for example, could be standardised to zero mean, and unit variance. This means that the sample mean score on that state would be expressed as zero, and the range of scores across the sample would range from -1 to +1.

Everitt also points out that if raw data is used, the clustering patterns which emerge are dependent on the scaling of the parameters. (E.g. radically different results would be obtained if, e.g. height was expressed in inches, instead of feet). This is not a problem for the T.L.S., as social data is coded in binary not quantitative form, and linguistic parameters are mutually comparable. (I.e. the scaling problem does not apply

to linguistic variables, as all variants are scored according to their relative frequency of usage, and scales of measurement are not mixed.[FN])

FN. Problems of mixed scaling would occur, e.g. in biological taxonomy, where one variable might represent weight of specimen, and another, length of specimen. Here, measurements are made with different kinds of units.

Even if there is no scaling problem, however, different coefficients still impose their own constraints on the data. Even if a variable weighting scheme is not adopted, many similarity coefficients do not even preserve the original equal weighting of variables. As Burnaby (1966) remarks:

"the consequence of choosing one or other of a plethora of similarity indices is in many cases equivalent to the adoption of different schemes of variable weighting, and so the concept of 'equal weighting' is not as simple as it seems at first sight."

He points out furthermore that variables not included in the analysis are effectively assigned zero weight, which is a persuasive argument for exhaustive sampling of variables.

Cormack (1971) demonstrates that different coefficients yield widely differing results for the same set of data, and, the more damaging fact that the values for all paired comparisons over the different coefficients are not jointly monotonic, i.e. the ranking of individuals according to degree of association varies between coefficients, as well as the values of the association coefficients. This may be explained as the creation of different data sets according to which coefficient is selected, however it emphasises the critical nature of the decision to use a certain similarity or distance measure.

According to Everitt (1974), the choice of similarity index must rely on investigator intuition. He says of distance measures:

"In brief, the choice of the correct distance measure to use would

be much simpler if we had prior knowledge of the structure of the data..." That is, some measures will come closer to eliciting the real structure of the data in question than others. But this is circular, for it is the classification techniques which organise the data into a structure which is more, or less, relevant to the purposes to which the data configuration is to be put. The naturalness of a classification, providing the sampling procedures are valid, is relative to the purposes of the investigator; it is doubtful whether absolute objectivity in a classification, (c.f. Labov's (1972, p.98) 'first and right' principle) is realistic, or attainable.

The optimal similarity measure (or measures) must be found by trial and error, which will involve the investigator in comparing the results produced by using a number of different measures, and taking a decision as to which yields the best results.

There are also many kinds of clustering techniques available, each of which predisposes the data in a certain way. It is possible to rearrange clusters using optimisation techniques, but these require a great deal of computer time, and "are probably not suitable to use with very large data sets " (Everitt: 1974, p.64). According to Everitt, clustering techniques are useful for data reduction or dissection, but if the search is for real, discrete sub-types, there is a danger of getting spurious solutions as most are biassed towards producing spherical clusters. If the variety clusters within the Tyneside community do not naturally resolve themselves into the kind of patterns discovered by certain clustering methods, then the application of such techniques must result in distortion of the data and considerable loss of information. And there is no reason for assuming that the variation in Tyneside speech groups will be reasonably accurately characterised by well-defined clusters: it may be a variably dense, but non-divisible continuum which would nevertheless be forced into a cluster pattern by the techniques applied to it. Some index of naturalness based on other criteria than the properties of the clusters needs to be devised, by which the results

of different clustering techniques can be assessed. One possibility is 'cross-validation' by hearer judgments.

Wishart (1969) has demonstrated that, of agglomerative techniques, single linkage is very sensitive to 'noise points', i.e. outlying individuals interfere with the clusters they are close to. Everitt suggests the detection and obliteration of these 'outriders', but this would be unsatisfactory on the T.L.S. Firstly, many of the informants are likely to be outriders, and their removal would disrupt the balance of the sample, secondly, outriders' speech is of interest to the survey.

Even amongst classification experts there is a great diversity of opinion concerning the relative merits of various clustering algorithms. Lance and Williams (1967) declare that nearest neighbour sorting (single linkage) should be 'regarded as obsolete' whilst Sibson and Jardine regard single linkage as 'the only agglomerative technique to be recommended.' (See ch.4, pp.94ff.)

Once again the only solution is to try several techniques and compare their results.

The other way of assessing the relative merits of the different similarity measures and clustering techniques is by the external evaluation procedures mentioned in Pellowe et al. 1972, namely by hearer judgement tests. However, hearer judgement tests are notoriously difficult to design and apply, and results are often erratic and difficult to interpret, and it would appear that, for this sort of test to be useful some sort of training of hearers would be requisite.

The question still remains, how natural will such a classification be, given the new sources of bias introduced by classification techniques and those introduced at certain stages of the analysis and computation? And how far can the precepts of exhaustiveness and inclusiveness of variables be sacrificed to 'the exigencies of definition and computation' before all hope of achieving a natural classification is lost?

These questions may only be answered by inspection of the results of processing the data.

CHAPTER 4

PROGRAMMING AND DATA PROCESSING

This chapter describes the computational processes applied to a sub-set of the T.L.S. data. This sub-set consists of:

a) The segmental phonological data of 52 informants;

b) The social data (for the same 52 informants).

The programming language used is PL/1[FN], (Programming Language 1, originally to be called N.P.L. - New Programming Language.)

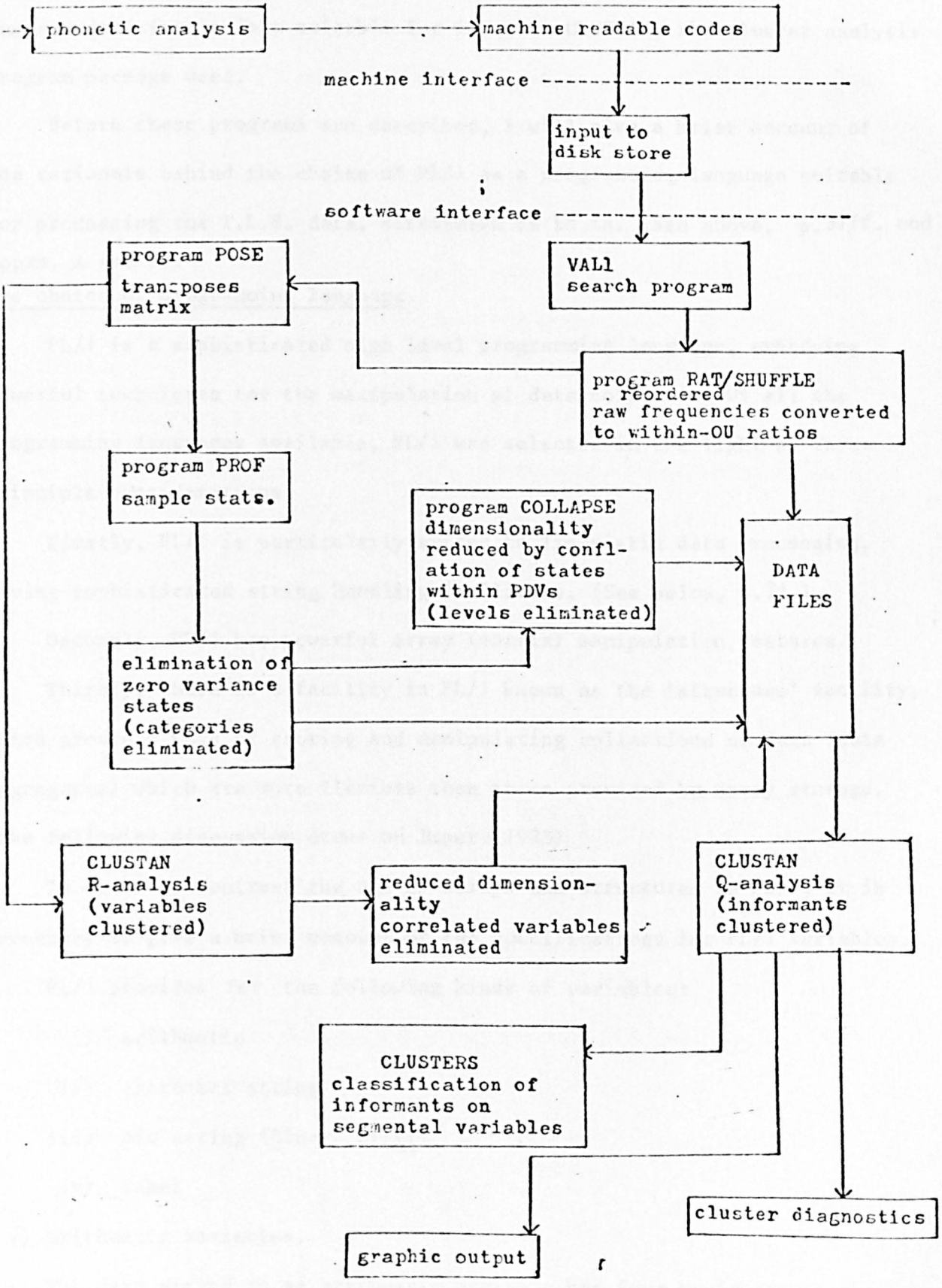
FN. The PL/1 compiler supported at NUMAC is the F-level compiler.

All programs referred to, with the exception of the CLUSTAN suite, were written by the author, and implemented at NUMAC, (Northumbrian Universities' Multiple Access Computer,) under the MTS, (Michigan Terminal System) operating system.

The sequence of processes which are applied to the segmental phonological data which is analysed here is given in flow chart form in Fig.5.

Fig. 5:

Flowchart - sequence of processes applied to segmental phonological data.



(a) Linguistic data - processing of.

The programs described below (pp. 79ff.) were designed to transform the raw data into a form suitable for input to CLUSTAN, the cluster analysis program package used.

Before these programs are described, I will give a brief account of the rationale behind the choice of PL/1 as a programming language suitable for processing the T.L.S. data, structured as it is. (See above, p.37ff. and Appxs. A and B).

The choice of programming language.

PL/1 is a sophisticated high level programming language, embodying powerful techniques for the manipulation of data collection. Of all the programming languages available, PL/1 was selected in the light of three principle considerations.

Firstly, PL/1 is particularly suited to linguistic data processing, having sophisticated string handling facilities. (See below, p.71).

Secondly, PL/1 has powerful array (matrix) manipulation features.

Thirdly, there is a facility in PL/1 known as the 'structure' facility, which provides ways of storing and manipulating collections of data (data aggregates) which are more flexible than those provided by array storage. (The following discussion draws on Roper: 1973)

In order to contrast the use of arrays and structures in PL/1, it is necessary to give a brief account of the specifications for PL/1 variables.

PL/1 provides for the following kinds of variables:

i) arithmetic

ii) character string

iii) bit string (Binary digit)

iv) label

i) Arithmetic variables.

The data stored in an arithmetic variable has four basic attributes: scale, base, mode and precision.

"The scale of an arithmetic data item is either fixed-point or floating-

point. In fixed-point data the position of the decimal or binary point is specified, e.g. 1.234. In floating-point data a fixed point number is followed by an exponent which specifies the true position of the point relative to the position in which it appears, e.g. 1.234E4 is equivalent to 12340.0.

"The base of an arithmetic data item is either decimal or binary..."

"...mode...is either real or complex..."

"precision...is the number of digits (either binary or decimal) a fixed point data item may contain or the number of significant digits (excluding the exponent) in a floating-point data item " (Roper: 1973, section 3.2).

ii) Character string variables.

A character string is a connected sequence of characters consisting of alphabetic, numeric or 'special' characters, (e.g. quotes, stops, commas, dollar signs.)

iii) Bit string variables, and iv) Label variables:

these variable types will not be discussed here.

Arrays

Arrays (or matrices) are "'collections' of data (or data aggregates) that can be referred to by a single name " (Roper: 1973, section 3.12).

Array manipulation comprises a powerful set of techniques for handling collections of data arranged according to one or more dimensions.

In PL/1, a one-dimensional array is called a 'list'.

One of the uses of arrays in some of the programs described below (pp. 79ff.) involves setting up a two-dimensional array to hold scores across states for a number of informants (or 'cases') (See Fig. 6)

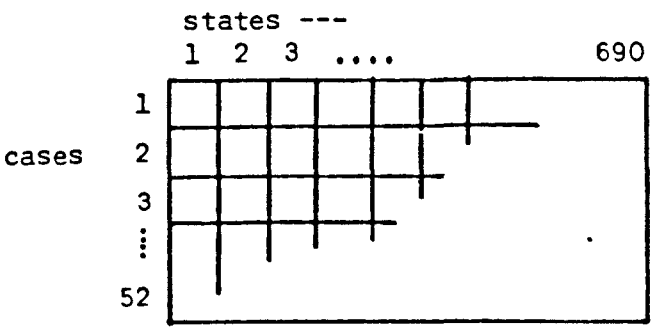


Fig. 6

An array.

If the array is called 'ARR', single cells may be accessed by the use of subscripts thus:

ARR (1,1) refers to the score recorded for the first case on the first state. The following constraints are imposed by the standard array form: firstly, data items stored in an array must be homogeneous in terms of their attributes, and, secondly, the array must be symmetrical.

Homogeneous data items share (in PL/1 terminology) the same data type (character or arithmetic), and the same scale, base, mode and precision.

Symmetry of dimensions requires that all rows must have the same number of columns, and all columns the same number of rows, (rectangularity).

PL/1 Structures

Structures in PL/1 permit the storage and manipulation of data aggregates, but are not subject to the constraints imposed by the array form mentioned above (p.66).

PL/1 structures are defined thus:

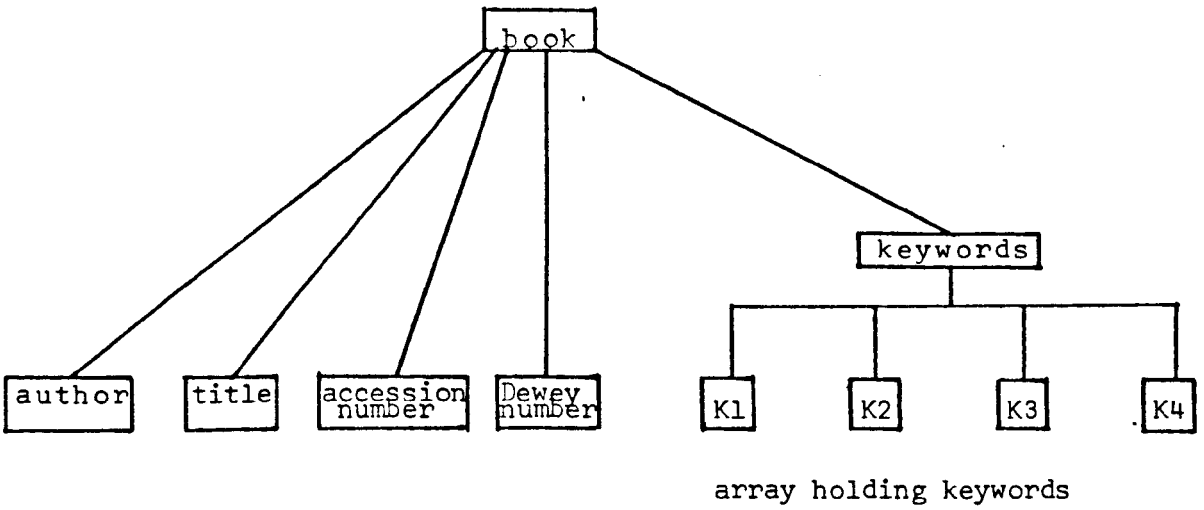
"A structure is a hierarchical collection of names. At the bottom of the hierarchy is a collection of elements, each of which represents either a single data item or an array." (Roper: 1973, section 6.2).

These elements at the bottom of the hierarchy are known as 'terminal elements'. Because a structure has a tree form, the constraint of rectangular symmetry obviously does not apply. Moreover, mixed attribute data may be stored in a structure.

Very complex forms can be constructed: structures may contain substructures consisting of arrays, or arrays of structures, related at different levels.

For instance, in an automated library cataloguing system, the catalogue entry for each document may include information such as author's name, title, accession number, Dewey classification number, and possibly a series of keywords describing the subject and content of the document. Some of these data are of character string type, some are numeric. A structure such as that shown in Fig. 7 could hold this information. (K = keyword).

Fig. 7.
An example of a structure used to store information.



Referring back to Fig. 4 , (coding frame structure - ch. 2) it appears that each informant's sociolinguistic profile, as analysed according to the T.L.S. coding frame, could be usefully stored in a PL/1 structure. The cases analysed could be stored in an array of such structures, with one structure per case. For reasons explained below (p.136), the social data is treated as binary (bit) (presence/absence) and the linguistic data as floating point decimal. As noted above (p.67), data attributes can be mixed within a structure. Moreover, the non-symmetry of the segmental phonological scheme would present no problems. (It is asymmetrical because different OU's have different numbers of PDV's, and PDV's different numbers of states. Also the PDV level does not exist for many of the consonantal OU's.)

Each informant's profile would consist of a multi-level structure, formally identical to that shown in Fig.4 (p.44):

Level 1 - structure name, 'PROFILE'

Level 2 - would be divided into linguistic, and x-linguistic (social) data.

At level 3 the linguistic data would split into segmental phonological, 2∝ and 3∝ data.

The segmental data substructure would consist of a 3-level hierarchy: OU's at level 4, PDV's at level 5, and states at level 6. The last of these levels,

(the 'terminal' elements of the structure, (see above, p.67)) would hold the actual frequency scores for states.

The 2-alpha data could be stored in a one-dimensional array (or 'list') and the 3-alpha data as a 3-dimensional array, each cell of which represents a specific combination of terms from the three linguistic systems, intonation, pitch range, and grammatical form class.

(E.g. PROFILE. LING. 3-ALPHA (2,2,1))would contain the frequency of occurrence, in one informant's interview, of the combination: falling tone, on common noun, preceeded by booster).

This plan, however, proved to be impractical.

Firstly,(a consequence of the history of the development of the coding frame), the organisation of codes does not completely reflect the hierarchical nature of the segmental variables. If it did, the accumulation of state scores could be efficiently managed.

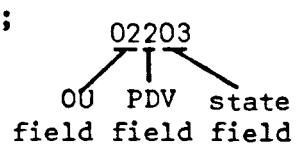
If, for instance, the five digit codes were structured in such a way as to reflect the 3 level hierarchy of OU's, PDV's, and states, the use of the PL/1 structure could have been an effective programming tool.

For instance, the first programming task is to count up all the instances of each state type in each informant's raw data file.

If the first 5-digit code in the file represents the 3rd state of the 2nd PDV of the 2nd OU, the program must find the corresponding counter FN in the structure, and increment it by one.

FN. The 'terminal elements' at state level are storage locations to hold number of occurrences - i.e..they can be used to count.

If, for instance, the codes used consisted of an OU field, a PDV field, and a state field, thus;



then the 5-digit code could be segmented into 2, 1 and 2 characters from left to right, the 3 resulting values then being used to traverse[FN] the structure to reach the correct counter.

FN. 'Traversing' means following a path down through the structure, via the appropriate nodes specified.

'02203' would be segmented into 02 = OU2

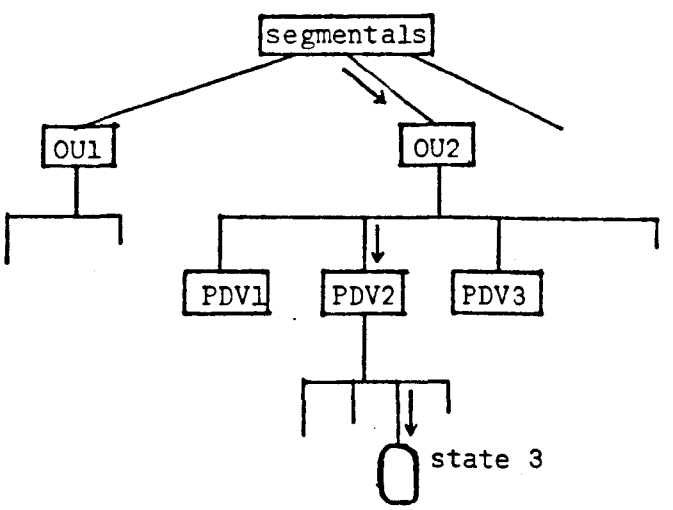
2 = PDV2 of OU2

03 = state 3 of PDV2 of OU2

(See Fig. 8)

Fig. 8.

Traversing a structure holding the segmental data.



In fact, however, the numeration of the 5-digit codes only reflects the hierarchy from PDV level downwards.

The 6 PDV's of OU1, for instance, are coded

0002, 0004, 0006, 0008, 0010, 0012, respectively. (The fifth digit, as in 00021, represents the state of the PDV).

The PDV's in OU2 then carry on with the even numbering, from 0014, onwards.

Thus, members of a given OU have no unique string associated with their codes. (Cf. my post hoc suggestion above (p.69) where the first 2 characters identify the OU.)

The hierarchy is reflected from PDV level downwards: all states which

are members of the same PDV share their first four digits, the fifth designating the particular state of that PDV.

eg. 00021
2
3
4 etc.
} states of PDV1 of OU1.

If the coding strategy suggested above (p. 69) had been adopted, then all states would have been uniquely identified with reference to their superordinate PDV and OU.

Input would have been treated as character string rather than arithmetic data. (See below (p.79) for a description of the actual implementation of the search program, VAL1.) The PL/1 SUBSTR[FN] (substring) function would then have been used to segment each 5-digit code into its OU, PDU, and state components.

FN. "SUBSTR extracts a substring of user-defined length from a given string and returns the substring to the point of invocation." (IBM System 360 Operating System:PL/1 (F) Language Reference Manual.) Thus if CODE(1) is a string variable, and has the value '01101' assigned to it, then SUBSTR (CODE (1), 1, 2) will return the value '01', which would refer to OU1, and could be used to point to the corresponding node (OU1) in the structure.

This method (allocation of codes to structure elements by segmentation of the 5-digit codes) I, in fact, attempted (on one OU only). However, the logic of the numbering of 5-digit codes, as it stands, made the programming too unwieldy in terms of the amount of program code to be written, and in terms of central processing time. This is because, for any 5-digit string input, in order to take the path to the appropriate OU node in the structure, the program must recognise not one, but a set of substrings associated with that OU. ("Members of a given OU have no unique string associated with their codes". See above, p. 70).

For example, the section of program which causes a move along the path to OU1 (see Appx. A and discussion of Program VAL 1 below) must recognise 6 different substrings, which must be specified explicitly in the

program. (The six are 0002, 0004, 0006, 0008, 0010, 0012.)

Similar specifications must be built into the program for each of the 52 OUs. Obviously this would have resulted in a very messy and uneconomical program.

Program VAL1 (see below, p. 79) provides a simpler solution to the frequency counting problem, without embodying the coding frame hierarchy in a PL/1 structure.

So, because of the numbering convention adopted in the coding of segmental variables, the PL/1 structure facility could not be exploited in the search program as was originally anticipated in the early 1970s.

It had also been hoped to use the flexibility of PL/1 structures to expedite the iterative application of some general process to the contents of successive sub-parts of the structure. For example, Program RAT (described below, p.96) converts state scores from raw frequencies to within -OU percentages. This process of conversion to %ages can be expressed in a general algorithm, which could be applied to the sections of the structure 'Profile' subordinated by each OU node, in turn. It is easy to apply a process iteratively to parts of arrays.

It is possible to access elements of arrays iteratively, by using a control variable and altering its value. For instance, a one-dimensional array can be set up, called, say, RA, and containing 10 elements, RA(1), RA(2). ...RA(10). The bracketted subscript '(1)' refers to the first element of the array, and so on. If, instead of referring to a specific element, a control variable N is used, the elements of the array can be accessed in turn by specifying RA(N), and changing the value of N. N can be given an initial value of 1, and then incremented by steps of 1 until it reaches the value 10.

Thus, if we want to perform the same operation on the contents of each of the array elements, it is only necessary to write one piece of program using the subscript N instead of a specific number, and LOOP through it 10 times, once for each value of N from one to ten.

If the array contains 10 numbers, we can, for instance, double each of these numbers using a loop. (In PL/1, a DO loop may be used, or a GOTO statement).

The following 3 lines of program would perform this operation 10 times: i.e. on the contents of each array element in turn:

a) DO N = 1 TO 10;

RA(N) = RA(N)*2;

END;

The '=' is the assignment operator (equivalent to := in Algol). The same symbol functions as the equals sign: the 2 functions are disambiguated by the syntactic context of the PL/1 statement containing the symbol.

The '*' is the multiplication operator . This symbol is also used to designate all elements of an array, and is known as the 'asterisk function'.

b) RA(*) = RA(*)*2;

performs the same operations as a) above.

(Once again, the syntax of a PL/1 statement determines which meaning '*' ha in a given context.) This technique of accessing the contents of an array iteratively is a powerful and useful programming tool. However, arrays have limitations which structures do not have, (see discussion of symmetry and homogeneity above, (p.66).

However, it is not possible to access sub-structrues of a structure iteratively in the same way as the elements of an array can be accessed.

To return to the problem introduced above (p.72) (see also p.92ff.) it would be useful to be able to take the sub-structure subordinate to each OU node in turn, and apply to it a piece of program which converts raw state frequencies to within-OU percentages. This would require a loop of the form:

step 1. Set up a DO loop in which N takes the values (1,2,3...52);

step 2. For each value of N, sum the contents of the state counters subordinate to OU(N);

step 3. Store sum in variable (eg. OU. TOTAL), divide the contents of each state counter by OU.TOTAL x 100.

This would be an economical method of performing the operation of conversion to percentages, requiring only a few lines of program code.

However, it is not possible to access parts of structures in the same way as one can access elements of arrays. Unlike array elements, nodes in a structure are names, which cannot carry subscripts. i.e. 'OU1' is a string which names a node in the structure, but it is not possible to substitute a control variable, eg. as in OU(N). This kind of strategy was attempted, and was rejected by the PL/1 compiler. On the failure of this strategy, another was attempted. This involved building the strings, OU1, OU2... etc. by concatenating[FN] the string 'OU' with the current value of a string variable valled N, which, by a D loop, took the values 1, 2, 3...52.

FN. "Concatenation of strings is a device for building longer strings from shorter strings by joining them end to end." (Roper: 1973, Section 3.8A).
Thus if STRING 1 has the value 'BRAIN'
and STRING 2 has the value 'STORM'
the assignment statement, ('|| ' is the concatenation operator)
STRING3 = STRING1 || STRING2;
will assign to the variable STRING3 the value 'BRAINSTORM'.

This program step was:

```
DO  N = 1 BY 1 TO 52;
OU.TOTAL = SUM(OU||N);
```

This kind of strategy is also rejected by the compiler.

The team who designed the PL/1 language stated six objectives: objective (1):"Nothing to be illegal which makes clear and unambiguous sense" has perhaps not been fully implemented in practice (Radin & Rogoway : 1965).

Parts of a PL/1 structure, then, have to be specified explicitly. A whole structure can be compared with another whole structure, if the two are morphologically identical: thus informants' profiles could be compared as <u>wholes</u>. However, operations on parts of structures require the specific

naming of the part of the structure. Therefore, the percentage conversion program could only operate on an informant's data stored in structure form if each OU is specified and processed individually. This is not economical from a programming viewpoint. An alternative strategy (described below p.96 - Prog. RAT) had to be found.

The original expectation that the structure facility of PL/1 could be usefully applied to the data processing problems of the T.L.S. was an expectation that could not be realised because the design of the coding frame failed to take account of these limitations on the manipulation of structures. A further difficulty pertaining to the use of structures relates to the PL/1/CLUSTAN interface (see below, p.105ff). Firstly, CLUSTAN operates on a 2-dimensional array of cases (informants) x variables. Data stored and processed in a PL/1 structure would have to be reduced to 2-D array form for input to CLUSTAN.

Secondly, constraints on CLUSTAN 1A as implemented at NUMAC include a limit on the maximum number of variables. No more than 200 'numeric' variables, or 400 'binary'[FN] variables can be processed in one CLUSTAN run.

FN. In CLUSTAN terminology, 'numeric' variables take continuous arithmetic values, either integer or decimal; and 'binary' variables represent simply presence/absence of features. An example of a numeric variable is falling tone as a percentage of all tones realised; an example of a binary variable is, SEX, where MALE is coded 1, and female, Ø.

Thus, the information contained by the hypothetical structure of Fig.4 (p. 44) could not be processed in one CLUSTAN run. The segmental phonological substructure alone holds 690 values, one for each state type.

The 3-alpha matrix, representing co-occurrences of 8 intonational features, with 9 grammatical form classes, with 7 pitch range features, yields a total of 504 combinations. The 2-alpha array holds 188 variables. These last 2 data collections (2-alpha and 3-alpha) are the basis for the calculation of various ratios, but the final number of variables derived from them is still quite large.

The segmental phonological data, then, cannot be processed altogether in one CLUSTAN run, but must be split up into several batches of variables. (Reasons for treating the segmental data as numeric rather than binary are given below, p.117ff.).

This separation of the segmental variables into several batches produces some interesting results (see below, Ch. 6), as several classifications of the sample are generated, each based on a different sub-set of variables.

To summarise, the reasons underlying the choice of PL/1 as the programming language to use included:

1. the sophistication of the language; its powerful data manipulation techniques, built in functions, especially string handling facilities;
2. its powerful array manipulation features;
3. the structure facility.

As explained in the foregoing, the structure facility proved to be less useful than was anticipated. Limited use, is, however, made of structures. (e.g. see VAL1, below, p.79).

Regarding string handling features: some use is made of strings, however, the computer programming phase of the T.L. S. data analysis involves mainly numerical data. Lamb (1965) recognises 3 categories of linguistic data processing (L.D.P.)

"1. processing of linguistic data (whether for linguistic or non-linguistic purposes) (LDP_1);

2. processing of data (whether verbal or not) for linguistic purposes (LDP_2);

3. linguistic processing of data (i.e. operating on data with linguistic processes.) (LDP_3)."

The T.L.S. involves activities covered by LDP_1 and LDP_2, but only the second type is automated.

The initial analysis and coding of data is related to LDP_1, but at this stage the human analyst is in charge, (e.g. auditory phonetic transcription) and at that stage the transcriptions are converted into numeric codings.

The programming stage, therefore, involves essentially arithmetic operations performed on numeric data. The automated processes do not involve handling data which is linguistic in nature, although it represents linguistic material. This is Lamb's LDP_2. Thus the advantages of PL/1 string handling facilities are not of central importance.[FN]

FN. Some use is made of PL/1 string manipulation, especially with 3-alpha data.

Regarding array manipulation: several of the programs which I have written take advantage of the PL/1 array manipulation facilities. Many features unique to PL/1, such as the INITial function, and the DEFINED[FN] function, also the asterisk function (see above, p.73), have been used, with the advantages of programming convenience, and compactness of program code.

FN. The INIT function assigns an initial value to a variable (or the contents of an array), e.g. zero, at the time of declaration. (Variables to be used in a program are declared at the beginning, by their identifier or name, and their data attributes are specified. A DECLARE statement announces the existence, of a variable, and causes storage space to be set aside for it.) Use of the DEFINED function avoids duplicating attribute specifications where one variable is part of (overlays) another. E.g. if a substring of a string is to be used as a variable, for instance, if the initial letter of a word is to be extracted and processed, a string variable is declared to hold the word, and another to hold the initial letter of the word; the latter variable can be declared as DEFINED by the former variable.

Other features of PL/1 proved to be useful, for instance, the option of 'implicit declaration.' This is one example of the exploitation of PL/1 default mechanisms. In the absence of explicit specifications in the program, the PL/1 compiler takes certain default actions. Where a variable

is not declared explicitly, the first usage of that variable in the program triggers the compiler to recognise and define it. It is defined according to the default attributes associated with its form. For instance, an undeclared variable identifier whose initial character is in the range I, J...N, is defined by default as having the attributes: arithmetic, fixed point, binary, real, and precision of 15 binary digits (on an IBM machine).

If the initial character of the variable identifier is in the ranges A - H, or O - Z, then the variable is defined by default as arithmetic, float, decimal, real, with precision of 6 significant decimal digits (on an IBM machine).

Even where variables are explicitly declared, unspecified attributes are supplied by default. Providing the relevant rules for default specifications are borne in mind, the use of default specifications saves the programmer time and program code.

The following section describes the sequence of processes applied to the TLS segmental phonological data, and describes the operation of my programs in detail.

Fig.5(p.64) charts the sequence of processes applied to the segmental phonological data, from the phonetic analysis through to the derivation of variety clusters and associated diagnostic statistics.

The input data for each informant consists of a deck of punched cards bearing a continuous string of 5-digit codes. Each 5-digit code represents a token of a state type. The stream of codes represents the order of occurence of segments, through time, in the interview. Each informant's deck is read into disk files. MTS line files are used. (Line files consist of a series of numbered lines, each of which holds a punched-card image.)

The first programming task consists of searching each informant's file and accumulating the absolute number of occurrences of tokens of each state type.

Program VAL1

Program VAL1 performs this searching and counting operation. Fig. 9 (pp.81 - 84) reproduces part of the program VAL1 for convenience in the text. (The full specification appears as the first program in Appendix X.) The figure shows (lines 8 -10) the variables for use in the program, and (lines 13 - 15) how a structure called CODECOUNT is set up.[FN]

FN. In the full program listing Appx. X of VAL1, all lines of program code are numbered. Regrettably this is not reproducible in the extract in Fig.9.

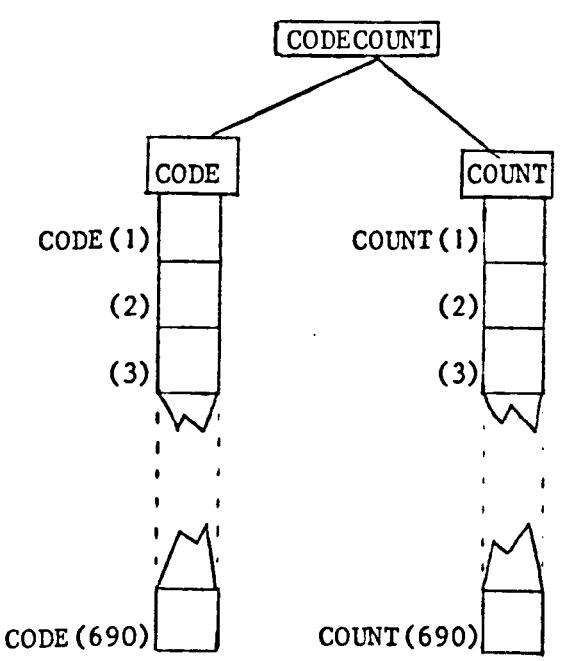
The structure CODECOUNT had two substructures, called CODE and COUNT respectively: each of these is a one-dimensional array consisting of 690 elements.

See Fig. 10.

Fig. 10.

The structure CODECOUNT

(Program VAL1)



Each element of the array CODE is assigned the value of one of the 5-digit codes; the 5-digit codes are assigned in ascending numeric order to the elements of array CODE from CODE(1) to CODE(690). This is performed by the assignment statements (lines 24 to 713 of the program listing). NB. '00021' appears as '21', as the PL/1 compiler removes non-significant zeros.

The array COUNT is used to accumulate the number of occurrences (tokens) of each 5-digit code type in an informant's raw data file.

COUNT(1) is the counter for the 5-digit code type stored in CODE(1), and so on. For every instance of the 5-digit code 00021 which is encountered in an informant's input file, COUNT(1) is incremented by 1.[FN]

FN. Text resumes on p.85.

Fig. 9.

```
/**************** VAL1 SEARCH PROGRAM*****************/

 VAL1:PROC OPTIONS(MAIN);
/*SEARCH FOR TOKENS OF 5-DIGIT CODES, ACCUMULATE TOTAL OCCURENCES */
/* 5-DIGIT STRINGS INPUT TO VARIABLE 'CHUNK' */

8   DCL NOTFOUND BIT(1),(X,CARD) FIXED BIN REAL,
9   (BOT,MID,TOP)FIXED BIN REAL,CHUNK FIXED BIN(17);
10  DCL NAME CHAR(8);

    /* STRUCTURE DECLARED TO HOLD 5-DIGIT CODE TYPES, AND FREQUENCY COUNT
13      DCL 1 CODECOUNT,
14          2 CODE(690) FIXED BIN(17),
15          2 COUNT(690)  FIXED BIN;
16   CODECOUNT.COUNT(*)=0;
     ON ERROR PUT DATA;

    /* ASSIGNMENT OF 5-DIGIT CODES TO CODE(N) FOR MATCHING */
    /* WITH INPUT STRINGS */


24  CODE(1)=21;
    CODE(2)=22;
    CODE(3)=23;
    CODE(4)=24;
    CODE(5)=25;
    CODE(6)=26;
    CODE(7)=41;
    CODE(8)=42;
    CODE(9)=43;
    CODE(10)=44;
    CODE(11)=45;
    CODE(12)=61;
    CODE(13)=62;
    CODE(14)=63;
    CODE(15)=64;
    CODE(16)=81;
    CODE(17)=82;
    CODE(18)=83;
    CODE(19)=84;
    CODE(20)=101;
    CODE(21)=102;
    CODE(22)=103;
    CODE(23)=121;
    CODE(24)=122;
    CODE(25)=123;
    CODE(26)=141;
    CODE(27)=142;
    CODE(28)=143;
    CODE(29)=144;
    CODE(30)=145;
    CODE(31)=161;
    CODE(32)=162;
    CODE(33)=163;
    CODE(34)=164;
    CODE(35)=165;
    CODE(36)=166;
    CODE(37)=181;
          .
          .
          .
          .
          .
          .
```

(most of the declarations, of which there are 690 for the elements of array CODE, are omitted.)

Fig.9. cont.

```
      :
      :
      :
      :
CODE(578)=2803;
CODE(579)=2804;
CODE(580)=2805;
CODE(581)=2806;
CODE(582)=2821;
CODE(583)=2822;
CODE(584)=2823;
CODE(585)=2824;
CODE(586)=2825;
CODE(587)=2841;
CODE(588)=2842;
CODE(589)=2843;
CODE(590)=2844;
CODE(591)=2845;
CODE(592)=2861;
CODE(593)=2862;
CODE(594)=2863;
CODE(595)=2864;
CODE(596)=2865;
CODE(597)=2866;
CODE(598)=2867;
CODE(599)=2868;
CODE(600)=2869;
CODE(601)=2881;
CODE(602)=2882;
CODE(603)=2883;
CODE(604)=2901;
CODE(605)=2902;
CODE(606)=2903;
CODE(607)=2921;
CODE(608)=2922;
CODE(609)=2923;
CODE(610)=2941;
CODE(611)=2942;
CODE(612)=2943;
CODE(613)=2944;
CODE(614)=2945;
CODE(615)=2946;
CODE(616)=2947;
CODE(617)=2948;
CODE(618)=4321;
CODE(619)=4322;
CODE(620)=4323;
CODE(621)=4324;
CODE(622)=4325;
CODE(623)=4341;
CODE(624)=4342;
CODE(625)=4361;
CODE(626)=4362;
CODE(627)=4363;
CODE(628)=4364;
CODE(629)=4381;
CODE(630)=4382;
CODE(631)=4383;
CODE(632)=4384;
CODE(633)=4385;
CODE(634)=4386;
CODE(635)=11201;
CODE(636)=11202;
CODE(637)=11203;
```

Fig. 9. cont.

```
CODE(638)=11204;
CODE(639)=11401;
CODE(640)=11402;
CODE(641)=12601;
CODE(642)=12602;
CODE(643)=14001;
CODE(644)=14002;
CODE(645)=14003;
CODE(646)=14004;
CODE(647)=14401;
CODE(648)=14402;
CODE(649)=14601;
CODE(650)=14602;
CODE(651)=15001;
CODE(652)=15002;
CODE(653)=15003;
CODE(654)=15004;
CODE(655)=15201;
CODE(656)=15202;
CODE(657)=16001;
CODE(658)=16002;
CODE(659)=16003;
CODE(660)=16004;
CODE(661)=16005;
CODE(662)=17201;
CODE(663)=17202;
CODE(664)=17801;
CODE(665)=17802;
CODE(666)=17803;
CODE(667)=20001;
CODE(668)=20021;
CODE(669)=24210;
CODE(670)=27601;
CODE(671)=27602;
CODE(672)=27603;
CODE(673)=28610;
CODE(674)=28611;
CODE(675)=28801;
CODE(676)=28802;
CODE(677)=28803;
CODE(678)=28804;
CODE(679)=43801;
CODE(680)=43802;
CODE(681)=43803;
CODE(682)=56001;
CODE(683)=56002;
CODE(684)=56003;
CODE(685)=56004;
CODE(686)=56005;
CODE(687)=58001;
CODE(688)=58002;
CODE(689)=58003;
713     CODE(690)=58004;

716       X=1;
          CARD=1;

/* INPUT, AND OUTPUT, INFORMANT'S MNEMONIC */
GET LIST (NAME);                PUT LIST(NAME);
```

```
/* CONVERSION ON CONDITION USED TO SENSE END-OF-RECORD MARKER */
 ON CONVERSION BEGIN;
 CARD=CARD+1;
 X=1;   GOTO JUMP;   END;

/* 'JUMP' LOOP INPUTS DATA AND PERFORMS BINARY SEARCH */
/* TO INCREMENT APPROPRIATE COUNTER */
/* 3 POINTERS USED IN BINARY SEARCH -'BOT','MID', 'TOP'*/

 JUMP:DO M=1 BY 1;
 BOT=1;
 TOP=699;
 MID=1;
 IF X=31 THEN DO;
 X=1;
 CARD=CARD+1;
 END;

/* INPUT ONE 5-DIGIT STRING TO VARIABLE 'CHUNK' */
741    GET EDIT(CHUNK)(COLUMN(X),F(5));
 NOTFOUND='0'B;

/* TEST FOR END-OF-FILE MARKER */
/* SWITCH 'NOTFOUND' SET OFF UNTIL CODE IS MATCHED*/

IF CHUNK=-9999 THEN DO;X=1;GOTO FINISH;END;
748  ELSE DO K=1 BY 1 WHILE(NOTFOUND='0'B &TOP>=BOT);
      IF CHUNK=CODE(BOT) THEN DO;
            COUNT(BOT)=COUNT(BOT)+1;
            NOTFOUND='1'B;
            X=X+5;
            END;
      ELSE IF CHUNK=CODE(TOP) THEN DO;
            COUNT(TOP)=COUNT(TOP)+1;
            NOTFOUND='1'B;
            X=X+5;
            END;
         ELSE IF CHUNK=CODE(MID) THEN DO;
            COUNT(MID)=COUNT(MID) +1;
            NOTFOUND='1'B;
            X=X+5;
            END;
         ELSE IF (TOP-BOT)<2 & NOTFOUND='0'B THEN DO;
               X=X+5;
               GOTO JUMP;
               END;

/*POINTERS MOVED UNTIL CORRECT CODE IS CONVERGED ON */
            ELSE DO;
               MID=(BOT+TOP)/2;
            IF CHUNK>CODE(MID) THEN BOT=MID;
            ELSE IF CHUNK<CODE(MID) THEN TOP=MID;
            END;
      END;
776  END;

/* RAW FREQUENCIES OUTPUT IN ASCENDING NUMERIC ORDER BY CODE TYPE
/* OUTPUT DIRECTED TO SEPARATE FILES IN BATCHES OF */
/* 200 VARIABLES FOR PROCESSING BY CLUSTAN */



 FINISH:DO J=1 TO 200;
X=1; PUT EDIT (NAME)(COL(X),A(8));
PUT EDIT(COUNT(J),'.')(COLUMN(X),F(3),A(1));
X=X+4;
IF X>74 THEN X=1;
END;

X=1;   PUT EDIT(NAME)(COLUMN(X),A(8));
DO J=201 TO 400;
PUT EDIT(COUNT(J),'.')(COLUMN(X),F(3),A(1));
X=X+4; IF X>74 THEN X=1; END;

X=1;   PUT EDIT(NAME)(COLUMN(X),A(8));
DO J=401 TO 600;
PUT EDIT(COUNT(J),'.')(COLUMN(X),F(3),A(1));
X=X+4; IF X>74 THEN X=1;END;

X=1;   PUT EDIT(NAME)(COLUMN(X),A(8));
DO J=601 TO 699;
PUT EDIT(COUNT(J),'.')(COLUMN(X),F(3),A(1));
X=X+4; IF X>74 THEN X=1; END;
END;
```

Line 16 shows the use of the asterisk function. This statement:

CODECOUNT. COUNT(*) = Ø ;

initialises the value of each element of the array COUNT to zero. This is necessary because when storage space is set aside for the structure CODECOUNT, there may be values already stored which must be removed. Otherwise, the accumulation of the number of instances of 5-digit codes would produce spurious results.

It is not necessary to initialise the contents of array CODE, as the assignment statements (lines 24-713) give the correct values to elements of this array. (The process of assigning a value to a variable effectively overwrites the previous contents stored in that variable's storage space. The process of incrementing the value of a variable, however, adds on to the previous contents of the storage space associated with that variable.)

The program so far (lines 1-713) has set up the structure CODECOUNT, and allocated a unique five-digit code type to each element of the array CODE. Each element of the array COUNT has the value zero. The processes of inputting an informant's file, and counting the number of tokens of each state type found therein occupies lines 716-776 on the program listing.

Two strategies could be adopted. As mentioned above (p. 78), the stream of 5-digit strings on an informant's deck of punched cards (and therefore, in his input file), represents the order of occurrence of segments through time during the interview. One strategy which suggests itself as a method of counting the number of instances of each code type is to take each code type in turn, and search through the file for instances of it. This would be very uneconomical, as the file would have to be input, and searched, 690 times, once for each code-type.

The strategy used in VAL1 involves a search of the array CODE, rather than a search of the input file.

The first five-digit code token in the file is input to a variable called 'CHUNK'. (line 741) This input string is matched with values stored in the array CODE. When the match is successful, the correct code has been

identified, CODE(n) and the corresponding counter COUNT(n) is incremented by 1. The process is then repeated with the next 5-digit string in the input file.

The simplest way to program this search is to attempt to match the input string with the contents of CODE(1), and then CODE(2), and so on, until the match is successful. This is a 'sequential' search. However, this is not an efficient method of programming. If, for the purposes of the argument, we make the assumption (which is actually untrue) that all 5-digit code types occur with the same relative frequency, then for each input string in the informant's file there will be, on average, 690/2 = 345 attempted matches, before the input string is successfully matched with the correct cell of the array CODE.

As each of the 52 informants processed so far has, on average, 2349 5-digit code tokens in their input file, a search based on this principle would require something like 2349 x 345 = 810405 matches per informant.

The strategy used is that of the 'binary search', also known as the 'split-half search'. The use of this technique depends on the list of items to be searched being arranged in ascending numeric order. (See above, p. 79).

A given input value is compared with the item in the middle of the list: if the input value has a greater numerical value than this item, then the search can be restricted to the second half of the list. If the input value has a smaller numerical value, the search can be restricted to the first half of the list. If the input value has the same numerical value, then a successful match has been made.

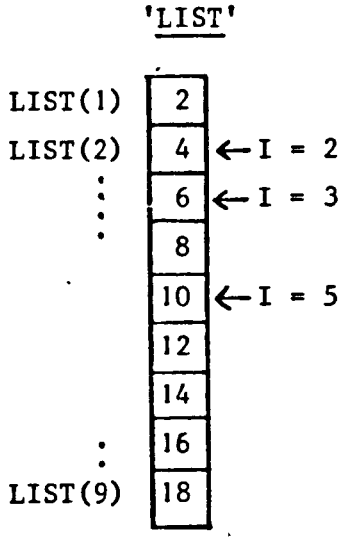
If, e.g., this input value is smaller than the middle item in the list, the process is repeated on the first half of the list. A match is attempted with the item half way through the first half of the list: if the match is unsuccessful the range of the search is restricted to half of the first half of the list. This process of halving is continued until the correct

item is converged upon.

For example, the first nine numbers of the series of even numbers can be stored in a one dimensional array called LIST. If we want to attempt to match the value stored in a variable N with this sub-set of the series of even numbers, in order to determine whether the number stored in N is a positive even number, less than 20, we can do this by means of a binary search.

LIST is set up as an array of 9 elements: LIST(1) = 2, LIST(2) = 4, etc. A 'pointer' variable, I, is set to the value 5, and thus LIST(I) points to the middle item in the list. See Fig.11.

Fig.11.



The value of N, (4), is compared with the value showed in LIST(I). As it is smaller, the search is restricted to the first half of the list, and the pointer is moved to the middle item of the first half of the list (I = 3).

Once again, the first half of this section of the list is identified as the search domain , and when the pointer is moved to the new mid-point (I = 2), the input item (N = 4) is matched with the contents of the array element. (LIST(2) = 4).

This technique is used in VAL1 (lines 748-776) to match each 5-digit code input with the 5-digit code types stored in the array CODE.

Three pointers are used: 'BOT', 'MID', and 'TOP'. BOT has the initial value 1, TOP, 690, and MID = TOP + BOT/2. (It is necessary to have 3 counters, as each new value taken by MID is calculated as half of the sum of the current values of TOP and BOT, which define the highest, and lowest, points in the sub-list currently being looked at.)

This binary search technique is more efficient than a sequential search through each element of the array in turn. (See above, pp.85 - 86). For each 5-digit token input, the DO loop (lines 748-755) is executed no more than 9 times. If the loop is executed, on average, 5 times for each 5-digit string input, then the expected average number of passes through the loop, per informant, is 11,745. (Cf. the expected average of 810,405 iterations with a sequential search - see above, p.86).

The binary search is executed for each 5-digit string input from the informant's data file. When the match is made, the counter in array COUNT, corresponding to the element of array CODE which the pointer points at, is incremented by 1. When all input has been processed in this way, the array COUNT holds the raw total number of occurrences of each CODE type.

The list now held in array COUNT is output to four separate files, for processing by CLUSTAN. (See above, pp.56 - 61 and below pp.104ff.).

It has already been mentioned (p. 52) that provision was made for 'new' states, and PDV's, to be incorporated into the coding frame, when and where an item appears in the data for which there exists no coding category.

For example, the original specification for OU46, ŋ , involved one PDV, with 6 states:

(ŋ , ŋg, ŋ̆, n, ng, nk).

During linguistic analysis of the tape recordings, another realisation of this item occurred: viz. [k] . A seventh state, k, therefore, was added to the state list associated with this OU. An extra five-digit code was created for this state, expanding the state list

from 02761 through 02766,

to 02761 through 02767.

This addition does not disrupt the sequential ordering of codes; however, when PDV's are inserted, the logic of the numeration of the codes is disrupted. This poses a computational problem.

The five-digit codes were designed so that the first four digits represent a unique PDV, and the fifth digit specifies the state of the superordinate PDV with which a segment is realised. (See coding frame specifications above, p. 37) PDV codes are arranged in ascending numeric order in increments of two. (Even numbers only. The first PDV is coded 0002, the second 0004....).

This convention (even numbering) was adopted with the use of Lector machine readable sheets in mind. Thus, when a new PDV was inserted into the coding frame, the odd number in between the two existent PDV codes was not used . Instead, a left-shift was used. For example, a new PDV value was invented to cope with realisations encounted in the data which merited the inclusion of an extra PDV associated with OU j This new PDV /j/[FN], was inserted between PDVs coded 0286 and 0288 respectively. Rather than creating a code of 0289 in between, the new PDV was given the code 2880.

FN. Actually, j / alveolar $\begin{Bmatrix}\text{stop}\\\text{fricative}\end{Bmatrix}$ —— u...
i.e. j, preceded by alveolar stop or fricative, followed by u, "which have a potential for affricationor deletion." (Pellowe, Nixon & McNeany, 1972a p.29).

In the event, the Lector was not used, so it would have been feasible (although this is only clear in retrospect) to use the intervening odd number as the code for newly inserted PDV's.This would not have presented any problems for the binary search program. However, as noted above (p.86), the binary search strategy requires that the values of items in the list to be searched must be in ascending numerical order. Because of the policy adopted of left-shifting the PDV digits by one place, the inserted PDV code is numerically greater than the preceding PDV code by a factor of

10; for example, the sequence of PDV's 0286, 2880, 0288. Thus, in the assignment of values to elements of the array CODE in program VAL1 (see above p.79) the five-digit codes associated with the inserted PDV's appear at the end of the list, and not in sequence with the codes belonging to the same OU as they do.

For example, the states of the inserted PDV discussed above, PDV 2880, must be placed lower down in the array CODE than the states of the other PDV's in the same OU,i.e. lines 698-701 of the program listing of VAL1 would be more in place (in terms of the linguistic logic of the coding frame,) between lines 623 and 624. In the example of an inserted state (see above, p. 88), this disruption does not occur.

However, in two instances, there are more than 9 states in one PDV.

PDV 0242 has 10 states, and PDV 0286 has eleven states. As state zero was not used, only 9 digits (1-9) are available to refer to the state of a PDV. Only one digit position is allocated to the designation of the state in the five digit code, and where state 10 or state 11 of a PDV is concerned, this one byte[FN1] range is overflowed.[FN2]

FN1. A 'byte' is an 8-bit (binary digit) quantity, capable of holding one character (e.g. 1 decimal digit).

FN2. If the coding scheme suggested above (p.69) had been used, with a 2-byte state field, this problem would not have arisen.

In these two cases, the policy of left-shifting was again adopted, thus the codes of the state list of PDV 0286, for example, are: 02861 through 02869, 28610, 28611. Thus, with the five digit codes organised in numerically ascending order, there are not only displaced PDVs, but also displaced states of PDVs. Because output from VAL1 is later processed OU by OU (see below p.96 RAT) , this disruption of the linguistically logical sequence of codes presents computational problems. For the purposes of program RAT (see below, p.96) it is necessary to restore the

ordering of scores with respect to the OU scheme.

Another program was written to perform this reordering of the output of VAL1.

Program Shuffle

Program SHUFFLE[FN] takes the output from VAL1 for each informant, and reorders the list of state scores to correspond with the ordering of states, within PDVs, within OUs, i.e. (OU1, PDV1, state 1); (OU1, PDV1, state 2);.... etc.

FN. This program, and program ELIM are not reproduced here, as the (late) version of program RAT shown here combines the functions of these programs.

The reordering is performed by the application of a lengthy series of control variable specifications in a DO loop. This is a simple programming exercise: however all such minor reshufflings of the data increase the likelihood of programming errors. Anticipation of such technical considerations at the initial stage of coding frame design could have avoided this problem, (although such problems often only emerge clearly in retrospect.) The importance of close collaboration from the outset between linguists and programmers is highlighted by this issue.

Comparison of informants on the basis of their respective raw total scores (number of realisations of each state type) is not very useful. There are two reasons for this which I deal with at length, in turn:

i) individual informants varied in terms of their overall productivity (total number of segments realised during the interview). (Some informants are more loquacious than others).

ii) OU's have different relative frequencies of occurrence in the sample.

i) On average, each member of the sub-sample of informants dealt with here

(52 Tyneside speakers), produced 2349 segments during the interview. However, the least productive speaker (ELIOT) and the most productive speaker (MARSH) produced 1084, and 3571 segments respectively. Even if these 2 speakers had identical linguistic profiles (i.e. their realisation of sound features were distributed across the same states, in the same proportions) MARSH's raw scores would exceed ELIOT's by a factor of 3. In order for their linguistic profiles to be compared, some kind of standardisation technique must be applied to the raw data, in order to compensate for informants' variable productivity.

Standardisation Statistic - S_1

This problem can be overcome by transforming raw scores with reference to the population mean value for total number of segments elicited per interview. The following formula would achieve this standardisation:

$$\text{score}(Uij)_{S_1} = \text{Score}(Uij)_r \times \frac{\text{popn. mean total segments}}{\text{total segments realised } i}$$

for the ith informant, on the jth variable, (variables being state scores), where

$(Uij)_{S_1}$ = indiv. i's standardised score on jth variable by 'standardisation statistic.'

and

$(Uij)_r$ = indiv. i's raw score on jth variable.

This formula multiplies all one informant's state scores by a constant multiple (i.e. a monotonic transformation.) Comparability between individuals in terms of their relative overall productivity is thus ensured.

However, this measure accounts for variation only at the level of representation of the state, without reference to the superordinate structure of OU's and PDV's, of which state scores form the terminal elements (see above, p.67).

OU's are, in a sense, the variables, and the list of states associated with each OU is the paradigm of its variants, grouped under the intermediate level of structure, the PDV level.

However, it is not simply a question of which states an informant uses to realise an abstract phonological entity (represented by an OU) but in what relative proportions he used the states which he uses.

As Labov (1966) points out, (p.129 f.), speakers' usages of variant realisations of phonological features are contrastive in terms of relative frequency of occurrence, and not in terms of "all-or-none" signals. Whether or not we give assent to Labov's notion of style, his further observation is pertinent:

> "whether or not we consider stylistic variation to be a continuum of expressive behavior, or a subtle type of discrete alternation, it is clear that it must be approached through quantitative methods..."

Thus, states are themselves quantitative variables, as well as being variants of the superordinate variables, OU's.

A measure of normalisation to be applied to state scores, then, must normalise with respect to the internal structure of the OU.

The 'standardisation statistic' was therefore rejected in favour of the within-OU percentage ratio.

Within OU-percentage ratio

The raw score for each state is converted to a percentage of the sum total of occurrences of state variants within the superordinate OU. This gives a picture of the proportional distribution of state variants used for each OU.

$$(Uij)_{S_2} = (Uij)_r \,/\, \sum_{n=1}^{m} (Uin) \times 100$$

where there are m states in the superordinate OU, $(Uij)_r$ is informant i's raw score on state j.

Because raw scores are transformed to percentages, differential overall productivity between informants is automatically compensated for.

Regarding the second reason for transforming raw scores, the within-OU percentage ratio measure represents one solution to the difficulties raised here.

ii) OU's have different relative frequencies of occurrence in the sample.

Figure 12 shows the relative proportions (by percentage) with which the sum of variant states of each OU are represented in the corpus. (Data analysed for 52 informants). (OU's 1-51 are shown, from left to right their specifications appear in Appx. A.)

8.10% of the speech segments recorded (in total the 52 informants produced 122184 segments) are phonetic variants of the abstract phonological entity, OU29, t. This phonological entity has the highest relative frequency of all the OU's. Three OU's are tied in the lowest position, with a relative frequency of 0.05%. These are the OU's (ʊə, ə4b and aɪə).

In the total corpus, then, the proportion of segments coded under e.g. OU ʊə , to those coded under OU t, is 8.10:0.05 = 1:162.

As this inequality, in terms of frequency, is reflected in the scores at state level, we find that the population mean frequencies for states of OU t are very much large/numerically than scores on states of OU ʊə. This means that the range of absolute numeric values is much larger for states of OU t , than, say, for states of OU ʊə.

The distance coefficient used in the CLUSTAN classification[FN] computes distance between individuals with respect to their scores on a give state using the square of the difference between their scores.

FN. The coefficient used here is Squared Euclidean Distance:

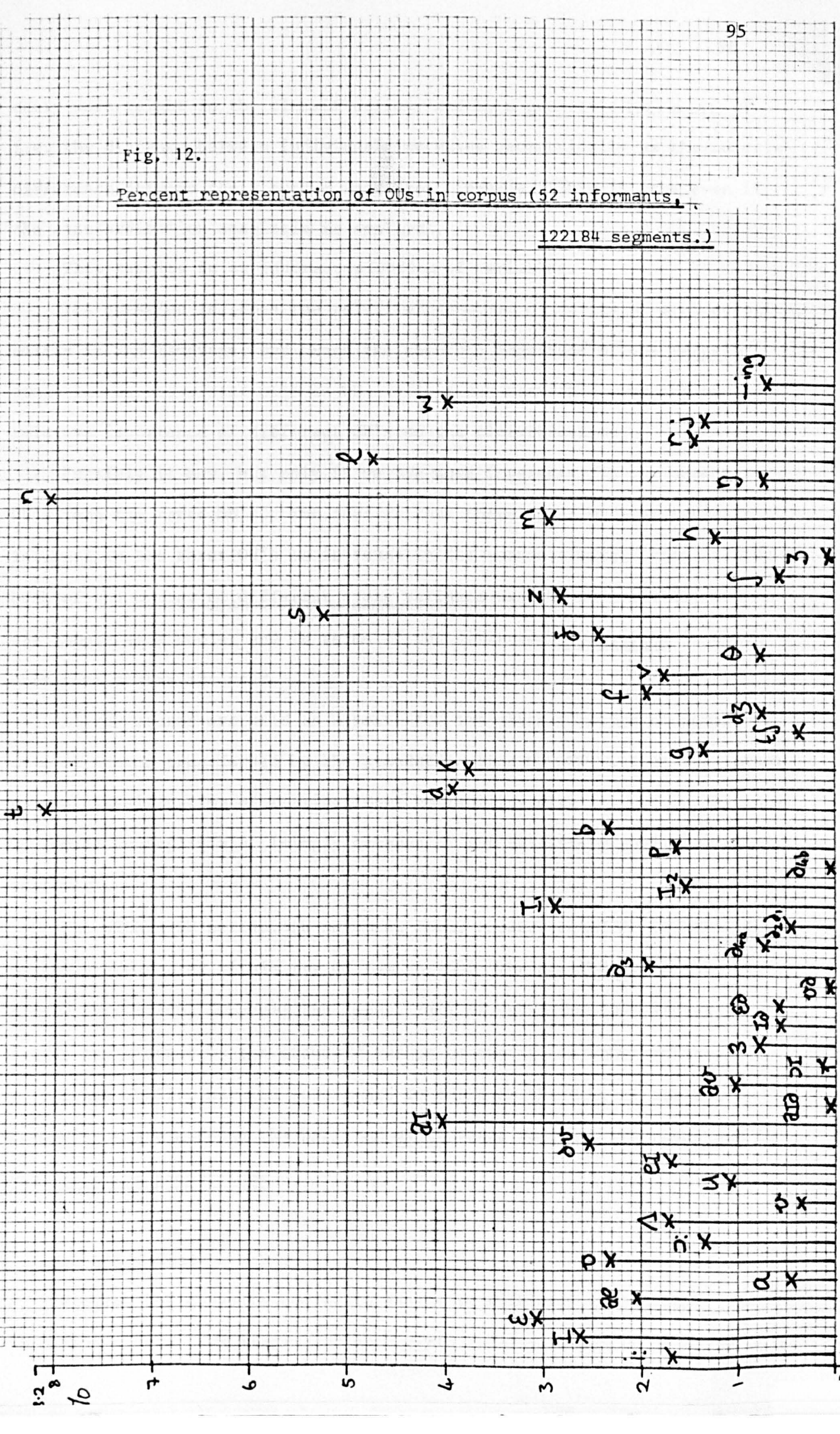
$$D^2_{pq} = \frac{1}{M} \sum_1^m (Ujp - Ujq)^2$$

where p,q are informants (cases), compared on a total of M variables (state scores) and U is the score on the jth variable.

This means that states whose range of values across the population is larger

Fig. 12.

Percent representation of OUs in corpus (52 informants, 122184 segments.)



carry more weight in the classification, and by a geometric increment : i.e. the distance between individuals increases geometrically as the numeric difference between their scores increases linearly. (This is true even if the standardisation statistic is applied).

This results in an inherentweighting of variables.

The question is, does the relative commonness or rarity of a phonological entity in the language, and hence the absolute magnitude of the range of state scores associated with an OU, reflect any desirable ranking of the relative importance of different OU's to the classification?

Possibly it may, but only if we are more interested in the commoner OU's than the rarer OU's. The fact is that there is as much variability internal to the rarer OU's as is found in the commoner ones. This is true both of the number of different state variants used by individuals, and by the sample as a whole; and in terms of the different proportions in which different states are used.

If raw scores are used as a basis for the classification of speakers, or even scores transformed by the standardisation statistic, then the impact on the classification of realisational variability with respect to the rarer OU's will be negligible, (simply because of the relatively small numbers involved).

The within-OU percentage measure, then, ensures equal weighting of OU's, as the state scores within one OU sum to 100. The possible range for all states is 0 - 100, (whereas the upper limit of unstandardised scores, or scores transformed by the standardisation statistic, is indeterminate).

Comparability between informants is based on the unit of the OU. Comparisons are made on the basis of which state variants are used, and in what relative proportions.

Program RAT (RATio)

Program RAT converts raw state scores to within-OU percentages. Each informant's data (raw scores for each state type) is input to an array

called 'CODE'. Another array, called 'N' stores the number of states in each OU.[FN] (Program RAT is given in Fig. 12A and Appx. X.)

FN. Because the data is not held in a structure which reflects the morphology of the coding frame, the structure of the data has to be embedded in the programming, as in the DO loop which operates under the control of the contents of array N, (line 57 of program RAT) which specifies the extent of each OU, in terms of state scores in the array CODE.

A variable called OUT (OU-total) is used to hold the result of summing all the scores for states within one OU. The loop 'NEWOU' is executed once for each of the 52 OU's, and this loop performs the conversion of each raw state score to a within-OU percentage score. Lines 56-68 of the program listing show the application of the within OU-percentage transformation formula shown above, (p. 93).

(Lines 34-51 of this version of program RAT actually perform the reordering operation described above (p.91), under program SHUFFLE).

The percentage frequencies are stored in an array called STATE. Thus if a given informant always realises the abstract phonological entity i: (OU1), with the third state of the first PDV of that OU, PDV i:,state i, (CODE (3) = 00023), then the third element of the array STATE will hold the value 100. (100% of instances of segments coded under OU1 are realised by state 3).

If, however, that informant distributes his realisations across the first three states in the proportions 2:1:1, then

STATE (1) = 50

STATE (2) = 25

STATE (3) = 25

Thus the actual magnitude of the raw frequencies does not skew the results, and the problems associated with different speakers having different overall productivity, and OU's being variably represented, are by-passed.

Output consists of a list of state scores, as within OU percentages,

```
/*** PROGRAM VAT - WITHIN OU RATIOS COMPUTED ***/

VAT: PROC OPTIONS (MAIN);
ON ENDFILE SYSIN DATA;

DCL NAME CHAR(3);
/* ARRAY 'CODE' SET UP TO HOLD SCORES FOR ALL 5-DIGIT CODE TYPES */
DCL CODE(690) FIXED BIN;

/* ARRAY 'N' STORES NUMBERS OF STATES IN EACH OF THE 52 OU'S*/
DCL N(52) FIXED BIN INIT(15,16,23,12,10,20,16,22,
11,12,32,24,15,10,29,9,22,14,9,7,6,6,5,5,5,2,18,12,
21,15,22,10,5,7,5,3,6,5,4,2,2,2,4,3,6,6,21,3,10,6,3,15);
DCL STATE(690) FLOAT;

/* 'OUT' (OU-TOTAL) HOLDS SUMMED RAW SCORES FOR ALL STATES */
/* WITHIN EACH OU */
/* FOR CALCULATION OF WITHIN-OU % REPRESENTATION PER STATE */
DCL OUT FLOAT;

/* CASE LOOP */
  DO CASE=1 TO 52;
  L=2;

/* INPUT RAW SCORE (AS OUTPUT FROM VAL1) */
  GET EDIT(NAME)(COL(1),A(3));
    DO I=1 TO 690;
    GET EDIT (CODE(I))(COL(L),F(4,0));
    L=L+4;
    IF L>74 THEN L=2;
    END;

/* DO LOOP RE-ORDERS OUT-OF-SEQUENCE CODES */
/* AND COPIES RE-ORDERED LIST TO ARRAY 'STATE' */
34    DO I=1 TO 5,8,10,16 TO 19,23 TO 37,39,40,42,44 TO 53,
      55 TO 63,65 TO 72,75 TO 79,85,86,89 TO 94, 99, 101,
      102,104 TO 111,114,116 TO 118,121 TO 124,682,683,686,120,
      638,128,130,133 TO 143,149,152,155,156,158 TO 169,175 TO
      182,185 TO 195, 197 TO 202,204,206 TO 209,211,216 TO 224,
      227,231 TO 234,635 TO 638,235,236,639,640,237 TO 240,
      242 TO 254,261 TO 264,641,642,265 TO 273,275 TO 283,286
      TO 289,643,644,646,290 TO 298,647,648,299 TO 302,649,650,
      303 TO 310,651 TO 654,311,313,314,655,656,315 TO 317,319,
      323 TO 326,328,661, 329 TO 334,336,338 TO 346,348 TO 351,
      652,354,355,357 TO 360,664 TO 666,363 TO 374,376,380 TO
      409,667,668,410 TO 496,669,497 TO 502,505 TO 533,
      535 TO 546,548 TO 559,561 TO 565,567 TO 574,576 TO 580,
      582 TO 597,599 TO 600,674 TO 676,601 TO 604,607,609 TO 614,
      617, 670 TO 672,618,619,621,623 TO 634;
      STATE(J)=CODE(I);
      J=J+1;
51    END;

/* NEWOU SEGMENT SUMS TOTAL STATE SCORES PER OU */
/* AND CONVERTS EACH STATE SCORE TO WITHIN-OU %AGE */
  K=0;
56 NEWOU:     DO I=1 TO 52;
57    M=N(I);
      OUT=0;
        DO J=K+1 TO K+M;

        OUT=STATE(J)+OUT;
        END;
        DO J=K+1 TO K+M;
          IF STATE(J)¬=0 THEN DO;
          STATE(J)=STATE(J)/OUT*100;
          END;
        END;
      K=K+M;
68    END;

    J=1;
      DO I=1 TO 690;
      PUT EDIT(STATE(I))(COL(J),F(5,1));
      J=J+5;
      IF J>75 THEN J=1;
      END;
    END;
77 END;
```

for each informant.

This collection of data can be thought of as a 2-dimensional matrix, where each column represents a state type , and each row, an informant. Thus row 1, column 1, contains the first informant's score on the first state: row 1 column 2 contains the same informant's score on state 2, and so on. (See Fig. 6, p.66.)

This matrix holds the data in a form which can be input to CLUSTAN, for a classification of informants.

Before cluster analysis is performed, however, there are several strategies which can be used to improve the classificatory scheme. Two of these are described here.

It is possible that some of the variables (state scores) to be input to CLUSTAN do not, in fact, vary across the particular sample of speakers which is here dealt with. Any variables which have zero variance across this sample are redundant, and will inhibit the discrimination of groups (clusters).

That is, the discriminating effect of any one variable which <u>does</u> display variation across the population is an inverse function of the total number of variables included in the classification. If many redundant variables are included, the discriminating effect of truly varying variables is proportionally reduced, and the measure of similarity between individuals is artificially inflated.

Therefore, it is advantageous to identify, and eliminate, zero variance variables before the classificatory process is applied.

A second source of classificatory distortion occurs if variables are interdependent, or correlated. If two, or more, variables are logically interdependent, then they will always co-occur. They will cluster together <u>as a group</u> of variables, and will perhaps distort the classification, by inflating the similarity levels between all pairs of individuals which have realisations of them.

In order to identify zero-variance variables, a program was written to compute sample statistics, including means, and standard deviations, for each variable (i.e. each state). This is program PROF, which stand for profile, and the output constitutes a profile of the sample with respect to all states.

In order to discover dependencies between variables, an R-analysis was planned. In an R-analysis, variables are clustered, rather than cases. Where variables cluster tightly together, correlations or covariances between variables are evidenced: if this occurs, there may be grounds for eliminating dependent variables. In order to perform these two processes, identification of zero variance variables, and identification of correlations between variables, the 2-dimensional matrix was transposed. For an R-analysis, input consists of a list of values taken by each variable, across cases, rather than a list of values for each case, across all variables. Transposition of the matrix, then, involves turning it through ninety degrees, so that, instead of rows representing cases, and columns, variables, now rows represent variables and columns cases.

C.f. Fig. 13 with Fig. 6 - p.66 above).

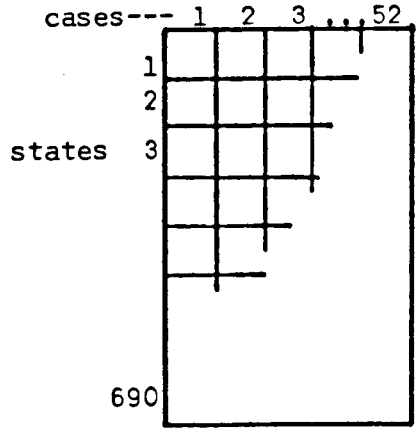


Fig. 13.

The transposed matrix.

This transposition of the matrix is performed by program POSE. (See program listing. Fig. 14 (p.101) and Appx. A.)
The transposition is achieved by inputting the matrix in row major order, and outputting in column major order. (i.e. (row 1, col. 1); (row 1, col.2)

Fig. 14.

```
/*** PRIMARY PURPOSE - TRANSPOSES MATRIX ***/

TABLE:          PROC OPTIONS(MAIN) ;
                DCL X FIXED BIN ;

/*2-D MATRIX SET UP TO HOLD SCORES */
/* CASE X STATE */
/* LOOP THROUGH 52 CASES */

                DCL MAT(690,52) FLOAT ;
                   DO J=1 TO 52 ;

/*SET COLUMN CONTROL FORMAT VARIABLE FOR INPUT */
                   X=2 ;
/* READ STATE SCORES INTO MATRIX */
                      DO I=1 TO 690 ;
                      GET EDIT(MAT(I,J))(COL(X),F(5,1)) ;
                      X=X+5 ;
                         IF X>76 THEN
                         X=2 ;
                      END ;
                   END ;

/* TRANSPOSE MATRIX */
                   DO I=1 TO 690 ;
/* SET COLUMN CONTROL FORMAT VARIABLE FOR OUTPUT */
                   X=1 ;
                      DO J=1 TO 52 ;
                      PUT EDIT (MAT(I,J))(COL(X),F(2)) ;
                      X=X+4 ;
                         IF X>76 THEN
                         X=1 ;
                      END ;
                   END ;
                END ;

FILE
```

and so on, and (row 1, col. 1); (row 2, col. 1)) (lines 11 and 16 of listing for input, lines 25 and 28 for output).

The transposed matrix is now input to program PROF.

Program PROF

One row of the matrix is input at a time. This row holds the scores on one variable (state scores) for each of the 52 cases. The sample mean, and standard deviation, is calculated for that variable.[FN]

FN. The program listing shown here (p.103) also calculates cluster means and standard deviations for each variable. This version of the program was run after the clustering programs had been implement, and cluster membership was known. The clusters, (shown here as 'X' and 'V', represent the 2 groups discovered by the linguistic classifications described below, ch. 6.(See below pp. 196ff.) taken at the 2-K (2-cluster) level. Output from PROF appears in Appx. X.

Output from program PROF showed that 113 of the 690 states were zero-variance variables. These were eliminated from the classification.

A CLUSTAN R-analysis was then attempted, using the transposed matrix as input. In an R-analysis, variables are treated as 'cases', and cases as 'variables', as far as the program is concerned.

The implementation restrictions on the data file input to CLUSTAN require that no more than 200 variables, and 999 cases, be processed.

For the R-analysis, the 690 variables are treated as 'cases', and the 52 cases as 'variables'; thus it is possible, in theory, to cluster all the variables in one run. For technical reasons two different attempts to compute these dependencies both failed.[FN]

FN. The temporary data files created by CLUSTAN to hold intermediate data sets exceeded the NUMAC system maximum file size. That is, the MTS maximum limit (for line files) of 255 disk pages was overflowed. 1 disk page = 4096 bytes, 1 byte = 8 bits, 1 disk page can hold 4096 characters. An attempt was made to specifically create the file (normally created automatically by CLUSTAN), and to create it as a sequential file (sequential files have larger maximum size limits) but this also resulted in program failure, as CLUSTAN performs indexed operations on this file, and sequential files can

(Text resumes at p.105.)

```
/***** PROGRAM PROF - SUMMARY STATISTICS - POPULATION PROFILE *****/
PROF:       PROC OPTIONS(MAIN) ;
/* COMPUTES      MEANS, SD'S , FOR ALL VARS, ALSO MEANS, */
/* SD'S FOR TWO MAJOR CLUSTERS, DESIGNATED 'V', AND 'X' */
            DCL (VSUM,XSUM,MEAN,VMEAN,XMEAN,SUMSQDIF,SD,VSSD,VSD,XSSD,XSD
) FLOAT ;
/* ARRAY SET UP TO HOLD SCORE ON STATE(I) FOR CASE(K) */
            DCL AR(52) FLOAT INIT((52)0) ;
/* PRINT HEADINGS */
            PUT EDIT('STATE','POPN MEAN','X MEAN','V MEAN','SD','X SD','V
 SD')(LINE(1),COL(1),A(5),COL(10),A(9),COL(20),A(6),COL(30),A(6),COL(40)
,A(2),COL(50),A(4),COL(60),A(4)) ;
/* LOOP THROUGH EACH STATE IN TURN */
                DO I=1 TO 690 ;
                VSUM=0 ;
                XSUM=0 ;
                MEAN=0 ;
                VMEAN=0 ;
                XMEAN=0 ;
                SUMSQDIF=0 ;
                SD=0 ;
                VSSD=0 ;
                VSD=0 ;
                XSSD=0 ;
                XSD=0 ;
                N=2 ;
/* INPUT SCORE FOR EACH CASE */
                   DO K=1 TO 52 ;
                   GET EDIT (AR(K))(COL(N),F(4,0)) ;
                   N=N+4 ;
                      IF N>76 THEN
                      N=2 ;
                   END ;
/* COMPUTE POPN. AND CLUSTER MEANS, ON EACH STATE */
/* CLUSTER V HAS 45 CASES, CLUSTER X HAS 7 CASES */
                   DO J=1 TO 45 ;
                   VSUM=VSUM+AR(J) ;
                   END ;
                   DO J=46 TO 52 ;
                   XSUM=XSUM+AR(J) ;
                   END ;
                MEAN=(VSUM+XSUM)/52 ;
                XMEAN=XSUM/7 ;
                VMEAN=VSUM/45 ;
```

Fig. 15.

```
STDEV:                  DO J=1 TO 52 ;
                        SUMSQDIF=SUMSQDIF+(AR(J)-MEAN)**2 ;
                        END ;
                    SD=SQRT(SUMSQDIF/52) ;
VINSD:                  DO J=1 TO 45 ;
                        VSSD=VSSD+(AR(J)-VMEAN)**2 ;
                        END ;
                    VSD=SQRT(VSSD/45) ;
NIXSD:                  DO J=46 TO 52 ;
                        XSSD=XSSD+(AR(J)-XMEAN)**2 ;
                        END ;
                    XSD=SQRT(XSSD/7) ;
/* OUTPUT TABLE: STATE NUMBER; POPN. MEAN; CLUSTER MEANS; */
/* POPN. SD'S, CLUSTER SD'S */
                    PUT EDIT(I,MEAN,XMEAN,VMEAN,SD,XSD,VSD)(COL(1),F(3),COL(10
),F(6,2),COL(20),F(6,2),COL(30),F(6,2),COL(40),F(5,2),COL(50),F(5,2),COL


(50),F(5,2)) ;
                END ;
/* END STATE LOOP */
            END ;

FILE
```

Fig. 15 cont.

only be accessed sequentially. 'Indexed' read, or write, operations involve a line number specification in the I/O subroutine call. Thus the CLUSTAN read operation must access a line file, and the restriction of maximum file size = 255 pages holds.

Thus the pathway shown on the flowchart, (Fig. 5 , p.64) from POSE through CLUSTAN R-analysis was not successfully implemented.

So input to CLUSTAN consists of informants' state scores (transformed to within-OU percentages), on 577 of the 690 original state types. Zero-variance states are eliminated, but the attempt to identify correlated variables was not successful, therefore no variables were eliminated on the grounds of mutual dependency.

In the CLUSTAN runs discussed here, (see below, ch. 6), OU52 was not included, as this OU is not strictly a segmental phonological variable, but a lexical item (OU52 covers variant realisations of the lexical item 'yes', (See Appendix A).

With this OU eliminated, the segmental phonological variables input to CLUSTAN number 542. As a maximum of 200 variables (of the 'numeric' type, (see FN. above, p.75) can be used in a CLUSTAN run, the data was split up into 3 sections, corresponding to

1) monophthongal vowel OU's, OU1 - OU10,

2) diphthongal, triphthongal and reduced vowel OU's, OU11 - OU26,

and 3) consonantal OU's, OU27 - OU51.

These three segmental phonological subspaces are designated %FON1, %FON2, and %FON3 respectively, and account for 154, 189 and 199 states respectively.

The output for classifications of informants in each of these subspaces is the result of taking a path through the flowchart (Fig.5, p.64), from VAL1, through SHUFFLE, RAT, POSE, PROF, and into CLUSTAN, from a Q-analysis. (i.e. informants are clustered on the basis of their scores on states.)

The sequence of CLUSTAN programs used was:

1. FILE
2. CORREL

3. HIERAR

4. RESULT

5. PLINK

(A full description of these programs is found in NUMAC Document 37: Cluster Analysis in MTS: CLUSTAN 1A User Manual).

1. The CLUSTAN progam FILE reads in the data, and sets up a data file for the succeeding programs to process. Permitted maxima are: number of cases $\leq$ 999; number of numeric variables $\leq$ 200; number of binary variables $\leq$ 400.

2. Program CORREL computes the similarity (or distance) between all pairs of cases, and stores these similarity (or distance) measures in the similarity matrix. 38 differently defined coefficients are available for computing similarity or distances.

The coefficient used in the CLUSTAN runs on the T.L.S. data is Squared Euclidean Distance, (see FN. p.94).

3. Program HIERAR operates on the similarity matrix, and builds clusters of indiviuals, on the basis of the mutual similarity levels between pairs. Initially, for n cases, there are n(n-1)/2 pairs. The most similar pair (those with the highest measure of mutual similarity) are fused, and thereafter treated as one case. The similarity matrix is correspondingly shrunk by one row, and all the similarity measures are recomputed. This process of fusion and recomputation of similarity measures continues for n-1 cycles (i.e. until all cases are fused into one cluster).

Of the 8 clustering methods available under HIERAR, Ward's algorithm was used in the CLUSTAN runs dealt with here (Ward: 1963). Everitt (1974) describes this method as follows:

> "Ward (1963) proposes that at any stage of an analysis the Dss of information which results from the grouping of individuals into clusters can be measured by the total sum of squared deviations of every point from the mean of the cluster to which it belongs. At

each step in the analysis, union of every possible pair of clusters is considered and the two clusters whose fusion results in the minimum increase in the error sum of squares are combined." (p.15)[FN].

FN. WARD'S ALGORITHM
If clusters P and Q were fused, then the similarity between any cluster R and the new cluster (P+Q) i.e. S(R,P+Q) is obtained from this formula:

$$S(R,P+Q) = \frac{NR+NP}{NR+NP+NQ} \times S(R,P) + \frac{NR+NQ}{NR+NP+NQ} \times S(R,Q) - \frac{NR}{NR+NP+NQ} \times S(P,Q),$$

where N stands for the number of members in a given cluster, so that NP is the size of cluster P.

---

4. Program RESULT prints out a summary of the classification process, and classification arrays from selected points in that process, (e.g. the 3-cluster level.) Print Options which can be selected include; printout of raw data, printout of similarity matrix, printout of selected classification arrays (which show cluster membership at given levels of the fusion process), and listing of cluster diagnostic statistics on variables.

5. Program PLINK takes output from program HIERAR, and draws a dendrogram, or fusion tree, on the graph plotter, which shows graphically the steps in the fusion process.

The results of the CLUSTAN runs on the segmental phonological data, and the social data are presented below (see Chs. 5 & 6).

The choice of distance coefficient, and clustering algorithm used was based on the results produced by a series of pilot runs on one section of the raw data: vocalic variables: CODE(1) to CODE(200), for the Gateshead subsample (45 speakers) .

Several combinations of CLUSTAN options for similarity coefficients and clustering algorithms were tried out on the raw data, and their performance was observed. Three of these runs are described here.

The similarity coefficient used first was CLUSTAN coefficient number 28, the Similarity Ratio measure.

This coefficient was originally chosen in the light of the problem of

'sparse matrices'. That is, the 2-dimensional matrix (cases x variables) contains a high proportion of empty cells. (Although 557/690 state types are used at least once across the whole sample of 52 speakers, each single speaker uses only 150 to 250 different state types altogether).

Thus, approximately 2/3 of the cells of the matrix input to CLUSTAN contain scores of zero.

These zeros represent states which are not realised at all by the informant in question. We may choose to regard these states as irrelevant variables, for that case. If this line is adopted, then when the distance or similarity measure between a pair of cases is computed, we want the value of the measure to depend more on the states which are realised (by one, or by both of the cases), than on those states which are not realised by either. Take, for example, the hypothetical situation where two cases are compared: each case realises 200 state types, but there is no intersection between the list of states used by the first case, and that used by the second case. We may want to say, seeing these two cases always use a different state realisation for a given phonological entity, that they are maximally different, as speakers. The similarity measure should reflect this difference. The (200+200) states actually realised by one or other of these cases will contribute to the overall measure of similarity by distancing there 2 cases from each other. However, for the remaining 290 variables (out of the total of 690), these two cases will have identical scores (i.e. each case scores zero.) These variables (shared zeros) will artificially inflate the degree of similarity between these two cases. (With quantitative data, the quantity zero is treated just like any other numeric quantity. Zero minus zero equals zero: which signifies zero distance (or maximal similarity) with respect to the variable in question, if these zero matches are not excluded by the coefficient used.)

The Similarity Ratio coefficient was chosen because zero matches on variables are discounted, the similarity between a given pair being computed

on the basis of only those variables which take a non-zero value for either one or both of the cases compared.

$$\text{Sim.Rat.} = \frac{\sum Ujp\ Ujq}{\sum U^2jp - \sum UjpUjq + \sum U^2jq}$$

where p,q are cases, and Uj is the score on the jth variable.

When Ujp and Ujq are both zero, no contribution is made to the sums of either the numerator or denominator. Thus summation occurs only over those variables which have a non-zero value for at least one of the pair.

Fig. 16 shows the dendrogram produced by CLUSTAN program PLINK, showing the fusion process which results from the application of the Similarity Ratio coefficient to the test data. (See above, p.108). The clustering method used here is Single Link or Nearest Neighbour (Sneath: 1957). With this technique, the criterion for a sample point (case) joining a cluster depends on the proximity of that sample point to any one member of the cluster. The newly joining sample point may be quite dissimilar to other members of the cluster, but the internal structure of the cluster depends on continuous inter-connectedness between adjacent points.

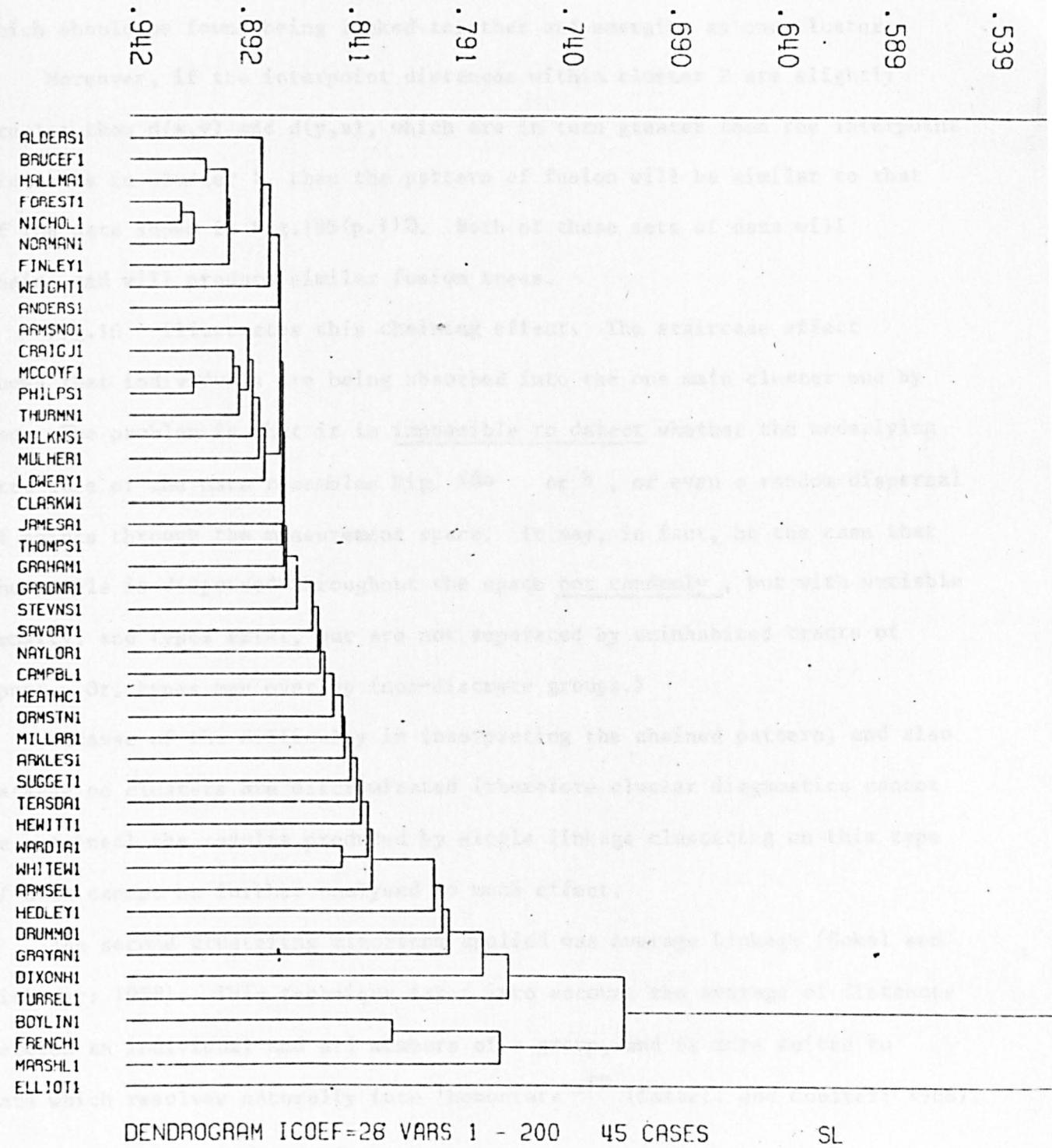
Fig. 17 (p.112) shows how sample points are joined into clusters by the single linkage method. (Clusters are depicted here in 2-dimensional space.) Members of clusters are 'chained' together by pair-wise links.

Note that point B is closer to a point in a different cluster (Point C) than to point A, which is in the same cluster. However, the distance between point B and point C is greater than the distance between any adjacent pair in either cluster.

Where the nature of the data corresponds to this type of 'straggly' cluster, or 'segregate' (Cattell and Coulter: 1966), then single linkage is an appropriate method for recovering the natural structure inherent in the data.

However, problems arise when clusters are not separated by interpoint distances larger than the maximum distance threshold which is the criterion

Fig. 16. Dendrogram based on Similarity Ratio Coefficient and Single Link Clustering.



for linkage of points.

Fig.18a(p.112) shows 2 clusters which ought to be separated, but which will not be discriminated by single link methods. The three 'outriders' or 'noise points' (Wishart: 1969b), x, y, z will result in the 2 clusters which should be found being linked together and emerging as one cluster.

Moreover, if the interpoint distances within cluster 2 are slightly greater than d(x,y) and d(y,z), which are in turn greater than the interpoint distances in cluster 1, then the pattern of fusion will be similar to that of the data shown in Fig.18b(p.112). Both of these sets of data will chain, and will produce similar fusion trees.

Fig.16 illustrates this chaining effect. The staircase effect shows that individuals are being absorbed into the one main cluster one by one. The problem is that it is impossible to detect whether the underlying structure of the data resembles Fig. 18a or b , or even a random dispersal of points through the measurement space. It may, in fact, be the case that the sample is dispersed throughout the space not randomly , but with variable density, and types exist, but are not separated by uninhabited tracts of space. Or, types may overlap (non-discrete groups.)

Because of the difficulty in interpreting the chained pattern, and also because no clusters are discriminated (therefore cluster diagnostics cannot be obtained) the results produced by single linkage clustering on this type of data cannot be further analysed to much effect.

The second clustering algorithm applied was Average Linkage (Sokal and Michener: 1958). This technique takes into account the average of distances between an individual and all members of a group, and is more suited to data which resolves naturally into 'homostats '[FN] (Cattell and Coulter: 1966).

FN. 'Homostats' are groups characterised by high internal homogeneity.

Since there are no sound reasons for predicting that this data is naturally structured, (with respect to the variables chosen,) either by

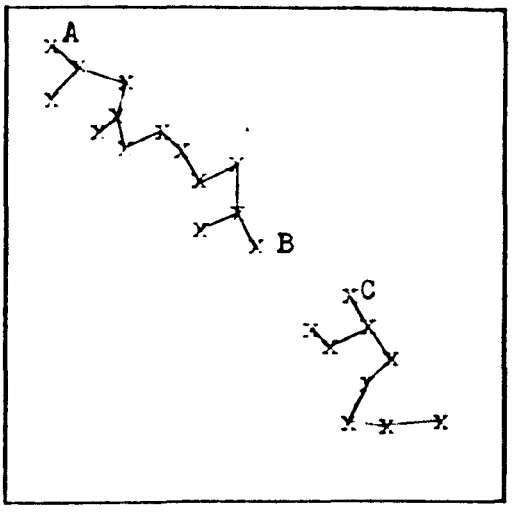


Fig. 17.

Single link clusters.

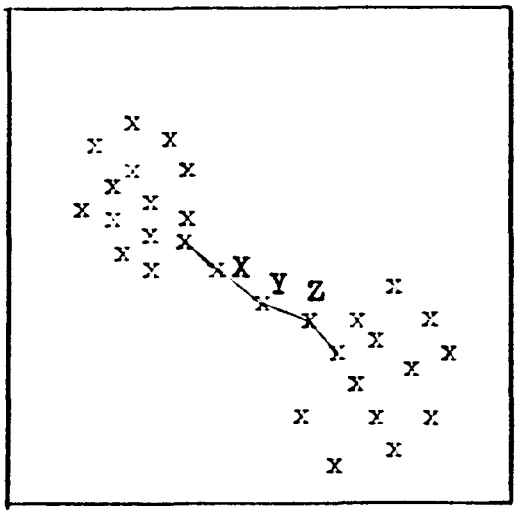
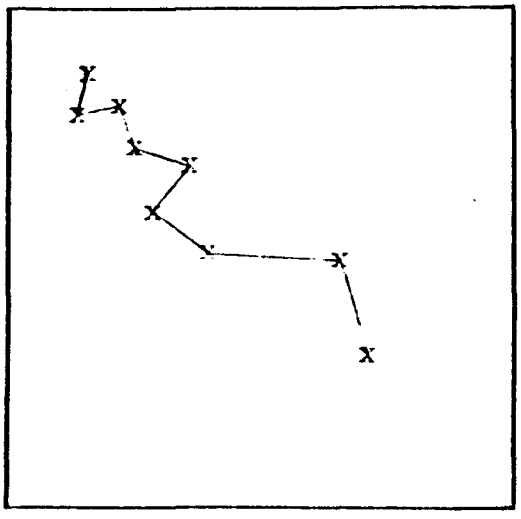


Fig. 18.

a) Spurious 'chaining' with single linkage.



b) Genuine chaining may produce apparently similar results.

homostat or segregate types, it is important to investigate the effects of techniques designed to discover both types of groupings.

Fig.19(p.114) shows the fusion process resulting from the application of the similarity ratio coefficient, and the average linkage clustering method, to the set of test data. (Cf. Fig. 16.)

With average linkage, small clusters begin to form, but these clusters chain together (as compared to the single linkage run, where individuals chain onto the main cluster).

Once again, there are no significant break points in the fusion tree (signifying distinct groupings), and, because of the chaining effect, there is no point on the similarity level scale at which the whole population is classified into groups. With this method also, then, a clear classification has not emerged, and the analysis cannot usefully proceed further.

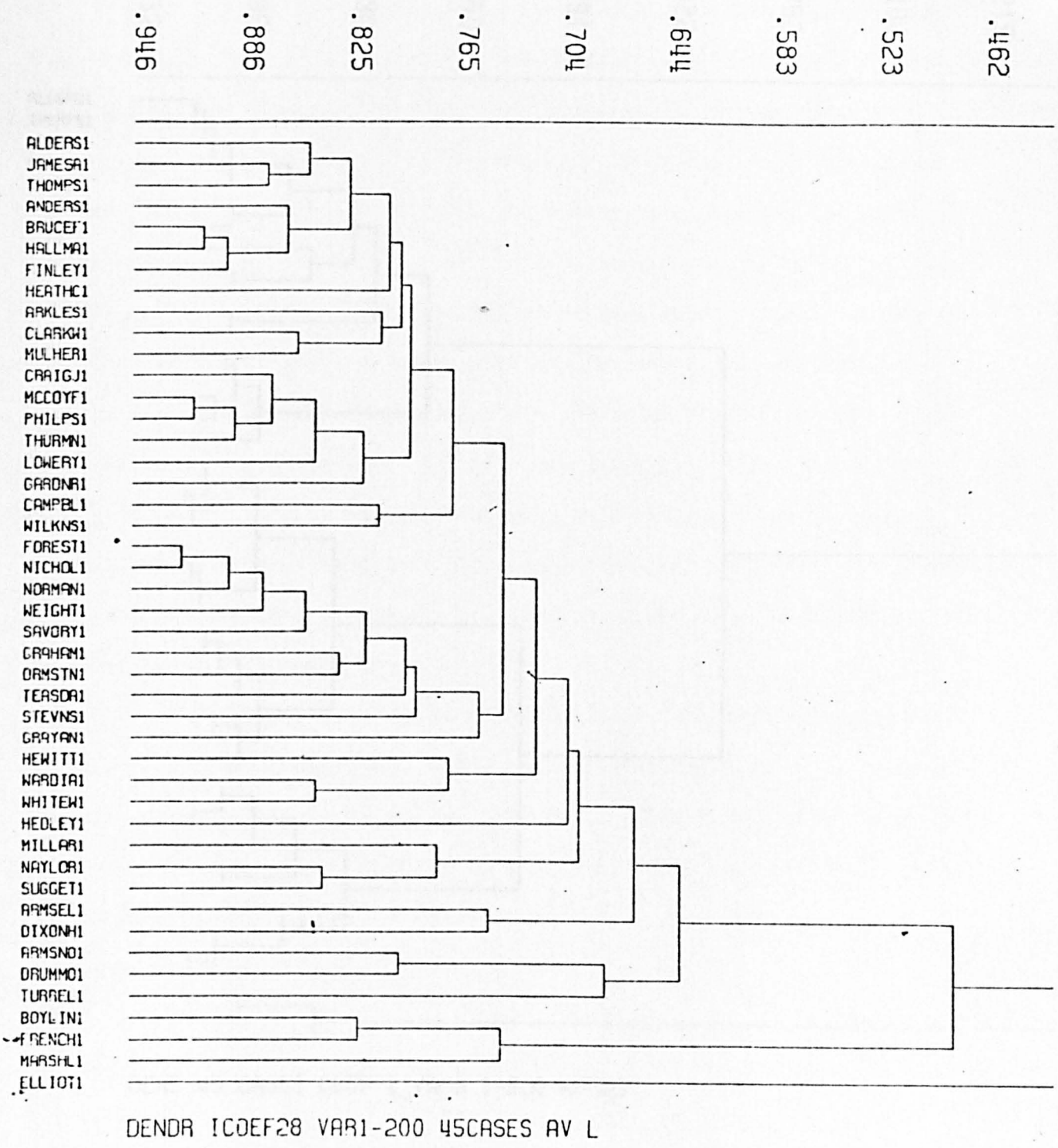
Fig. 20 (p.115) shows the same test data clustered by Ward's method. (See above, p.106). This method is incompatible with similarity coefficients, therefore a distance coefficient was used (Squared Euclidean Distance, see above, p.94 , FN.) This coefficient was chosen because all the information in the similarity matrix is retained, i.e. shape, elevation and scatter.[FN]

FN. 'Shape' refers to the contours of the indiviudal's profile across the variables; 'elevation' is an individual's mean score across all variables; and 'scatter' is a measure of each individual's deviation from the profile mean.

Of the other distance coefficients available, some are sensitive to the position of the origin, (e.g. Canberra (or non-metric) coefficient, size difference, shape difference), and the others reduce to Squared Euclidean Distance when used in combination with Ward's method (average distance, error sum, variance coefficients.)

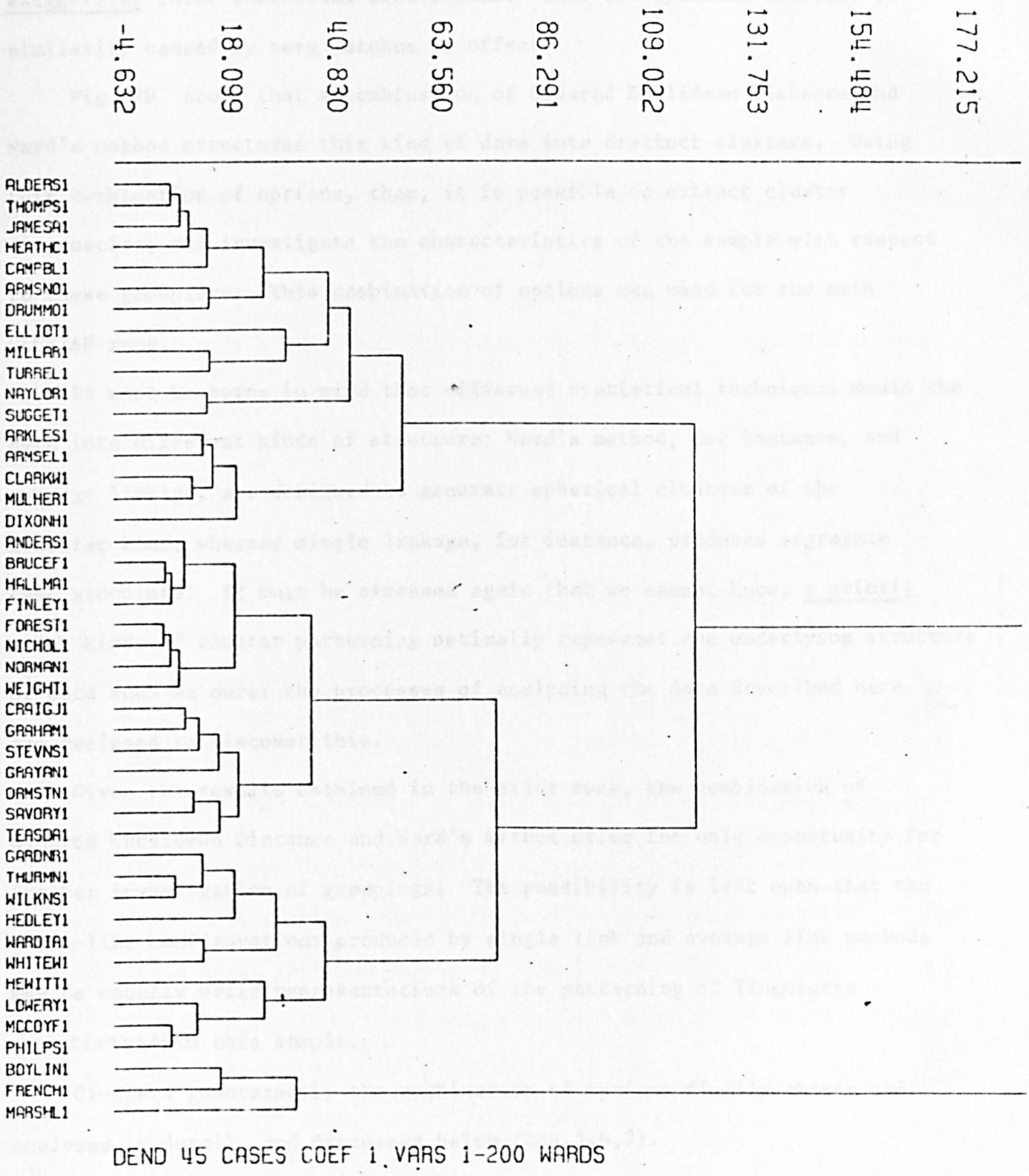
As remarked above (pp.107-109) the Similarity Ratio excludes the effects of zero matches, and thereby avoids the artificial inflation of inter-case similarity. The Squared Euclidean Distance coefficient does not exclude

Fig. 19. Dendrogram based on Similarity Ratio coefficient and Average Linkage Clustering.



DENDR ICOEF28 VAR1-200 45CASES AV L

Fig. 20. Dendrogram based on Squared Euclidean Distance and Ward's Method of Clustering.



these zero mathes; however, because the numeric difference between two individuals' scores on a given variable is squared, this coefficient exaggerates inter-individual differences. Thus the spurious increase in similarity caused by zero matches is offset.

Fig. 20 shows that a combination of Squared Euclidean Distance and Ward's method structures this kind of data into distinct clusters. Using this combination of options, then, it is possible to extract cluster diagnostics, and investigate the characteristics of the sample with respect to these groupings. This combination of options was used for the main CLUSTAN runs.

It must be borne in mind that different statistical techniques mould the data into different kinds of structure: Ward's method, for instance, and average linkage, are designed to generate spherical clusters of the homostat kind, whereas single linkage, for instance, produces segregate type groupings. It must be stressed again that we cannot know, a priori, which kinds of cluster patterning optimally represent the underlying structure of data such as ours: the processes of analysing the data described here are designed to discover this.

Given the results obtained in the pilot runs, the combination of Squared Euclidean Distance and Ward's method offer the only opportunity for further investigation of groupings. The possibility is left open that the chain-like configurations produced by single link and average link methods may be equally valid representations of the patterning of linguistic varieties within this sample.

Clusters generated by the combination of options finally chosen are analysed in detail, and discussed below (chs.5,6,7).

In addition to the programs shown on the flowchart (Fig.5, p.64), two other programs are of interest.

Program SCATTER (See listing, Appendix (X).)

This program inputs the transposed matrix, and produces graphic output

on the line printer, showing the distribution of values for a given state[FN] across the sample, in ascending numeric order.

FN. Scores are within-OU percentages.

The curves produced for state scores reinforce the significance of Labov's (1966) remarks concerning the necessity of the quantitative approach to linguistic variables.

Several interesting points emerge from an examination of the scatter plots. Regrettably, space does not permit the inclusion of all 577 here, however a selection are reproduced, and discussed briefly in the following pages.

There are many different shapes of curve amongst the 577; three types[FN] of curve shapes which occur frequently are discussed.

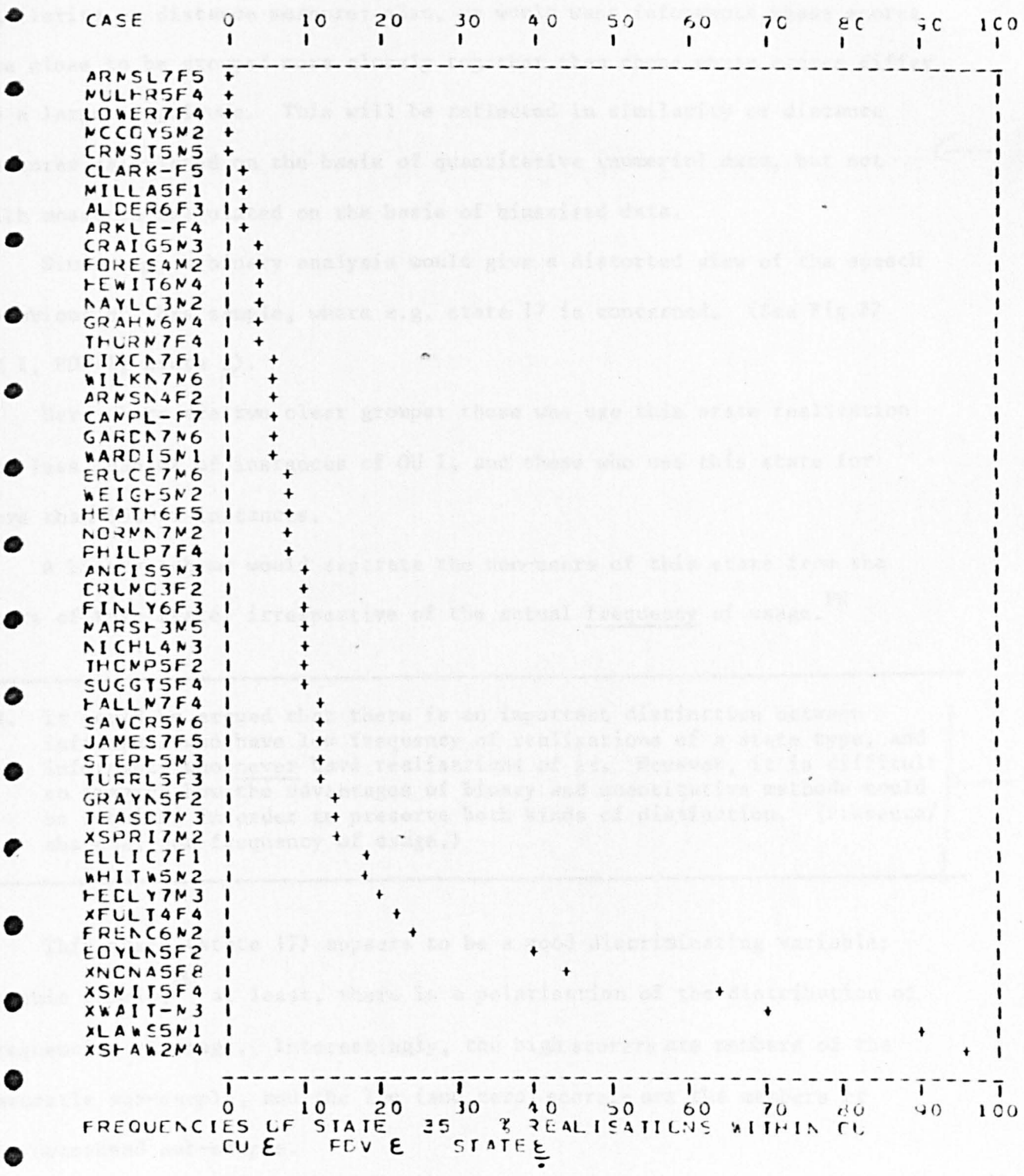
FN. Although several classes of curve shapes can be identified, the pattern of distribution of the sample across curves of similar shape differs, i.e. the sample is grouped differently by curves with the same formal properties. Thus these curve types have nothing to do with the 'typical patterns' referred to by, e.g. Trudgill (1974). See above (p. 20).

The distributions of informants' scores across these variables raises some important issues concerning how phonological variables should be treated.

The reasons for treating states as numeric rather than binary variables becomes clear when one examines the scatterplots showing the distribution of state scores across the sample.

Fig.21 (p.118) shows the distribution of state (35), (OU ɛ, PDV ɛ, state ɛ) across the 52 informants. If informants were compared in binary terms (i.e. presence or absence of realisations of this OU with this state) then the first five cases (from the top of the diagram) would score 0 for absence, and the remainder 1 for presence, of this state. Clearly this kind of approach is not satisfactory. CLARK (2%), and XSHAW(96%) would be counted

Fig. 21. Scatterplot of State 35.



similar, and CLARK (2%) and ORMST (0%) would be counted dissimilar, on this variable. Clearly, we would want the difference between CLARK and XSHAW to be reflected by the contribution this variable makes to the overall similarity or distance measure: also, we would want informants whose scores are close to be grouped more closely together than those whose scores differ by a larger magnitude. This will be reflected in similarity or distance measures calculated on the basis of quantitative (numeric) data, but not with measures calculated on the basis of binarised data.

Similarly, a binary analysis would give a distorted view of the speech behaviour of this sample, where e.g. state 17 is concerned. (See Fig.22 OU I, PDU I, state i).

Here there are two clear groups: those who use this state realisation for less than 4% of instances of OU I, and those who use this state for more than 65% of instances.

A binary scheme would separate the non-users of this state from the users of this state, irrespective of the actual frequency of usage.[FN]

FN. It could be argued that there is an important distinction between informants who have low frequency of realisations of a state type, and informants who never have realisations of it. However, it is difficult to imagine how the advantages of binary and quantitative methods could be combined in order to preserve both kinds of distinction. (Presence/absence, and frequency of usage.)

This state (state 17) appears to be a good discriminating variable: in this sample,[FN] at least, there is a polarisation of the distribution of frequencies of usage. Interestingly, the high scorers are members of the Newcastle sub-sample, and the low (and zero) scorers are the members of the Gateshead sub-sample.

FN. With a larger sample, however, the gap between the two distinct groups might be completely filled in, or not as distinct.

However, these two sub-samples are not distinguished by all variables.

Fig. 22. Scatterplot of State 17.

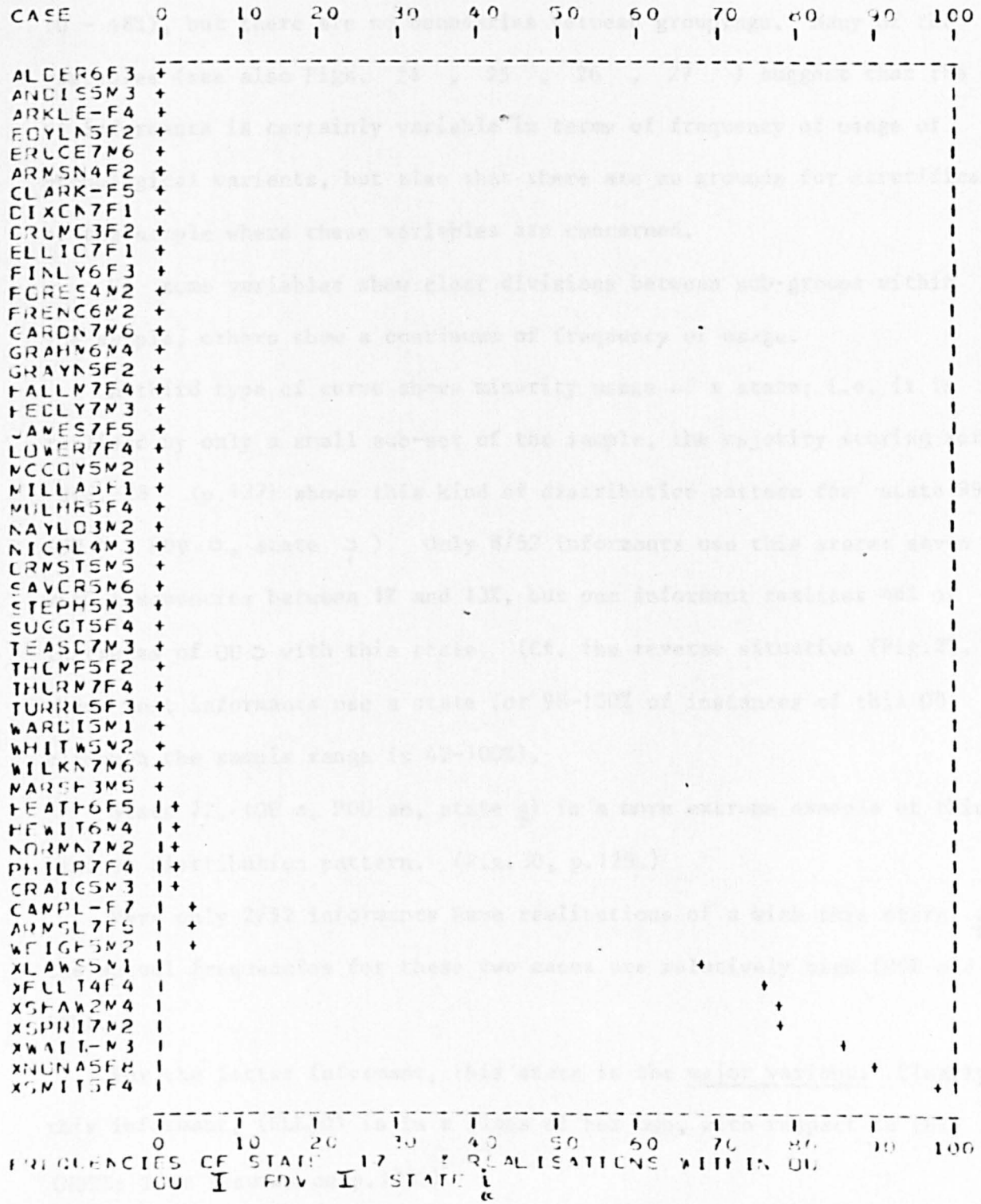


Fig. 23 (p.122) shows the distribution of state 77, (OU ɒ , PDV ɒ , state ɒ꜀). Here the Newcastle sub-sample (whose members can be distinguished by the X-prefix in the mnemonics) are interspersed with the Gateshead sub-sample. So, informants are distributed <u>differently</u> with respect to <u>different variables</u>.

This state also demonstrates the phenomenon of continuity: many of the scatter plots are continuous curves showing no break points between groups. There is a large range of variability with respect to scores on this state (0 - 48%), but there are no boundaries between groupings. Many of the variables (see also Figs. 24 , 25 , 26 , 27) suggest that the speech of informants is certainly variable in terms of frequency of usage of phonological variants, but also that there are no grounds for stratification of the sample where these variables are concerned.

So some variables show clear divisions between sub-groups within the sample, others show a continuum of frequency of usage.

A third type of curve shows minority usage of a state; i.e. it is realised by only a small sub-set of the sample, the majority scoring zero. Fig. 28 (p.127) shows this kind of distribution pattern for state 99, (OU ɔ , PDV ɔ, state ɔ̣). Only 8/52 informants use this state: seven use it with frequencies between 1% and 13%, but one informant realises 46% of instances of OU ɔ with this state. (Cf. the reverse situation (Fig.29, p.128.) where most informants use a state for 98-100% of instances of this OU, although the sample range is 42-100%).

State 72, (OU a, PDU ae, state a̤) is a more extreme example of this kind of distribution pattern. (Fig.30, p.129.)

Here only 2/52 informants have realisations of a with this state, a̤, but the actual frequencies for these two cases are relatively high (20% and 66%).

For the latter informant, this state is the <u>major variant</u>. Clearly this informant, (ELLIO) is in a class of her own, with respect to this

(NOTE: Text resumes on p.130.)

Fig. 23. Scatterplot of State 77.

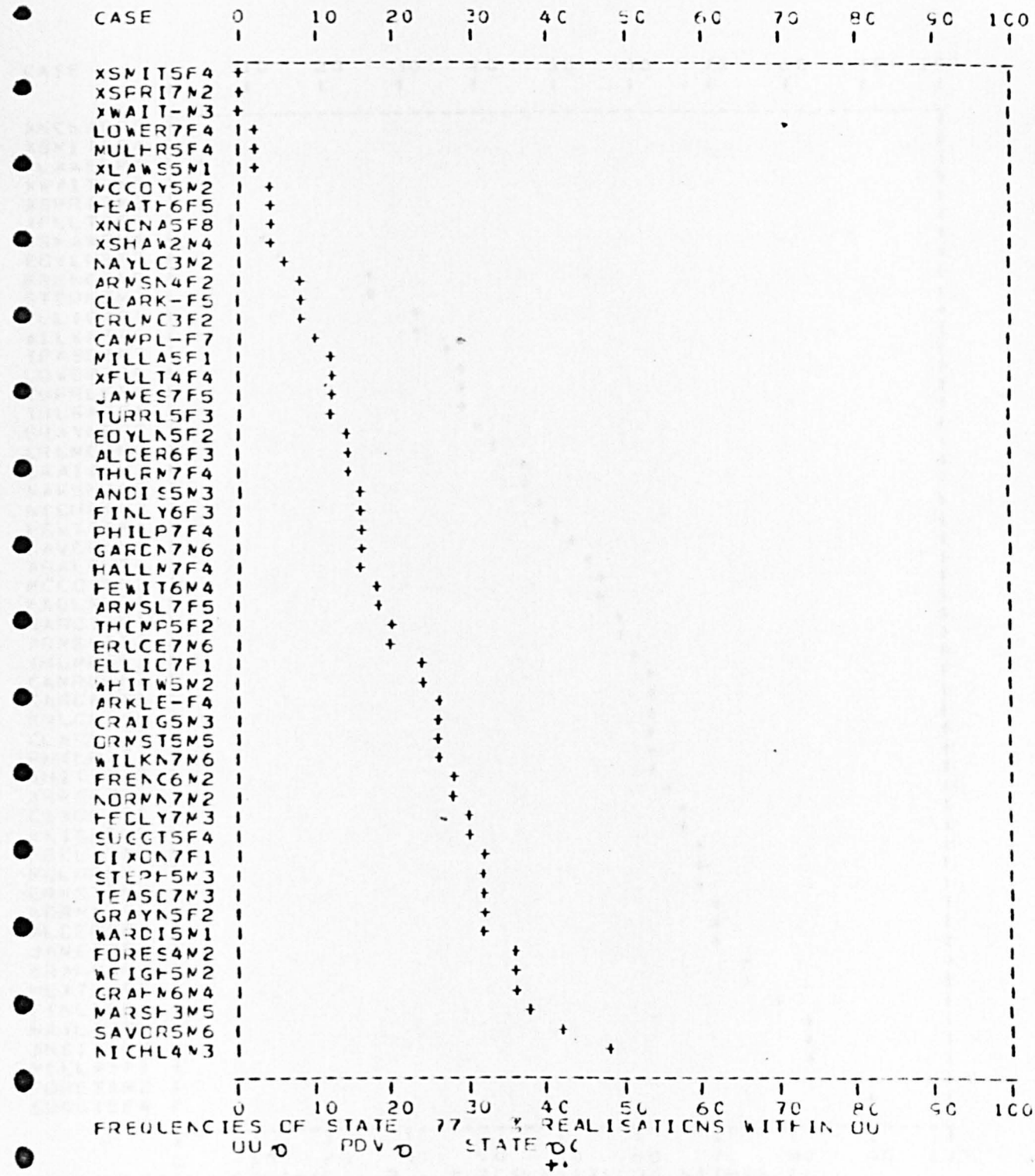


Fig. 24. Scatterplot of State 3.

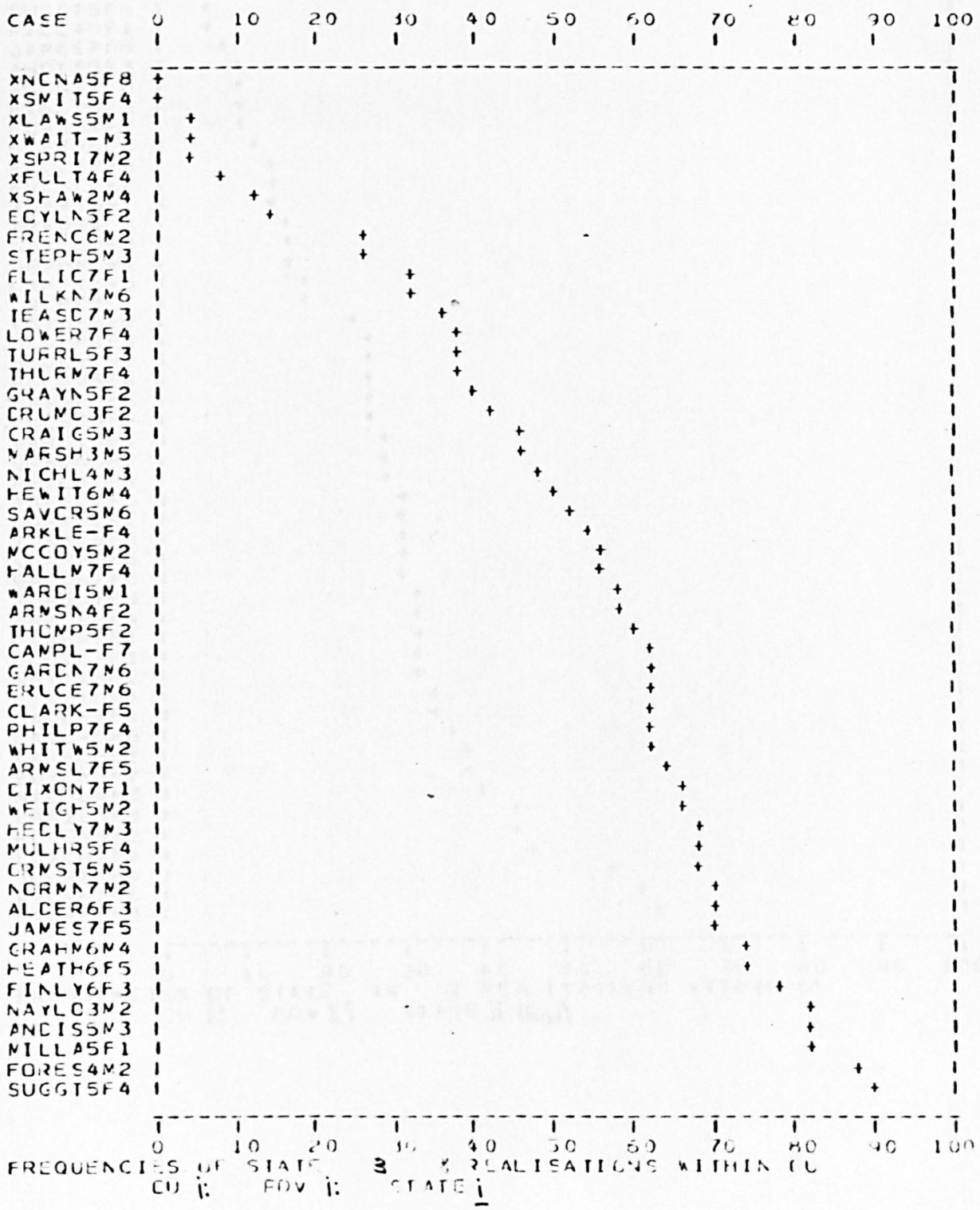


Fig. 25. Scatterplot of State 14.

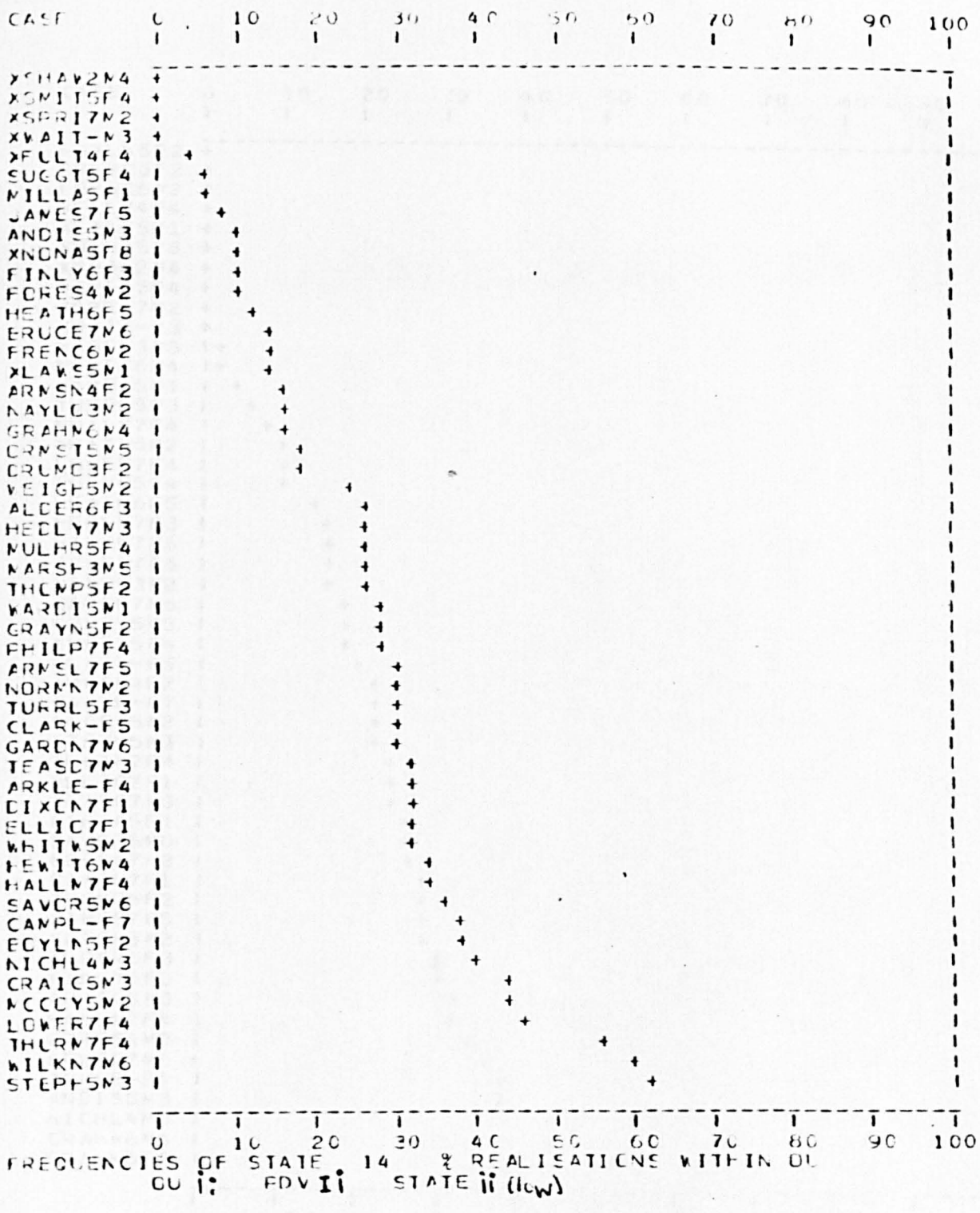


Fig. 26. Scatterplot of State 21.

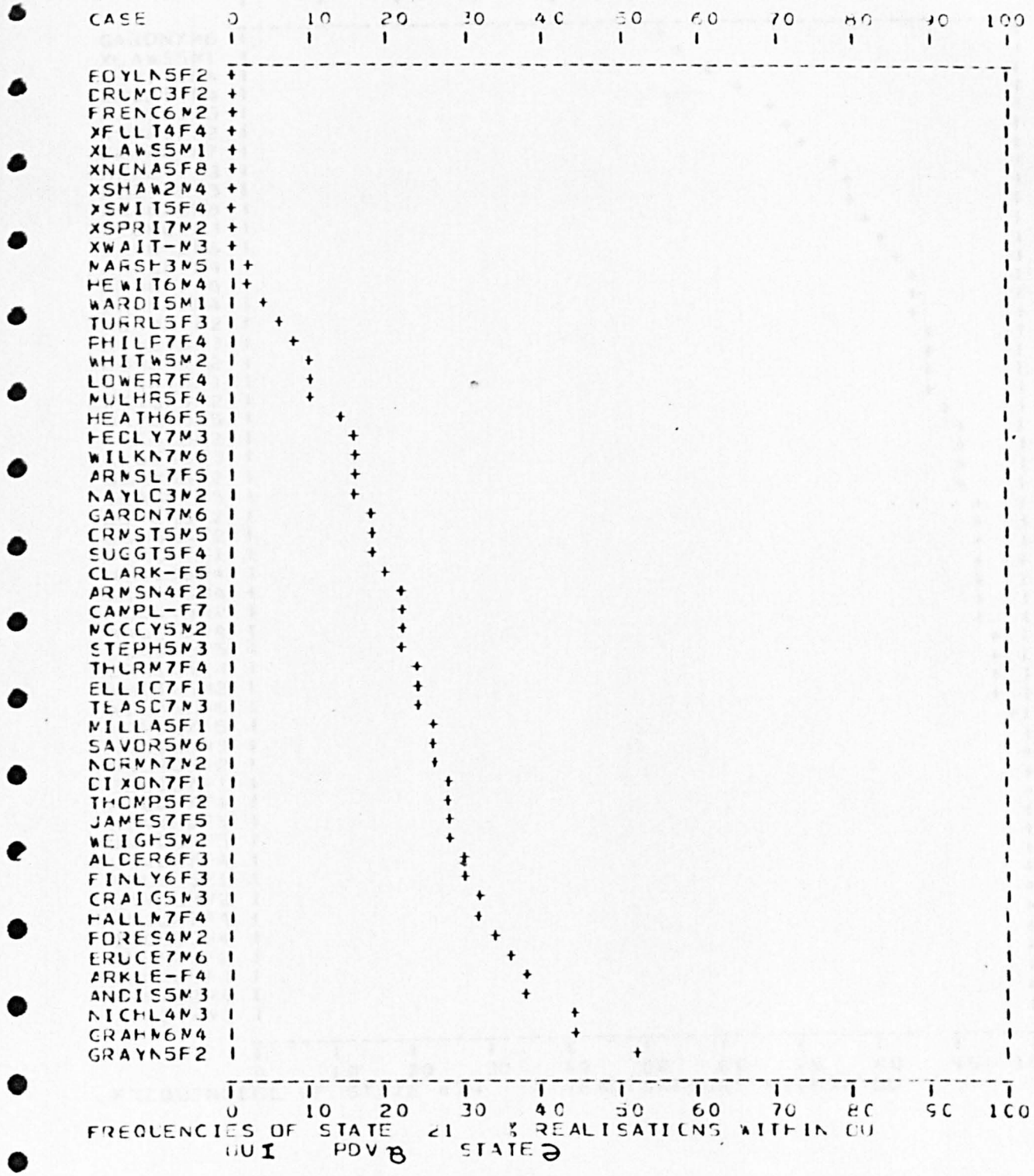


Fig. 27. Scatterplot of State 479.

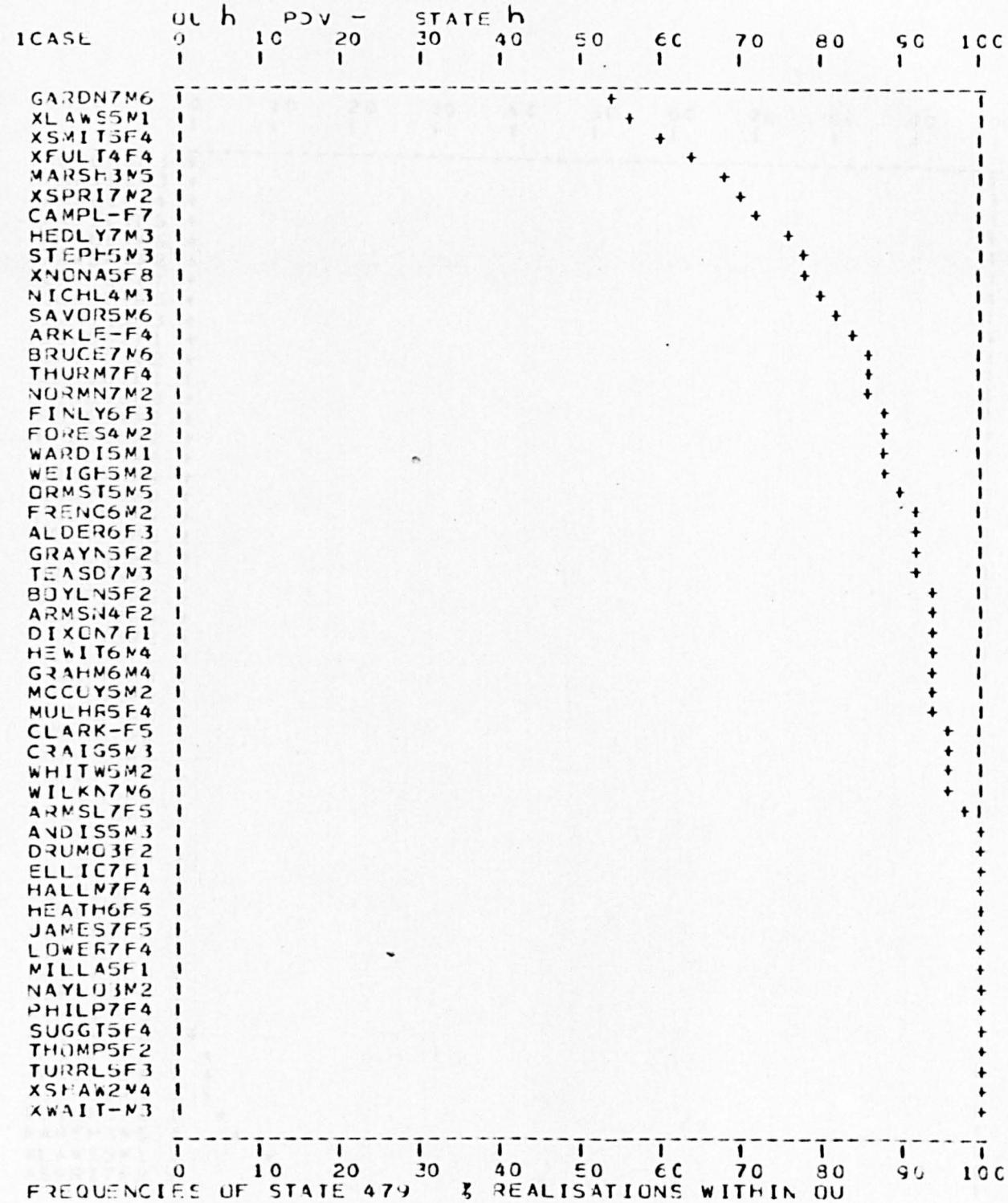


Fig. 28. Scatterplot of State 99.

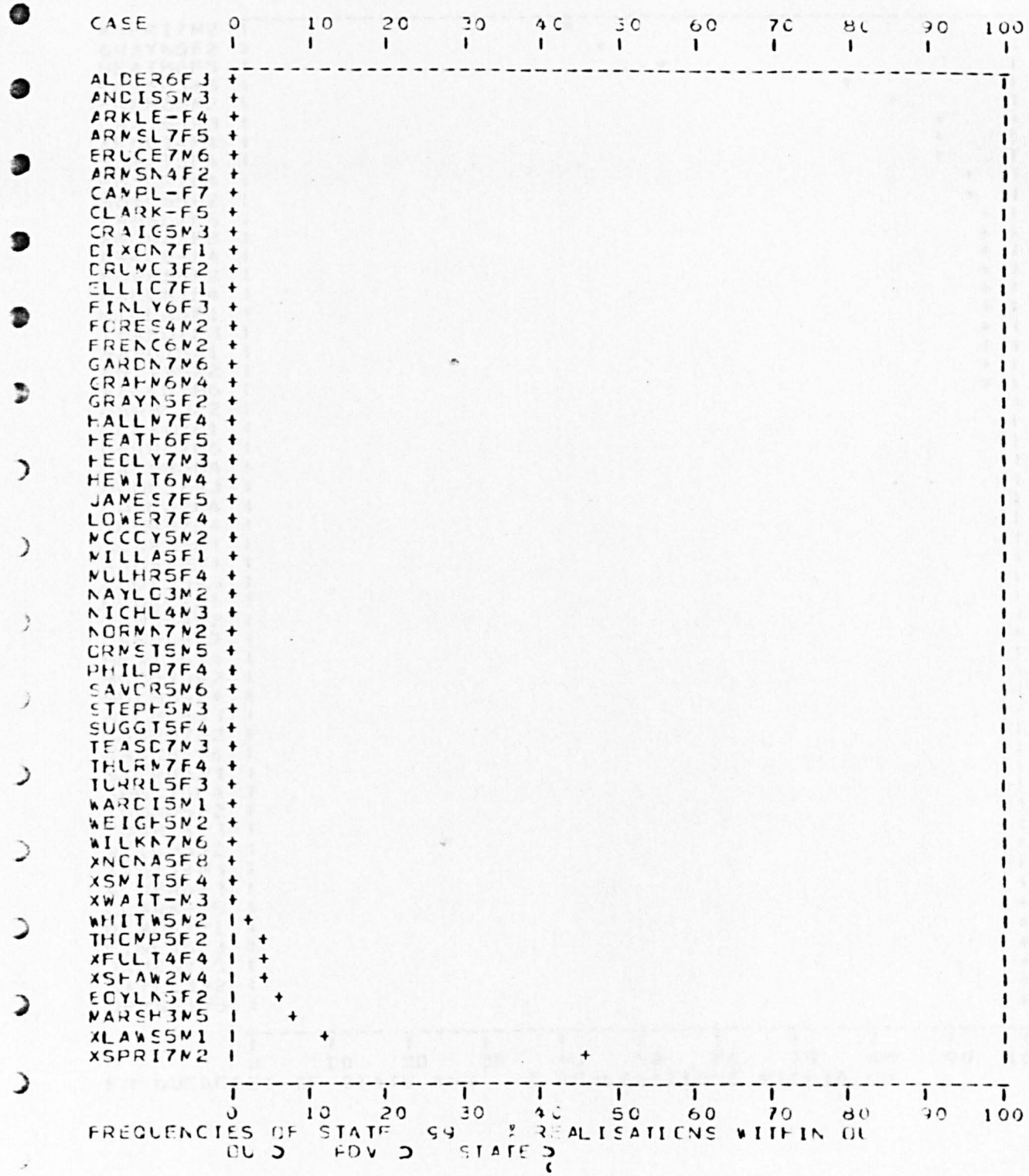


Fig. 29. Scatterplot of State 469.

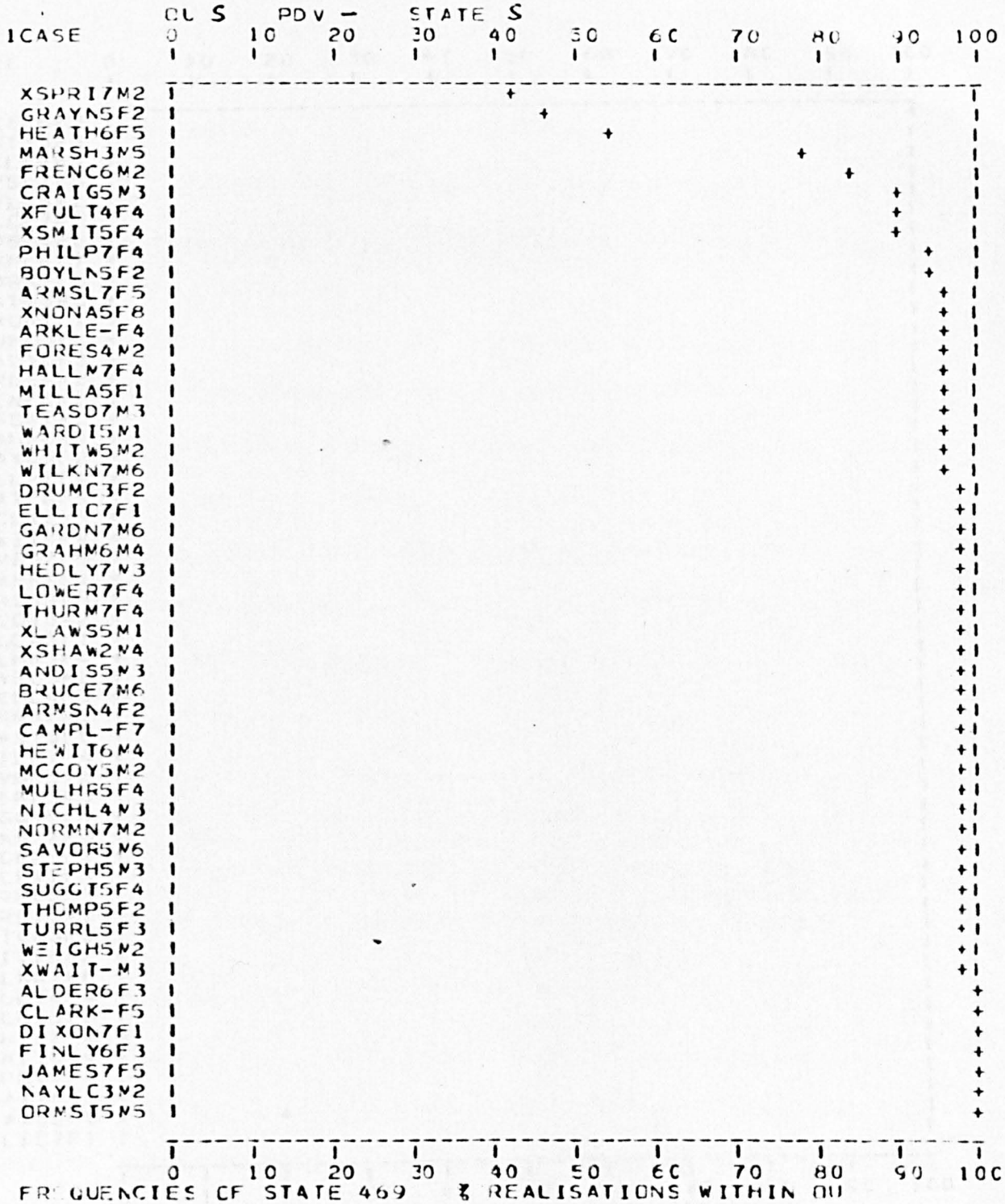
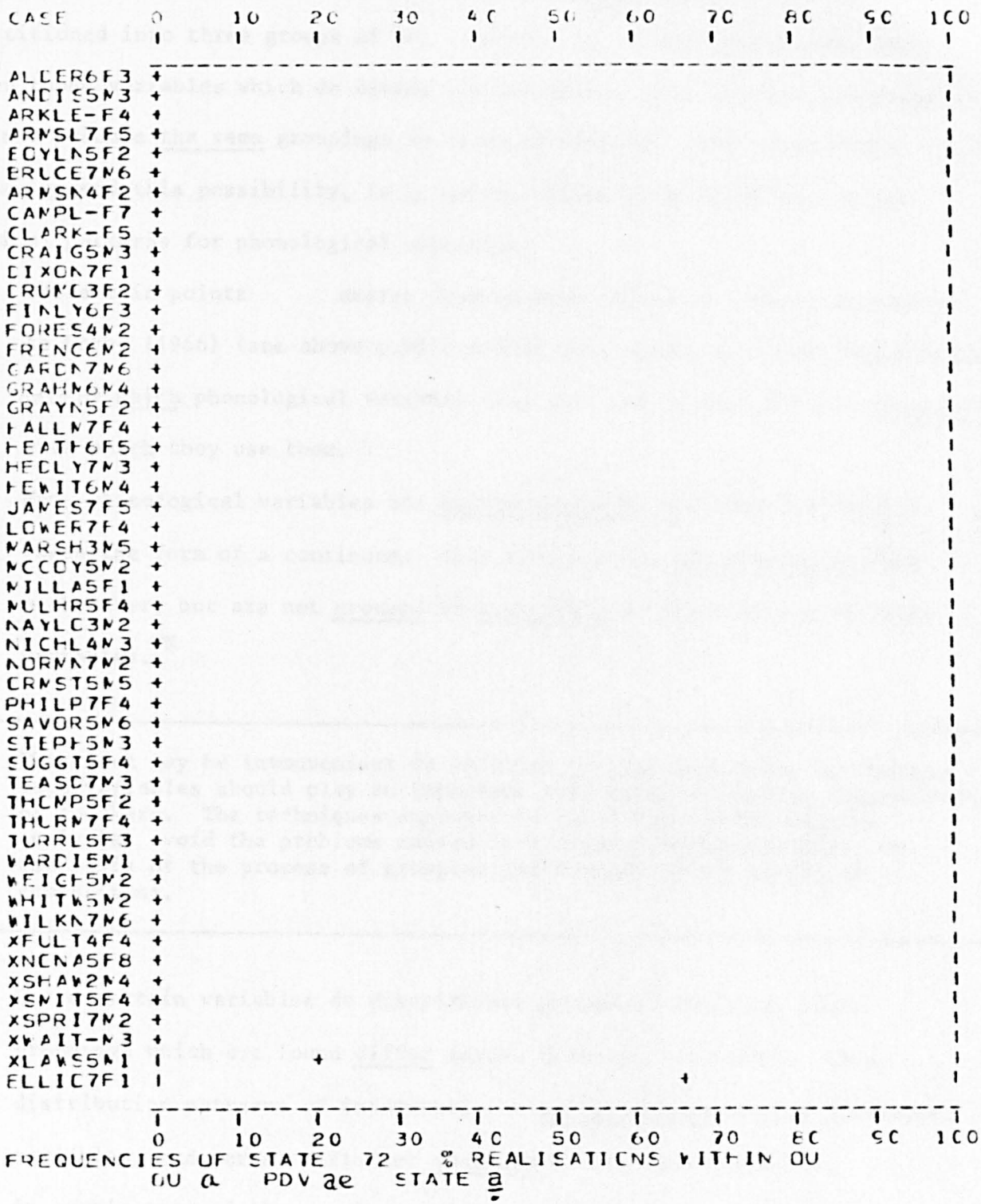


Fig. 30. Scatterplot of State 72.



state. XLAWS, too with 20%, has a score sufficiently different from both ELLIO, and the remainder of the sample, to constitute a group of one, with respect to this state. Thus, on this variable, the sample is clearly partitioned into three groups of 50, 1 and 1. It is becoming evident that even those variables which do divide the population into discrete groupings do not produce the same groupings in terms of members. Many researchers have ignored this possibility, (e.g. Labov, Trudgill) in their search for typical patterns for phonological variables.

Three main points emerge from an examination of these scatterplots.

1. As Labov (1966) (see above p.93) points out, speakers are distinguished not only by which phonological variants they use, but by the relative frequency with which they use them.
2. Many phonological variables are highly variable, yet that variability takes the form of a continuum. Thus speakers are differentiated from each other, but are not grouped or stratified by their scores on these variables.[FN]

FN. This fact may be invonvenient in relation to some methodologies; however such variables should play an important role in an exhaustive representation of speakers. The techniques employed in the T.L.S. (multi-variate analysis) avoid the problems caused by non-stratifying variables, as the basis of the process of grouping individuals relies on paired comparisons.

---

3. Where certain variables do discriminate groups of speakers, those groupings which are found differ across different variables. (i.e. distribution patterns of informants are non-homologous across different variables, and across different variants of the same variable).

An examination of these selected scatterplots provides some useful insights into the behaviour of the sample with respect to single phonological variants, and what forms of distribution the scores of single states show across the sample.

However, for the purposes of an inclusive and exhaustive classification

of speakers, the distribution of state scores across the superordinate OU, from all OUs, is the frame of reference.

Program TRAN

Given the complications which have arisen in the referencing of variables (e.g. re-ordering operations, elimination operations, see above, pp.79-92) , the original 5-digit code types are referred to by different variable names at different stages of the processing. Each 5-digit code type is referred to by a subscript in the array CODE, and by a different subscript in the array STATE, (after the re-ordering and elimination processes). Then, in a CLUSTAN run, variables are numbered sequentially from 1, so that the CLUSTAN variable numbers correspond to the STATE(n) subscript only in the first subspace (%FON1). So, e.g. STATE (155), the first state input to the %FON2 classification, is called VARIABLE 1 by CLUSTAN, and so on. (Program TRAN listing shown in Fig.31. p.132 & Appx. X.)

Program TRAN was written to produce a table to reference, showing, for each 5-digit code, which subscripts are used at different points in the processing. (See TRAN TABLE output - Fig.31A, p.133; als Appx. X.)

Thus CODELIST (1) (same as CODE(1)), corresponds to the 5-digit code 00021, and STATE(1), and %FON1 variable(1). This table gives complete cross-references between the original coding frame specifications, the output of the PL/1 programs, and the CLUSTAN output.

The PDV level of representation may also be a useful analytic level on which to base a classification, (i.e. one level higher in the structure). By comparing clusterings of the sample based on state scores, with a clustering based on PDV scores, we can test whether the fine phonetic distinctions made at the level of the state are too delicate to form the basis for a useful linguistic classification. (e.g. distinctions between one or two degrees of advancement: e.g. the third and fifth states of OU a , PDV a, which are a_{++} and a_{+}). For this OU, distinctions at

```
TRAN:PROC OPTIONS(MAIN);

/* CODELIST INPUT TO ARRAY */
/* 5-DIGIT CODES ASSOCIATED WITH CODELIST SUBSCRIPTS */
/* E.G. CODELIST(1)=00021 */
DCL CODELIST(600) CHAR(5);
DCL STRING CHAR(12);
DCL ST(557) CHAR(5) INIT((557)(5)'X');
GET DATA (CODELIST);
J=1;

/* DO LOOP REORDERS CODELIST INTO FINAL ORDER OF STATES */
  DO I=1 TO 6,8,10,16 TO 19,23 TO 37,39,40,42,44 TO 53,56 TO 63,
  65 TO 72,75 TO 79,85,86,89 TO 94,99,101,102,104 TO 111,114,116 TO 118,
  121 TO 124,682,683,686,126,688,128,130,133 TO 143,149,152,155,156,
  158 TO 169,175 TO 183,185 TO 195,197 TO 202,204,206 TO 209,211,
  216 TO 224,227,231 TO 234,635 TO 638,235,236,639,640,237 TO 240,
  242 TO 254,261 TO 264,641,642,265 TO 273,275 TO 283,286 TO 289,643,
  644,646,290 TO 298,647,648,299 TO 302,649,650,303 TO 310,651 TO 654,
  311,313,314,655,656,315 TO 317,319,323 TO 326,328,661,329 TO 334,336,
  338 TO 346,348 TO 351,662,354,355,357 TO 360,664 TO 666,363 TO 374,
  376,380 TO 409,667,668,410 TO 496,669,497 TO 502,505 TO 533,
  535 TO 546,548 TO 557,559,561 TO 565,567 TO 571,573,574,576 TO 579,582
  TO 597,599 TO 600,674 TO 678,601 TO 604,607,609 TO 613,617,670 TO 672,
  613,619,621,623 TO 634;
  ST(J)=CODELIST(I);

/* PRINT OUT TRANSLATION TABLE SHOWING, FOR EACH 5-DIGIT */
/* CODE, THE CORRESPONDING CODE(N), AND STATE(M), ARRAY SUBSCRIPTS */
/* AND ALSO THE CLUSTAN VARIABLE CODES */
    IF J>154 &  J<344 THEN DO;
    KK=J-154;
    STRING=' =%FCN2-VAR(';
    END;
        ELSE IF J>343 THEN DO;
        KK=J-343;
        STRING=' =%FCN3-VAR(';
        END;
    ELSE DO;
    KK=J;
    STRING=' =%FCN1-VAR(';
    END;
  PUT SKIP EDIT('CODELIST(',I,')=',CODELIST(I),'=STATE(',J,')=',ST(J),
  STRING,KK,')')
    (A(5),F(3),A(2),A(5),A(7),F(3),A(2),A(5),A(12),F(3),A(1));
    J=J+1;
    END;
END;
```

Fig. 31. Program listing for program TRAN.

Fig. 31A. Output from Program TRAN to show names of variables at different stages of processing.

```
******** TRANSLATION TABLE OUTPUT BY PROGRAM TRAN ********
CODELIST(   1)=21     =STATE(   1)=21     =%FON1-VAR(   1)
CODELIST(   2)=22     =STATE(   2)=22     =%FON1-VAR(   2)
CODELIST(   3)=23     =STATE(   3)=23     =%FON1-VAR(   3)
CODELIST(   4)=24     =STATE(   4)=24     =%FON1-VAR(   4)
CODELIST(   5)=25     =STATE(   5)=25     =%FON1-VAR(   5)
CODELIST(   6)=26     =STATE(   6)=26     =%FON1-VAR(   6)
CODELIST(   8)=42     =STATE(   7)=42     =%FON1-VAR(   7)
CODELIST(  10)=44     =STATE(   8)=44     =%FON1-VAR(   8)
CODELIST(  16)=81     =STATE(   9)=81     =%FON1-VAR(   9)
CODELIST(  17)=82     =STATE(  10)=82     =%FON1-VAR(  10)
CODELIST(  18)=83     =STATE(  11)=83     =%FON1-VAR(  11)
CODELIST(  19)=84     =STATE(  12)=84     =%FON1-VAR(  12)
CODELIST(  23)=121    =STATE(  13)=121    =%FON1-VAR(  13)
CODELIST(  24)=122    =STATE(  14)=122    =%FON1-VAR(  14)
CODELIST(  25)=123    =STATE(  15)=123    =%FON1-VAR(  15)
CODELIST(  26)=141    =STATE(  16)=141    =%FON1-VAR(  16)
CODELIST(  27)=142    =STATE(  17)=142    =%FON1-VAR(  17)
CODELIST(  28)=143    =STATE(  18)=143    =%FON1-VAR(  18)
CODELIST(  29)=144    =STATE(  19)=144    =%FON1-VAR(  19)
CODELIST(  30)=145    =STATE(  20)=145    =%FON1-VAR(  20)
CODELIST(  31)=161    =STATE(  21)=161    =%FON1-VAR(  21)
CODELIST(  32)=162    =STATE(  22)=162    =%FON1-VAR(  22)
CODELIST(  33)=163    =STATE(  23)=163    =%FON1-VAR(  23)
CODELIST(  34)=164    =STATE(  24)=164    =%FON1-VAR(  24)
CODELIST(  35)=165    =STATE(  25)=165    =%FON1-VAR(  25)
CODELIST(  36)=166    =STATE(  26)=166    =%FON1-VAR(  26)
CODELIST(  37)=181    =STATE(  27)=181    =%FON1-VAR(  27)
CODELIST(  39)=183    =STATE(  28)=183    =%FON1-VAR(  28)
CODELIST(  40)=201    =STATE(  29)=201    =%FON1-VAR(  29)
CODELIST(  42)=203    =STATE(  30)=203    =%FON1-VAR(  30)
CODELIST(  44)=222    =STATE(  31)=222    =%FON1-VAR(  31)
CODELIST(  45)=241    =STATE(  32)=241    =%FON1-VAR(  32)
CODELIST(  46)=242    =STATE(  33)=242    =%FON1-VAR(  33)
CODELIST(  47)=243    =STATE(  34)=243    =%FON1-VAR(  34)
CODELIST(  48)=244    =STATE(  35)=244    =%FON1-VAR(  35)
CODELIST(  49)=245    =STATE(  36)=245    =%FON1-VAR(  36)
CODELIST(  50)=246    =STATE(  37)=246    =%FON1-VAR(  37)
CODELIST(  51)=261    =STATE(  38)=261    =%FON1-VAR(  38)
CODELIST(  52)=262    =STATE(  39)=262    =%FON1-VAR(  39)
CODELIST(  53)=263    =STATE(  40)=263    =%FON1-VAR(  40)
CODELIST(  56)=281    =STATE(  41)=281    =%FON1-VAR(  41)
CODELIST(  57)=282    =STATE(  42)=282    =%FON1-VAR(  42)
CODELIST(  58)=283    =STATE(  43)=283    =%FON1-VAR(  43)
CODELIST(  59)=284    =STATE(  44)=284    =%FON1-VAR(  44)
CODELIST(  60)=285    =STATE(  45)=285    =%FON1-VAR(  45)
CODELIST(  61)=301    =STATE(  46)=301    =%FON1-VAR(  46)
CODELIST(  62)=302    =STATE(  47)=302    =%FON1-VAR(  47)
CODELIST(  63)=303    =STATE(  48)=303    =%FON1-VAR(  48)
CODELIST(  65)=305    =STATE(  49)=305    =%FON1-VAR(  49)
CODELIST(  66)=321    =STATE(  50)=321    =%FON1-VAR(  50)
CODELIST(  67)=322    =STATE(  51)=322    =%FON1-VAR(  51)
CODELIST(  68)=341    =STATE(  52)=341    =%FON1-VAR(  52)
CODELIST(  69)=342    =STATE(  53)=342    =%FON1-VAR(  53)
CODELIST(  70)=343    =STATE(  54)=343    =%FON1-VAR(  54)
CODELIST(  71)=344    =STATE(  55)=344    =%FON1-VAR(  55)
CODELIST(  72)=345    =STATE(  56)=345    =%FON1-VAR(  56)
CODELIST(  75)=361    =STATE(  57)=361    =%FON1-VAR(  57)
CODELIST(  76)=364    =STATE(  58)=364    =%FON1-VAR(  58)
```

PDV level, between PDV's ɑ, ɔ , and ae might prove to be more revealing as a basis for comparison between speakers.

On the other hand, its may prove to be the case that PDV distinctions are too gross to discriminate sub-groups of an urban speech community.

A classification based on PDV scores is, at present, being implemented. The results will not be presented here. It is anticipated, however, that interesting comparisons or contrasts may emerge.

Program COLLAPSE

This program reduces the array of state scores to PDV scores.

2 arrays are declared STARR (STate ARRay), and PARR (PDV ARRay).

For each case, the state array[FN], is taken in sections which correspond to the extent of one PDV, and the state scores within that section are summed.

FN. This sectioning applies to the vowel OU's only, as for OU's, the PDV level does not exist. Hence 343 states - see program COLLAPSE listing (Fig.32, p.135 and Appx. X.)

For example, the first PDV of OU1 has 6 states (after elimination of zero-variance states), so the first six states are summed, and the result is copied into the first element of array PARR.

The contents of array PARR are output: each element of this array represents a PDV score (expressed as a within-OU percentage).

As with program RAT, (cf. footnote, p.97), the structure of the linguistic coding frame, in terms of hierarchical organisation, is not represented in PL/1 structure form: hence, where the programming needs to operate according to the structuring of PDVs and states within OUs, this structure must be embedded in the program. (I.e. the DO loop, (lines 25-28), in effect imposes the PDV level of representation on the state array).

```
/****PROGRAM COLLAPSE- REDUCES STATE SCORES TO PDV LEVEL****/

COLLAPSE:PROC OPTIONS(MAIN);

DCL X FIXED BIN;
/* 1-D ARRAY STARR DECLARED TO HOLD STATE SCORES */
/* 1-D ARRAY PARR DECLARED TO HOLD PDV SCORES */
DCL STARR(343) FLOAT;
DCL PARR(112) FLOAT INIT ((112)0);

/* LOOP THROUGH 52 CASES */
DO KASE=1 TO 52;
   PARR(*)=0;
    X=2;

/* INPUT STATE SCORES FOR CASE(KASE) */
     DO I=1 TO 343;
     GET EDIT(STARR(I))(COL(X),F(5,1));
     X=X+5; IF X>75 THEN X=2;
     END;

   K=1; J=1;

/* SUM STATE SCORESWITHIN EACH PDV */
25   DO I=6,2,4,3,5,6,2,2,1,6,3,5,4,2,5,2,3,2,5,2,3,5,3,3,4,3,1,1,
26        2,7,4,1,2,3,5,5,5,4,4,3,4,6,2,3,1,4,4,1,1,4,4,2,2,
27        4,4,5,4,4,2,5,4,3,3,3,2,2,3,5,4,2,4,2,4,4,4,3,2,2,1,1,5,1,3,3,
28        1,3,1,4,1,2,3,1,3,3,4,2,2,2,3,2,4,2,4,2,2,1,3,2,3,2,1,1;

       DO KOUNT=1 TO I;
       PARR(K)=PARR(K)+STARR(J);
       J=J+1;
       END;
    K=K+1;
    END;

   X=1;

/* OUTPUT PDV SCORES (WITHIN-00 %) */
    DO L=1 TO 112;
     PUT EDIT(PARR(L),'.')(COL(X),F(3),A(1));
     X=X+4; IF X>76 THEN X=1;
    END;
  PUT SKIP(16);
  END;

/* END OF CASE LOOP */
 END;
FILE
```

Fig. 32. Listing for Program COLLAPSE.

(b) Social Data - Processing of.

(See above, chapter 2 and Appx. B, for a description of the social data collected on the informants).

The social data is transformed into a binary coding scheme for processing by CLUSTAN.

The reasons for this are that:

a) the data can be expressed numerically in binarised form as easily as in any other form (e.g. quantitative).

b) CLUSTAN permits mixed mode data input i.e. a mixture of binary and numeric data. Thus the 'numeric data' (in CLUSTAN terminology (see FN. above p.75)) representing scores on linguistic variables, and the social data, represented in binarised form, can be processed simultaneously in a CLUSTAN run.

Where mixed mode data is used, only one data mode can be used in the clustering process: however, diagnostic statistics can be generated for both data modes.

Thus it is possible to input a file of linguistic and social data for each informant, mask out the social data from the clustering process by generating a linguistic classification), and obtain social diagnostics directly for the linguistic clusters obtained.

Similarly, the same data file can be input, for a clustering on social variables, and linguistic diagnostics of social clusters can be obtained. (In the CLUSTAN runs described below (Ch. 7) the mixed mode data facility was used. There social diagnostics of linguistic clusters were analysed and are discussed.)

The following section describes the process by which the social coding frame was converted to a binarised form.

A binarised version of the social coding frame was constructed, where each possible response to a multiple choice question is represented by one binary variable. If the response is positive to a given category,

the variable takes the value 1 (true), otherwise zero (false).

Thus Q16 ('religion'), has three possible response categories, (which are mutually exclusive), so all responses given to this question (in terms of the categories available on the pink sheets) can be expressed using 3 binary variables.

e.g. Case (1) claims to be 'religiously active',

Case (2) " " " inactive,

Case (3) " " " anti-religious

Case (4) is coded 'NC'[FN] (i.e. data is missing.)

FN. The code of NC ('non-comparable'), means that data is missing, for this informant, on this question, because:
a) the question was not asked, or
b) the question was not answered.

Fig. 33 shows how these four cases would be coded according to their responses to this question.

Fig. 33. Binarisation of an unordered multistate variable.

'Religion'

	active	inactive	anti
Case 1	1	Ø	Ø
2	Ø	1	Ø
3	Ø	Ø	1
4	Ø	Ø	Ø

There is no need to have a binary variable for NC codes, as non-respondents are distinguished from respondents on at least 1 binary variable.

We are treating the three possible responses to this question as bearing no ordinal relationship to each other; that is to say, the three responses are not treated as points on a continuum from positive religiosit to hostility to religion, but as three distinct attitudinal states which are, socially, qualitatively different from each other. In other words,

the variable 'religion', is an unordered multistate (UM) variable.

Wishart (1969 (pp.3-4)) gives the example of hair-colour as an unordered multistate variable. If four qualitatively different categories are chosen (for coding purposes) out of the spectrum of possible shades: say,

white / red / brown / black,

then it must be made certain that no ordinal relationship is implied by the codes chosen to represent the four categories. If, for example, the numeric codes 1, 2, 3, 4, are used, then a stronger numeric relationship is implied between 1:2, than 1:4. But there is no stronger relationship between white:red, than white:black.

The coding scheme shown for the variable 'religion' does not imply any such ordinal relationship.

However, with quantitative[FN] variables, there is an ordinal relationship between the categories, e.g. age = 17-21 is more closely related to age = 21-30 than to age = 71-80.

FN. Here we are reducing a continuous scale to a series of binary variables, which correspond to 'bounded classes', e.g. age bands, 17-21 etc.

This type of variable is ordered multistate (OM). Using a similar binary scheme (i.e. one binary variable for each coding category) we can preserve the ordinal relationships by coding a 1 for all binary variables up to the category in which the response is coded. Fig.34 shows how 3 cases, aged 18, 22 and 73 respectively, are coded on age as an ordered multistate variable.

Fig. 34 Age: an OM variable

	17-20	21-30	31-40	41-50	51-60	61-70	71-80	80+	
Case 1	1	Ø	Ø	Ø	Ø	Ø	Ø	Ø	age=18
2	1	1	Ø	Ø	Ø	Ø	Ø	Ø	age=22
3	1	1	1	1	1	1	1	Ø	age=73

If there was a 1 only in the category in which the response falls, then these 3 cases would be equally similar in terms of matches and mismatches across these 8 binary variables (2 matches and 6 mismatches between each pair [FN]).

FN. A match is where both cases have a 1, or both have a zero, on a given variable. A mismatch is 1,Ø, or Ø,1.

However, if the variable is treated as OM, as shown here, then the differences between the pairs is preserved. i.e. Case 1 and Case 2 are different, but more similar to each other than to Case 3. And Case 3 is further from Case 1 than Case 2.

Case 1:Case 2 (1 mismatch, 7 matches)

Case 1:Case 3 (6 mismatches, 2 matches)

Case 2:Case 3 (5 mismatches, 3 matches)

This convention, then, preserves the relative distance between the pairs, on the grounds of e.g. age differences (See fig.35.)

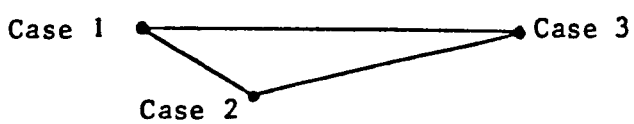


Fig.35. Distance on an ordered multistate variable.

N.B. With OM variables, the meaning of each binary variable can be reinterpreted as 'greater than'.

E.g. Age: the eight binary variables shown in Fig. 34 can be reinterpreted as:

>16, >20, >30, and so on,

so Case 1 and Case 2 are matched on the first variable (>16) score a mismatch on the second (>20) and are matched on the remaining six ($\neg>30$, $\neg>40$ etc.).

Multiple Coding (MC)

To return to the hair colour example, the coding categories for this variable are mutually exclusive, (i.e. we assume that a redhead by definition does not have black hair).

However, in some unordered multistate variables, more than one category may be applicable simultaneously.

For example, the (hypothetical) variable 'countries resided in for at least one year' could have as many responses as the informant has years. Here we can use multiple coding, that is, enter a 1 under as many categories as are appropriate.

Q1 'cityness of informant' is an MC variable.

The application of the social space to the sample.

Each possible response to the questions shown on the social coding sheets was treated as one binary variable. So, 1 is scored for every response which is coded positively, and zero is scored for each coding category to which the informant does not respond positively. There are, in total, 297 possible responses across the 38 questions, giving a total of 297 binary variables. Approximately half of these responses were coded at least once for this sample, (153 altogether). The remaining 144 categories are redundant for this sample. (A coding category is redundant for a given sample if all members of that sample score zero on it. E.g., if there is no case (informant) in the sample who has resided in UK Wales, response 6 of Q2 is redundant. (All cases score zero.) This coding category produces 'noise' in the classification, as all cases will score spurious similarity on this irrelevant variable. The inclusion of many redundant coding categories such as this leads to similarity levels across the whole sample being artificially inflated. There are two ways of avoiding this:

1. by omitting redundant variables;
2. by including them, but masking them from the clustering process.

The first option has the advantage of reducing the dimensionality of the measurement space, in that the total number of variables is reduced. Thus input data files are smaller, processing time is reduced, and only relevant diagnostics are produced. (The cluster analysis program used here, CLUSTAN, prints diagnostics for all variables, regardless of whether some of them have been masked.) If initial analysis shows that a number of variables are redundant, and will therefore contribute nothing to the classification process, then it is more convenient to have diagnostic information printed only for those variables which are actually used in the clustering. Thus the choice may be made to eliminate the redundant variables from the data, as far as that sample is concerned.

However, if one wishes to introduce more cases at a later date, and

cluster a larger sample, it may turn out that variables which were redundant for the original sample are not redundant for the new sample. Exclusion of these variables will then cause difficulties. The space, which was reduced, will then have to be extended to include these variables, or else the information contained by them will have to be discarded. However, this is not a very serious problem; it is possible to include new coding categories (actually old discarded ones) by a similar process to that used to incorporate 'new' variables into the linguistic coding frame. There is no problem of comparability between the original, and the new members of the sample, as it is known that the original members all scored zero for these variables.

I decided, then, to take the first option, and omit those variables which were redundant for this sample.

So, the sample was classified socially on the basis of the 153 binary variables derived from a reduced version of the coding scheme shown in appendix B. Table 1 (pp.142-146) lists the variables in this reduced coding frame, following the elimination of redundant coding slots. The leftmost column shows the question number, corresponding to the social coding sheets (appendix B), the next column shows the response code, then the CLUSTAN variable code is given, and then the definition of the response category is given in the right-hand column. Table 2 (p.147) shows the number of coding categories in the original coding frame, and the reduced coding frame, for each of the questions.

Questions 13, 16, 19 and 24 were also eliminated, as response rates for these questions were less than 20 (out of 52). Low response rates cause similar classificatory problems to those caused by redundant variables, in that informants are mis-classified as similar on the basis of absence of the same response. ('Zero matches'.) The lower the response rate, the more distortion is introduced. Table 3 (p.148) shows the response rates for this sample, for the 38 questions.

Table 1

Table showing values of CLUSTAN binary variable codes (social classification), and corresponding responses (social coding sheets.)

Questionaire number	response code	bin. code CLUS	response value
1. Citiness of informant	T	1	Tyneside
	L	2	London
	3	3	Market town
	4	4	other
	M	144	Merseyside
	LS	145	Leeds
2. Regionality of informant	1	5	UK Northern
	2	146	UK E & W Ridings
	3	148	UK NW
	5	147	UK Midland
	8	6	UK London, SE
3. Regionality of both parents	1	7	UK Northern
	2	8	UK E & W Ridings
	3	149	UK NW
	4	9	UK N Midland
	5	10	UK Midland
	8	11	UK London, SE
	11	12	UK Lowlands
4. No. moves per 5 yr. period before marriage	1	13	none
	2	14	less than five
	7	15	five or more
5. ditto (after marriage)	1	16	none
	2	17	one
	3	18	two or more
6. age	1	19	17-20
	2	20	21-30
	3	21	31-40
	4	22	41-50
	5	23	51-60
	6	24	61-70
	7	25	71+
7.	1	26	male
	2	-	female

Table 1 cont.

Questionaire number	response code	bin. code CLUS	response value
8. School leaving age	2	27	legal minimum
	3	28	" " + 1 yr.
	4	29	" " + 2 yrs.
	5	30	" " + 3 yrs.
	7	31	" " + 5 yrs.
9. Tertiary and further education	1	32	none
	2	33	Uni/Poly. full time
	3	34	Tech./nursing/secretarial, full time
	4	35	Coll. Ed.
	6	36	day release
	7	37	night school
10.Attitude to education (of self)	1	38	negative
	2	39	RRR, basic skills
	3	40	liberal
	4	41	job oriented
11.Attitude to education (of children)	1	42	negative
	2	43	RRR, basic skills
	3	44	liberal
	4	45	job oriented
12.Distinction education boys/ girls.	Y	46	distinction made
	N	47	no distinction made
14. Parental control of children	1	48	direct verbal
	2	49	indirect verbal
	3	50	direct physical
	4	51	indirect physical
15. marital status	1	52	married
	2	53	single
	5	54	widow
17. nuclear family size	1	55	1
	2	56	1+
	3	57	2+
	4	58	3+
	5	59	4+
	6	60	5+
	7	61	6+

Table 1 cont.

Questionaire number	response code	bin. code CLUS	response value
18. sex	1	62	zero bias
distribution	2	63	F bias
of children	3	64	M bias
20.distance of	1	65	same local authority
spouse's prim-	2	66	less than 50 miles
ary regionality	3	67	50+ miles
21. mic. envir.	1	68	neutral
preference	2	69	dissatisfied
(sentiment)	4	70	satisfied stable
22. ditto	1	71	neutral
(housing)	2	72	dissatisfied
	3	73	satisfied ambitious
	4	74	satisfied stable
23a. decor	1	75	good
'taste aspiration'	2	76	bad
	3	77	indifferent
23b. financial	1	78	low
commitment to	2-3	79	
23a.	4-5	80	
	6-7	81	
	8-10	82	high
25a. env. pref.	1	83	rural
(type/size)	2	84	smaller town
	3	85	same size
25b. ditto	1	86	south
(location)	2	87	north
	3	88	nowhere else
	5	89	abroad
26. positively	Y	90	yes
Tyneside conscious	N	-	no
27.	1	91	non-existent, unknown
social	2	92	" , known
integration	3	93	antagonistic
with	4	94	minimal, pleasant
neighbours	5	95	cordial
	6	96	intimate

Table 1 cont.

Questionaire number	response code	bin. code CLUS	response value
28.	3	97	inspectional, supervisory, non-manual, higher grade
father's	4	98	ditto, lower grade
occupation	5	99	skilled manual, routine non-manual
	6	100	semi-skilled manual
	7	101	unskilled manual
29.	2	152	managerial & executive
informant's	3	102	
present	4	103	
occupation	5	104	(see same response codes - Q. 28)
	6	105	
	7	106	
30.	2	153	
informant's	3	107	
first	4	108	
occupation	5	109	
	6	110	
	7	111	
31. job	I	112	prospects/thinking/self-deciding
preference	R	113	immediate gain/learnt/supervised
32.	1	114	high
job	2	115	↕
satisfaction	3	116	fairly low
33a. daily	1	117	predom. radio
exposure to	2	118	" TV
TV, radio	4	119	TV only
	5	120	non-own
33b. ditto	1	121	intense/non-selective
intensity/	2	122	intense/selective
selectivity	3	123	non-intense/non-selective
34. hobby-drinking	Y	124	yes
" housework	Y	125	yes
35. leisure	1	126	satisfied
satisfaction	2	127	partially "
	3	128	disgruntled

Table 1 cont.

Questionaire number	response code	bin. code CLUS	response value
36.	1	150	(see social coding sheets)
hobbies	4	129	
	5	130	
	7	131	
	8	132	
	10	151	
	12	133	
	15	134	
	16	135	
	22	136	
37.connection	1	137	approve
occupation/	2	138	accept
voting behaviour	3	139	dissaprove
38.	1	140	Conservative
voting	2	141	Labour
preference	6	142	refusal
	7	143	floater

Table 2.

Table showing the number of coding categories in the original social coding frame, (o), and the number of binary variables in the reduced coding frame, (r), by question number (q), as on the social coding sheets.

q	o	r
1	5	6
2	26	5
3	26	7
4	8	3
5	8	3
6	8	7
7	2	1
8	8	5
9	9	6
10	6	4
11	6	4
12	3	2
13	3	-
14	5	4
15	5	3
16	4	-
17	8	7
18	4	3
19	7	-
20	4	3
21	5	3
22	5	4
23a	4	3
23b	11	5
24	6	-
25a	4	3
25b	4	4
26	3	1
27	7	6
28	8	5
29	8	6
30	8	6
31	9	2
32	5	3
33a	6	4
33b	4	3
34	3	2
35	4	3
36	25	10
37	4	3
38	8	4
TOT	297	153

Table 3.

Response rates on social questions.

Q	RR		Q	RR	
1	52		23a	44	
2	52		23b	44	
3	51		24	16	*
4	48		25a	45	
5	42		25b	47	
6	52		26	43	
7	52		27	45	
8	50		28	46	
9	51		29	45	
10	48		30	49	
11	46		31a	16	
12	22		31b	16	
13	16	*	31c	11	
14	32		32	37	
15	52		33a	45	
16	10	*	33b	45	
17	52		34	40	
18	38		35	47	
19	19	*	36	52	
20	38		37	29	
21	51		38	37	
22	49				

Q= question number (see social coding sheets.)

RR= response rate out of 52.

* signifies questions with RR < 20, which were eliminated.

(Q's 31a, 31b, 31c were not eliminated. For explanation see p.).

The application of the social coding frame to the sample.

(Unless otherwise specified, the coding categories shown in appendix B are unchanged for the reduced SocSp, apart from NC slots: (see below Q1). (Cityness. (UM, MC) This variable refers to the size and type of community to which the informant belongs, and has belonged in the past. There were originally 5 coding categories for this variable:

1. city 2. big town 3. market town 4. other 5. NC.

In the sample under study, 49/52 were coded on the first category. However, these informants had resided in different cities, including Tyneside (the majority), also London, Merseyside and Leeds. The original coding scheme for this variable was therefore extended to cover these distinctions. The variable now includes two kinds of information; that specified by the original definition of its variants, and also information concerning the particular cities which informants have resided in. Those variables in the original paradigm which were not taken up by any members of the sample are eliminated (big town). The variants are now defined as: citiness = city (sub-categorised into Tyneside, London, Merseyside and Leeds), market town/other.

There is no need to include the NC (non-comparable) category. If it were included, then absurd matches would occur. Informants for whom this data was not available would score a positive match on the NC category, if it was included: this is obviously unnecessary and biassing. (This elimination of the NC slot applies to all questions).

Q2. Regionality. (UM, MC) Informants are coded according to the regions (of UK, and abroad) in which they have lived (for at least 2 years of their life). Some of this information will overlap with the categories added to Q1, however, information on particular cities which have been resided in could provide a finer classification than those regions specified under regionality.

Q3. covers the same information as Q2, but for the parents of the informant. (UM, MC).

Qs. 4,5. Number of moves per 5-year period before (and after) marriage. (OM) This information provides a measure of the geographic mobility of informants. There is a problem of logical dependency with this variable. Any positive responses on this variable must correlate with some variants of marital status (married, widowed, divorced) as single informants will not be coded on this variable. (One way round this problem is to reinterpret these two questions for single informants, e.g. by taking some comparable period such as the ten years after they left school).

There were originally 8 coding categories for both Qs 4 and 5. The only categories which were coded positively for any members of this sample were (Q4) number of moves = none, = 2, and = 5+. Q4 was restructured (for this sample) with 3 coding slots: zero moves/ 1 - 3/3+ moves. Q5 had 3 positively coded variants, number of moves = none, = 1, and = 2+. 3 categories were defined for this sample, then, which are zero moves/1 move/2+ moves.

Q6. Age. (OM) For the 7 age groups defined (see appendix B) we need only 6 binary variables to express all the distinctions. This is because age is an ordinal scale, which is coded as an ordered multistate variable (see above, p. 138) Since all members of the sample are over 16, the first variable (17-20) is redundant (i.e. everyone will score a 1 on this state). The 6 binary variables can be interpreted as 20+, 30+, 40+ etc. So, an informant in his thirties will be coded positively on 20+ and 30+, but will be coded zero on the remaining categories. Informants aged 17-20 are distinguished by scoring zeros across all age categories.

Q7. Sex. One binary variable is used, 1 is coded for male, 0 for female.

Q8. School leaving age. (OM) This variable indicates whether informants left school before the legal minimum age (lma), at the lma, or for how many years they stayed on at school beyond this point.

Q9. Tertiary and further education. (OM) This information, combined with that coded under Q3, provides a fuller picture of educational history than would, for example, an education index which only takes into account the point at which education ceased.

Qs. 10, 11. provide attitudinal data concerning the informant's attitude to his own education and to the education of his children. (UM, MC).

Q12. Distinction between education of boys and girls.
It is considered that the responses 'yes' and 'no' both indicate positive attitudes which differ qualitatively from the response 'don't know' (= indifference?), and from each other. Two binary variables are sufficient to express these 3 distinctions, which are coded thus:

Yes	No	
1	0	(yes)
0	1	(no)
0	0	(don't know)

Q13. Positive distinction between parental and school roles. With a response rate of only 15/52, this variable was not included for this sample.

Q14. Parental control of children. 'Direct verbal' control means, for example, 'don't do that!', whilst indirect verbal control involves, e.g. explanation of why 'doing that' is not a good idea. 'Direct physical' covers slapping, smacking, etc., whilst 'indirect physical' includes, e.g. not allowing the child to go out to play. (UM, MC).

Q15. Marital status. The only categories applicable to this sample are: married/single/widowed. (UM).

Q16. Religion. This variable was omitted. (Response rate = 10/52).

Q17. Nuclear family size. (OM) This factor is of sociological interest: social behaviour and psychology are affected by family structure. It may be that language behaviour is affected by the size and constitution of the family group. For instance, parents may have a qualitatively different lifestyle than childless couples of similar age; this may influence speech behaviour in ways which can be traced. Additionally, (apart from the inter-influence of the generations), members of large families may behave differently (linguistically) from, e.g., those living alone.

Q18. Sex distribution of offspring. (UM) Parents of girls may be differently influenced by interactions with their children than parents of boys.

Q19. Average age gap between offspring. This variable combined with information from Qs 15 and 17 gives a measure of the duration of parents' exposure to members of the younger generation within their own household. This variable was not included, however, as the response rate for this sample was only 19/52.

Q20. Distance of spouse's primary regionality. (OM) Marriage to someone from a different geographical background/speech community may influence the informant's speech behaviour. This information may also relate to the informant's (or spouse's) degree of entrenchment in the local community, or their geographical mobility, in the past.

Q21. Micro-environmental preference in terms of sentiment. (UM) The

informant is requested to indicate his response to the locale, the neighbours, and his degree of identification with the local community, or his aspirations towards a different (better?) local environment. (The facts concerning local environment are sociologically significant; so too are the informants responses to his situation).

Q22. relates specifically to the informant's response to his environment in terms of his satisfaction, or otherwise, with his housing conditions. (UM)

Q23. Interviewer's assessment of decoration, furnishing and domestic equipment. Section (b) of this question (see appendix B) originally provided a 10-point rating scale to represent informants' financial commitment to their 'taste aspiration'. If all these categories were included then we would need 10 binary variables to express these distinctions. Such a large number of binary variables for only one social feature would assign excessive weight to it in the classification; therefore this variable was reduced to a 5-point scale. These 5 distinctions cover the categories shown on the coding sheets as 1, 2-3, 4-5, 6-7, 8-10.

Q24. Rateable value of dwelling. This variable provides information relating to social status. However, the response rate was 16/52, and it was therefore not included in the classification of this sample.

Q25. Macro-environmental preference. (UM) Informants are asked what size/type of community they would prefer to live in, and also which part of the country they prefer. Several informants responded to section (b) by stating they would prefer to live abroad. This is a sufficiently different response to, e.g. the 'north', or 'nowhere else', to warrant the inclusion of an additional coding slot here.

Q26. Informants are coded on presence or absence of positive Tyneside

consciousness. A sense of identification with the (wider) community of the conurbation (cf. Qs 21, 22, 25, 27), may be a significant social psychological factor influencing degree of localisation of speech.

Q27. concerns the nature, and degree, of the informant's social involvment with his neighbours. (UM, MC).

Qs 28, 29, 30. Occupation. (UM) The informant's present occupation, and his first occupation after leaving school, and also his father's occupation are coded according to Hall and Jones' (1950) 'Social Grading of Occupations', which correspond to the Registrar General's Occupational Gradings thus:

Hall and Jones	Registrar General
1-2	I - upper and middle class
3-4	II - intermediate
5	III - skilled workmen
6	IV - intermediate
7	V - unskilled workmen

(Stevenson: 1911).
(See coding sheets (appendix B) for full definitions of Hall & Jones' categories).

These codings give a detailed picture of the informant's present occupational status, and whether this represents an increase, or a decrease, or stability in relation to first occupation, and in relation to his father's occupational status.

Occupation group 1 was not represented in this sample, therefore this category was not included in the classification.

Q31. Job preference. This question complements the analytic sociological information derived in Q28, by soliciting the informant's occupational desiderata. There were 3 polar distinctions specified in the original

coding scheme:

(a) prospects versus immediate gain;

(b) thinking (new elements) versus learned (no new elements);

and (c) supervised versus self-deciding.

These represent qualitatively different aspects of a job situation, concerning, respectively, motivation for work being related to career ambitions or financial returns; the degree of intellectual challenge preferred; and the desired degree of initiative attached to a job. However, the response rates on these 3 subsections were quite low: 16/52; 16/52 and 11/52 respectively. As more informants (24) responded to at least one of these sub-questions, though, it was possible to devise a single variable which represents a reduction of this information, but which avoids dispensing with it altogether. This variable was defined on the basis of the expectation that there would be an association between the first responses to sections (a) and (b), and the second response to section (c). I.e., informants who cite prospects rather than immediate gain as motivation are also likely to prefer a job which involves thinking and a degree of autonomy, to one which is routine, and supervised. The combination of responses (a)1/(b)1/(c)2 was arbitrarily labelled 'I' (for initiative), and the combination (a)2/(b)2/(c)1 was labelled 'R' (routine).

Criteria for assigning informants to categories 'I' or 'R' on the grounds of their responses to Q31 (a), (b) and (c) were then laid down thus: if all 3, or 2 out of 3 of the responses coded fit in with the typical 'I' pattern, then this informant is coded 'I', (and similarly for the typical 'R' pattern). If only 2 responses are coded (the 3rd being NC), and the responses conflict with the typical patterns, then the informant is assigned to 'I' or 'R' on a priority basis, where (b) takes precedence over (c), which takes precedence over (a). If 3 responses are coded, and there is a conflict between typical 'I' and 'R' responses, then the category ('I' or 'R') which is represented by 2 of the 3 codings is selected. If only one

response is coded across the 3 sections, then if that response is a typical 'I' response the informant is coded 'I', otherwise 'R'.

This scheme does not fulfil the original intentions embodied in the structure of Q31; however, it is a preferable alternative to omitting the information which was elicited.

Q32. extends the general attitudinal information derived from Q31 with respect to the informant's actual job situation. Job satisfaction, then, shows how far the informant's occupation as coded under Q29 fulfils his job preferences codedunder Q31. (OM)

Qs 33, 34 and 36 (UM, MC) deal with leisure activities and hobbies. Given the immense range of possible responses to Q36, a classification was devised which includes all combinations of the following distinctions: active/sedentary; expensive/cheap; rule-based/non-rule-based; and club/non-club. Also the attribute 'collecting' is included. These distinctions yield 24 combinations. Examples of the way in which various hobbies are classified according to this scheme are shown under Q36, social coding sheets (appendix B).

Q35. Leisure satisfaction, shows how far the informant is satisfied with his leisure activities. (OM)

Q37. shows whether the informant believes that political allegiance should be associated with occupational group membership. (UM)

Q38. Voting preference. In addition to indicating informants' political allegiances, this question, taken together with the previous one, and with Q29 shows whether the belief that occupation (or class membership) should be connected with voting behaviour is associated with particular levels of

occupational status in the sample, and whether, for instance, approval of such a connection is more strongly associated with the Labour than the Conservative vote. (UM)

The social coding frame which was applied to the present sample of 52 Tyneside informants, then, is defined by social features shown in appendix B, subject to the modifications outlined above. To summarise these modifications;

1. 4 questions were eliminated because of low response rates;
2. redundant categories were eliminated;
3. some questions were restructured, e.g. Q23b was reduced from a 10- to a 5-point scale, as the original distinctions were over-differentiated, and too much classificatory weight would have been assigned to this feature if all the 10 distinctions had been preserved;
4. the 3 distinctions in Q31 were collapsed into one, as response rates were too low on the separate sections of this question;
5. new categories were incoporated, where relevant, (e.g. to accommodate the response 'abroad' (Q25 - macro-environmental preference)).

These modifications were made specifically for this sample: they do not represent a restructuring of the TLS coding frame per se, but just a fitting of this coding frame to the particular sample under study.

The social data for the sample of 52 informants was coded with respect to the 153 binary variables of the reduced SocSp, according to the OM, UM and MC conventions (see above, pp. 136 - 140) punched on cards, and input to MTS line files.

The social data file was then appended to the processed linguistic data file for each informant, to be used in the CLUSTAN mixed mode data runs.

CHAPTER 5

THE SOCIAL CHARACTERISTICS OF THE SUB-SAMPLE

The sub-sample of 52 consists of 45 informants resident in Gateshead, (the largest centre of population on the south bank of the Tyne), and 7 informants from Newcastle upon Tyne, (north bank).

The Gateshead (Phase 3) sample was drawn as a stratified sample, the stratifying factor which was used is "rateable value per dwelling by polling district" (Pellowe, Nixon, Strang & McNeany (1972: p.27)). 150 informants were selected, from 5 strata. The 45 Gateshead informants (in the sub-sample of 52 under investigation here) exhaust the 2 lowest strata of the sample in terms of this stratifying factor. These strata are defined thus:

Stratum 4: rent= £4+ per week (11 informants),

Stratum 5: other council house (34 informants).

Our expectation, then, is that these 45 informants will display some characteristically working class social attributes, if the stratifying factor is a reliable index in relation to social class.

The other 7 informants (distinguished by the 'X' prefix to their mnemonics, e.g. 'XFULT'), are part of the Phase 2 sample, drawn from the C.B. of Newcastle upon Tyne. Unfortunately, the data for the entire Newcastle sample was not available for processing when this research was begun.

The Newcastle sample was drawn by the 'clustering method' of sampling, (Moser: 1958), where every kth name from some relevant list, (in this case, the Electoral Register) is selected.

The sub-sample of 52 informants dealt with here, then, in no way constitutes a random or representative sample of the statistical population. This investigation is concerned with sociolinguistic variability within one (broad) social class, (as defined by the index used).

The informants from Newcastle are included as a test component, for 2 reasons:

i) by introducing informants from the other side of the Tyne, we can test for geographical correlates of linguistic variation,

and ii) by introducing an element which is partially non-working class, we can discover whether and how far, the different social classes tend to cluster separately (on social, and linguistic features).[FN]

FN. If, for instance, the working class element of the Newcastle sample clusterswith the Gateshead informants, then we can surmise that geographical factors are less significant than social class factors in this case.

The following section gives a brief account of the social profile of the sub-sample of 52 informants, in terms of sex, age, education and occupation.

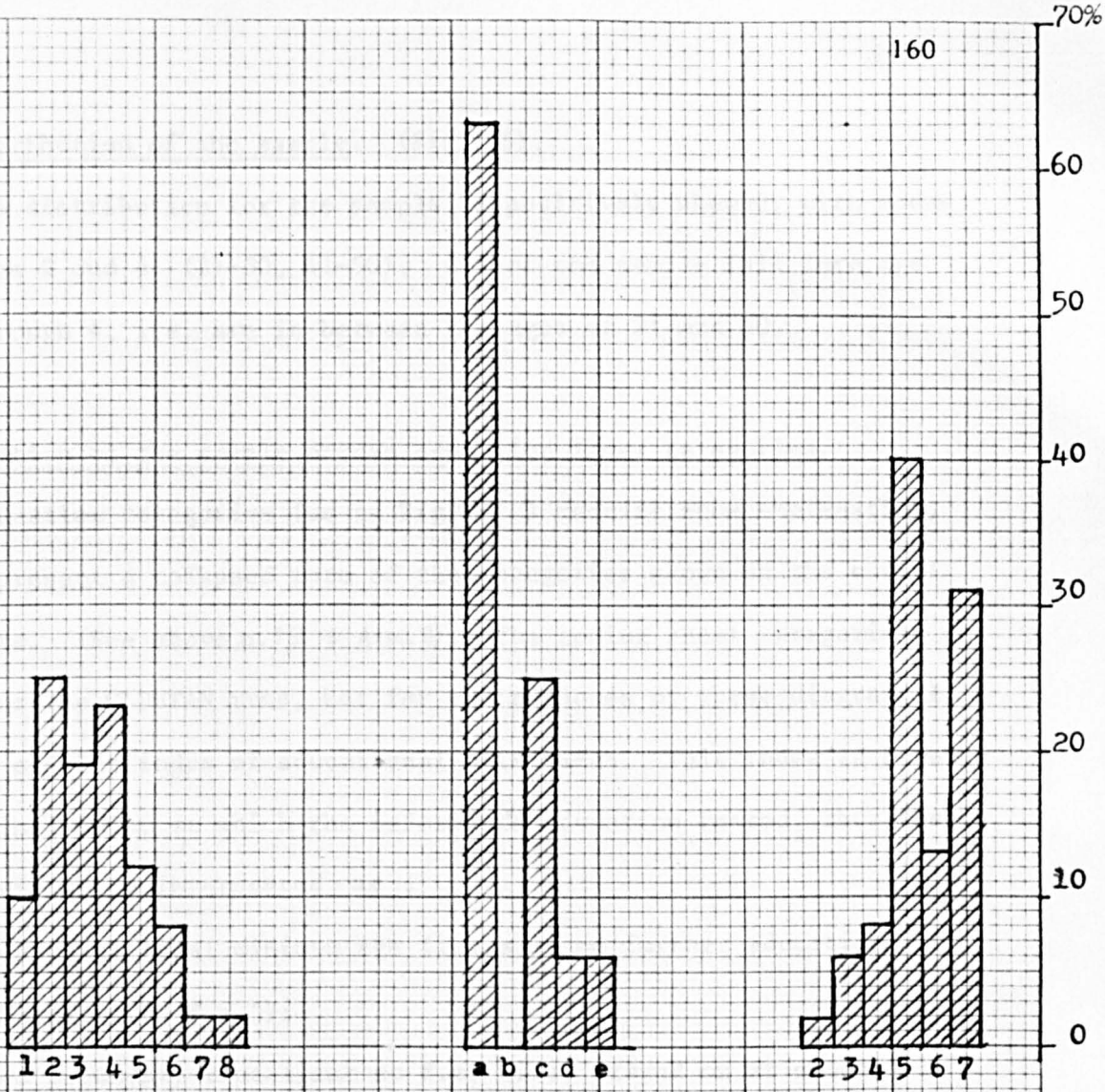
The sub-sample of 52 informants,[FN] consists of 25 females and 27 males, ranging in age from 17-21, to 80+ years, and ranging in educational background from those who left school at the legal minimum age (and have undergone no further education), to informants having taken full time university or polytechnic courses.

FN. Hereafter referred to as 'the sample'.

Six occupational groups are represented, groups 2 through 7. (See above p. 154 for definitions of occupational groups).

The shaded histograms (Figs. 36, 37, 38 (p.160)).
show the distribution of the sample with respect to age groups, education level, and occupational groups respectively, in terms of the proportion (%) of the sample belonging to each category.

Thus, 10% of the sample belong to age group 1, (17-20), 25% to age group 2 (21-30), and so on.



Age groups	Education index	Occupation groups
1 17-20	a lma, no fe	2 manag/exec.
2 21-30	b lma+, no fe	3 inspect/super (higher)
3 31-40	c fe - tech.	4 " " (lower)
4 41-50	d full time	5 skilled manual
5 51-60	e " "(acad.)	6 semi " "
6 61-70		7 unskilled "
7 71-80		
8 80+		

Fig. 36 Age distribution

Fig. 37 Education

Fig. 38 Occupation

Percentage frequency representation of age groups, education index categories, and occupation groups, across the sample.

The age distribution of the sample. (Fig. 36).

The age distribution for the sample is positively skewed, with modes at age groups 2 and 4 (21-30, 41-50). 67% of the sample fall into age groups 2 through 4, i.e. are in between the ages of 21 and 50.

The distribution of the sample across education index categories.

The education categories (as in Fig. 37) require some explanation, as these represent a collapsed form of the categories shown on the social coding sheets. (See above,p.151 & Appx.B). The coding sheet categories were used for the CLUSTAN runs, but for the purposes of these diagrams I have used a simpler index of educational attainment. This index is based on the terminal point at which the informant's education ceased, thus the categories are to be interpreted as .:

a - left school at legal minimum age (l.m.a.), no further education (henceforth 'f.e.');

b - extended secondary education: no f.e. (i.e. stayed on at school beyond legal minimum leaving age);

c - education continued into working life, (block release, day release, self-taught, correspondence courses, night school);

d - full time technical college, nursing, secretarial college;

e - full time academic training: college of education, university, polytechnic.

This index covers the information derived in Q's. 8 and 9 (see social coding sheets, Appx. B) but represents a reduction of that information in 2 ways

i) categories have been conflated;

ii) intermediate educational history is ignored.

In the social space which is the basis for the CLUSTAN classifications, important distinctions which are lost to this index are retained. For example (see ii) above), there may be a significant distinction between ORMST, who left school at the legal minimum age, but proceeded to study

at night school, and through day release, and TURRL who also proceeded to further study (night school), but who had stayed on at school two years beyond the legal minimum leaving age. This distinction is preserved in the social space, but such distinctions are ignored here for the sake of graphic simplicity.

The shaded histogram (Fig. 37) shows that the majority of informants in the sample left school at the legal minimum age (63%). Category b, (extended secondary education) has 0% representation because all informants who stayed on at school also proceeded to f.e. of some description. Thus, one or more extra years at school was not, in any case, the terminal point in an informant's educational history.

Of the 37% who were educated beyond the legal minimum requirements, 25% furthered their education by block release, day release, correspondence courses, night school, or were self-taught.

6% were trained full time at technical or secretarial college, or in nursing.

6% attended college of education, university or polytechnic (full time).

The distribution of the sample across occupation groups.

The shaded histogram (Fig.38p160)shows the distribution of the sample across occupation groups 2 through 7.

Here there is a negative skew, with group 5, (skilled manual and routine non-manual) and group 7, (unskilled manual) showing modal tendencies.

This bias in the sample results from the fact that, of the 200 informants (drawn originally as a representative sample), 45 of the 52 whose data were available for processing are overly representative of the working class strata of the Gateshead sub-sample.

Given that these 52 do not constitute a random sample of the population of Tyneside, then the social characteristics of sub-groups (clusters) must be constructed on the basis of expected values based on the characteristics

of this whole sub-sample.

Thus the histograms showing the within-cluster distributions across coding categories (for age, education and occupation) are meaningful in relation to the sample distribution across the same attributes. (See below, pp. 167ff.)

The Social Space: a clustering of 52 informants.

The sample of 52 informants was analysed by clustering methods, in terms of the social attributes outlined above, (pp. 149-157).

(See above, pp. 105ff. for a description of CLUSTAN). The CLUSTAN options used were:

(i) Distance coefficient: Binary Euclidean Distance (hereafter D), defined as:

$$D = b + c \,/\, a + b + c + d,$$

i.e. total number of mismatches in each pair-wise comparison, over the total number of attributes;[FN]

FN. where a, b, c, d refer to the standard 2-way contingency table, i.e.

	+	−
+	a	b
−	c	d

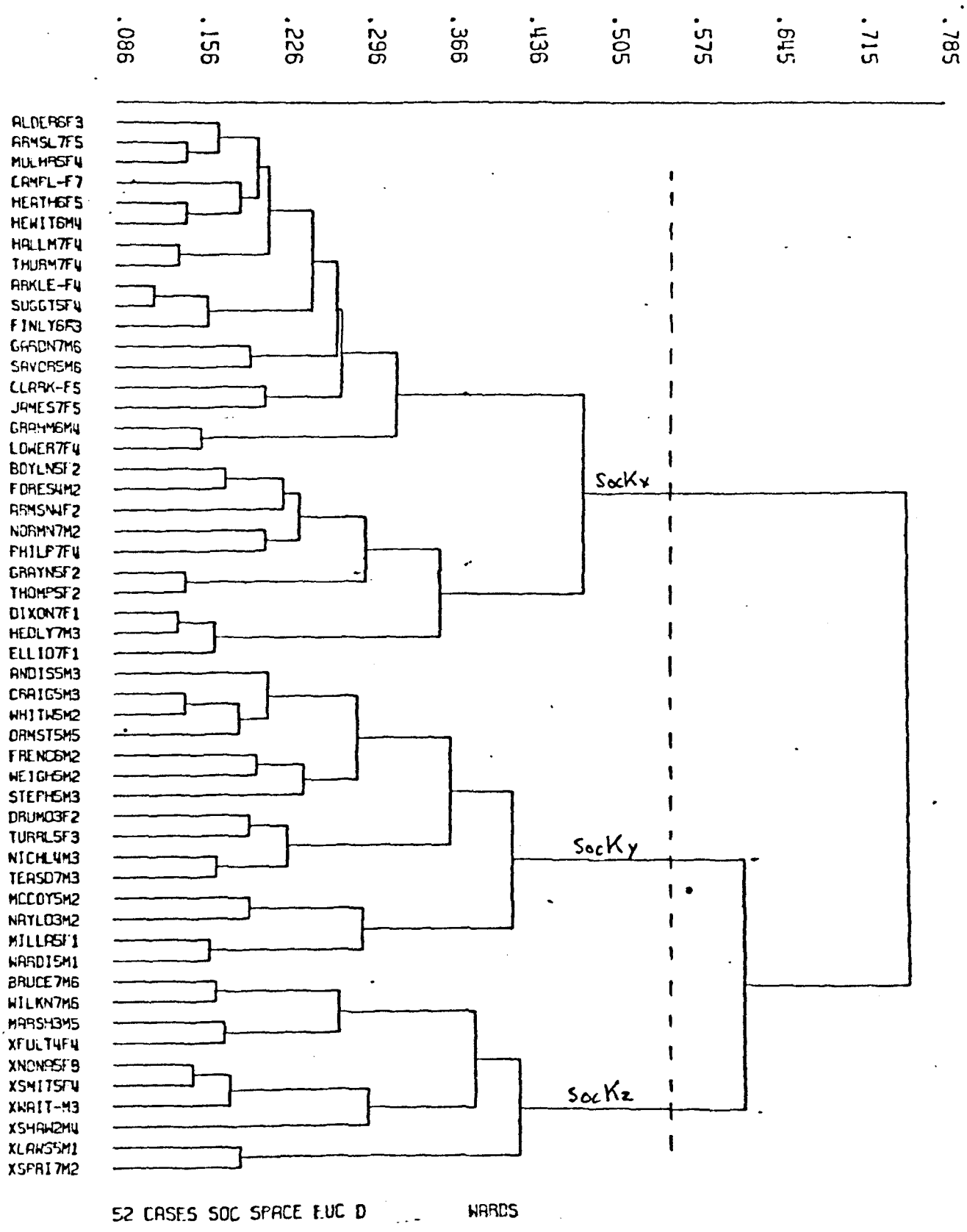
(ii) Clustering algorithm: Ward's method, (minimisation of Error Sum of Squares between clusters - see above, pp. 106-107).

Fig. 39 (p.164) shows the fusion tree output by CLUSTAN, and summarises the fusion process resulting from the classification of the sample on social attributes.

The first decision to be taken involves the question of how many clusters are present.

In other words, which point on the scale of increasing values of D should be taken as the cut-off point, (which defines the number of clusters

Fig. 39. Dendrogram of sample clustered on social attributes.



identified, and the membership of those clusters).

Given the known characteristics of the sample (see above pp.158ff.) with respect to the selection criterion for the Gateshead sub-sample, we may expect that these 45 informants will cluster into two groups. The inclusion of the Newcastle sub-sample may produce a third cluster. So, the 3-cluster (hereafter, 3-K) level may be a useful level at which to examine the constitution of clusters, (in terms of cluster membership, and diagnostic statistics for variables with respect to these clusters).

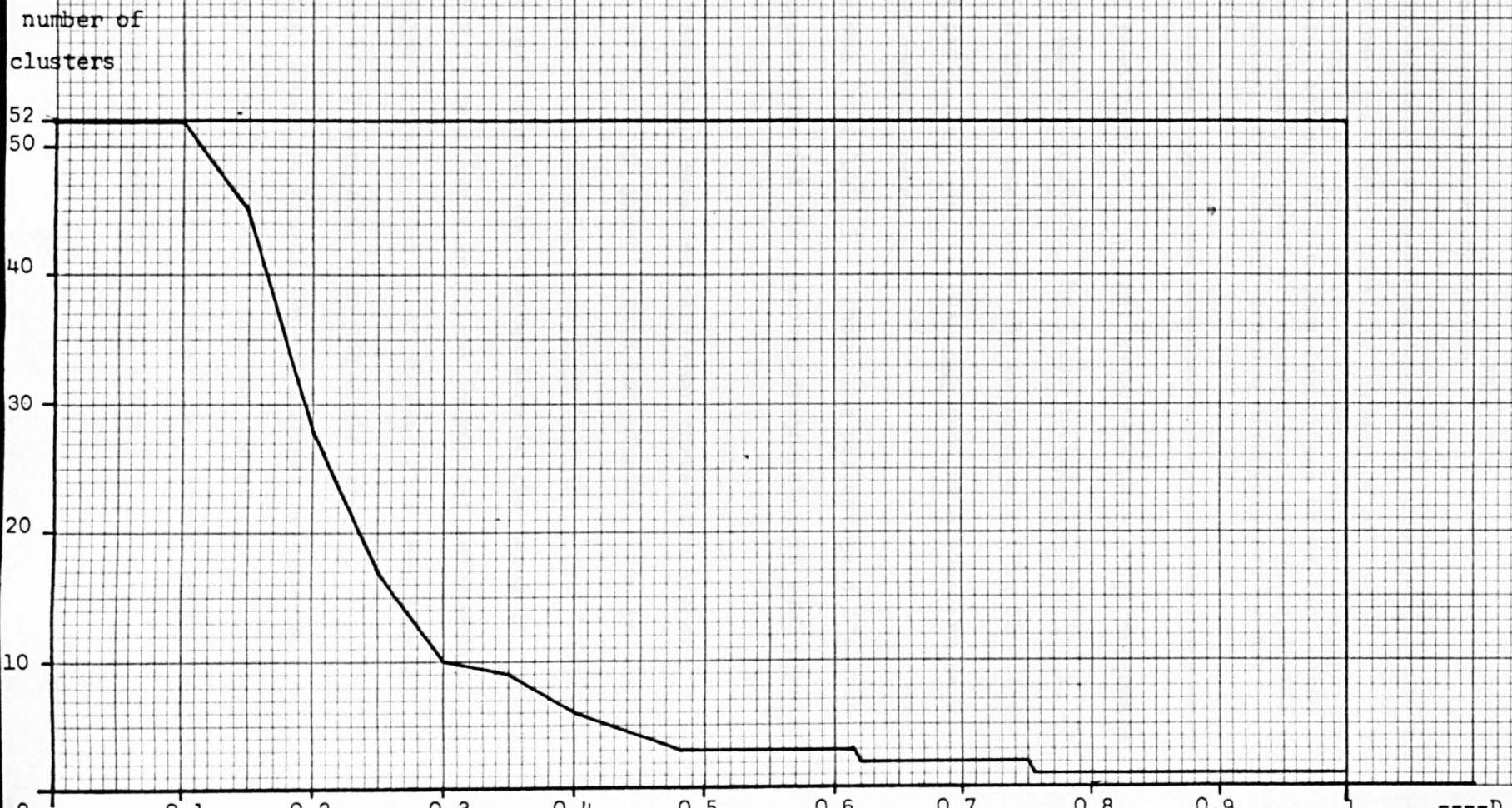
The optimum cut-off point may be identified where the number of clusters does not change across a wide range of increasing values of D, that is, the threshold value of D increases sharply before another entity (or cluster) can join any of the clusters currently existing. Fig.40(p.166) plots the number of clusters present for increasing values of D. It will be seen that the first plateau (indicating stability of the number of clusters currently present) occurs at the 3-K level. There is, then, a relatively large jump in the value of D, between the fusion resulting in 3 clusters, and that resulting in 2 clusters (0.134). This region, then, is the first break point in the dengrogram, (cf. Fig. 39 (p.164)).

There is also a significant plateau at the 2K level. This division of the sample into 2 groups, then, must also be considered as a potentially useful division. However, as will become apparent in the discussion of the properties of the social clusters, the two groups which fuse into one at the 2-K level (SocKy and SocKz) (see below, pp. 167ff.) display very different distributions with respect to age, occupation and education. These distinctions would be levelled if these 2 clusters are allowed to merge (i.e. at the 2K, rather than the 3K, level), and the distinctions of these two groups from each other, and from the other social cluster, would be submerged. So the 3-K cut-off point was chosen.

The 3 clusters are designated SocKx (Social K x), SocKy, and SocKz respecively. (See Fig. 39). There are 27, 15 and 10 informants

Fig. 40.

Number of clusters present at given levels of D (binary Euclidean Distance), plotted against ascending values of D, from 0 - 1. (Social classification.)



respectively in these three clusters.

These three clusters are now analysed with respect to their distributions across age and occupation groups, sex, and education (education index) as was the total sample (see above, pp. 159-163).

The distribution of age groups across clusters.[FN]

FN. See my remarks above (pp. 8ff.) on the inadequacy of social indices based on single (or few) social factors. The distributions of social attributes across the clusters identified shows that these remarks were well founded, in that different social variables divide the sample differently, and clusters are not discriminated by any of them absolutely.

Fig. 41 derived from Table 4 p.169) shows the percentage frequency representation of each age group across the three social clusters (x, y, z).

When compared with the age profile of the whole sample (Fig.36 p.160) these distributions suggest that the variable age must be associated with other variables in the social space.

Fig. 42 (derived from Table 5) shows the percentage difference between the sample expectation and the observed frequencies in each cluster, for each age group.

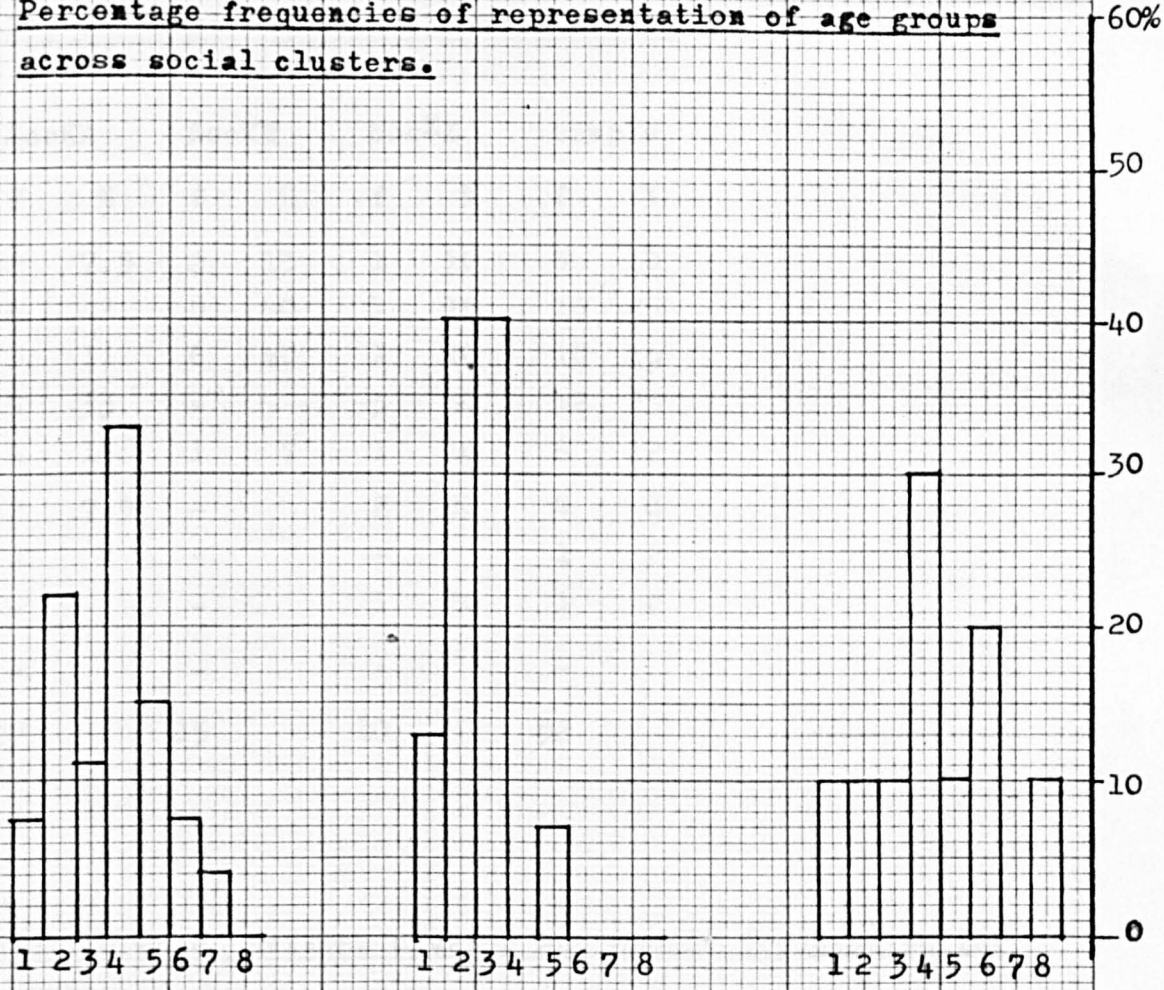
(Bars above the horizontal axis indicate that there is a higher percentage of members of that age group within the cluster, than in the whole sample; bars below the horizontal axis show that there are relatively fewer members of this age group in the cluster than in the sample.)

We find that SocKx is relatively deficient in younger informants, (17-40), and has a higher concentration of 41-50 year olds than sample expectation, (10% more 41-50 year olds than the whole sample).

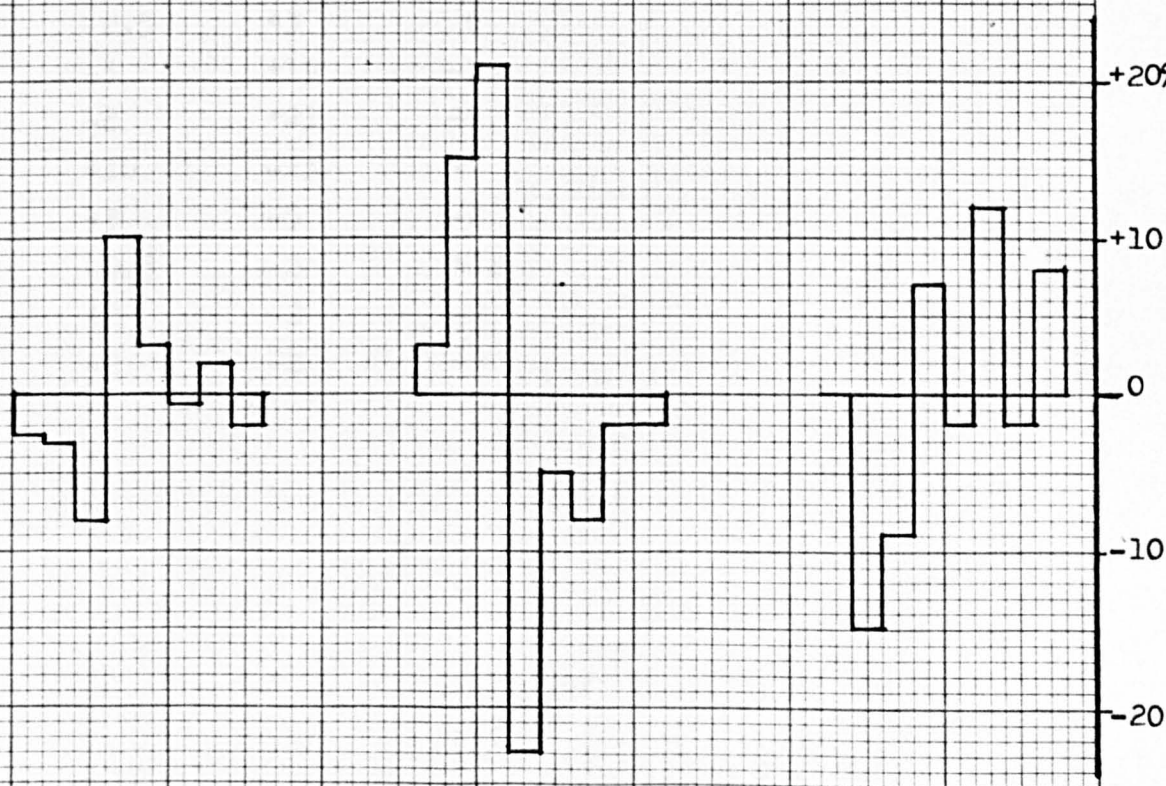
SocKy, on the other hand, is biased towards younger informants, particularly age groups 2 and 3, (21-40). All but one of the informants in SocKy is under 41 years old. (Actual frequencies, percentage frequencies and percentage differences are shown in Tables 4 & 5.)

Fig. 41.

Percentage frequencies of representation of age groups across social clusters.



SocKx SocKy SocKz



Percentage deviations from sample expectation of age distribution in three social clusters.

Table 4.

Raw and %age frequencies for age groups across social clusters, (X,Y,Z), and across sample (52 speakers.)

Age	SocKX		SocKY		SocKZ		sample	
gp.	f	%	f	%	f	%	f	%
1	2	7.5	2	13	1	10	5	10
2	6	22	6	40	1	10	13	25
3	3	11	6	40	1	10	10	19
4	9	33	-		3	30	12	23
5	4	15	1	7	1	10	6	12
6	2	7.5	-		2	20	4	8
7	1	4	-		-		1	2
8	-		-		1	10	1	2
NC	-		-		-		-	
TOT	27		15		10		52	

Table 5.

%age difference between cluster and sample frequencies for age groups.

age gp.	SocKX % diff.	SocKY % diff.	SocKZ % diff.
1	-2.5	+3	0
2	-3	+15	-15
3	-8	+21	-9
4	+10	-23	+7
5	+3	-5	-2
6	-0.5	-8	+12
7	+2	-2	-2
8	-2	-2	+8

There is a general dearth of octogenarians in the sample: the only one is found in SocKz, which cluster is relatively deficient in age groups 2 and 3, (21-40s, who tend to congregate in SocKy). SocKz has relatively more informants in their forties and sixties than statistically expected.

Generally, then, SocKx is characteristically middle aged, SocKy is a more youthful group, and SocKz is mixed, tending towards the middle aged, and the old, yet with a relative frequency higher than the sample frequency for age group 1, (17-20).

SocKy and SocKz show distinctly different age distributions. The decision to take the 3-K rather than the 2-K-level (at which point these two would have become one cluster) is supported by these age distributions.

SocKz, unlike the other two clusters, does not show a clear age trend: the shape of the histogram (Fig. 42) is quite chaotic. It is noteworthy that all age groups (except 7) are represented in this cluster; it may be that SocKz is a more socially heterogeneous group.

Evidently, though, for SocKx and SocKy, age trends do exist: thus we can conclude that responses to questions in the interview differed on the grounds of age, for these two groups, but did not differ so clearly for members of SocKz.

Sex distributions across clusters.

Of the 25 females in the sub-sample, 19 are found in SocKx, 3 in SocKy and 3 in SocKz.

Thus we have the sex = male ratios:

$$\text{SocKx} \quad \frac{8}{27} = 30\%$$

$$\text{y} \quad \frac{12}{15} = 80\%$$

$$\text{z} \quad \frac{7}{10} = 70\%$$

There is a clear sex distinction here, between SocKx, and SocKy, SocKz. Evidently the sociological facts, and social attitudes elicited in the interviews are to some extent sex differentiated as well as age differentiated.

Distribution of education index categories across clusters.

Fig.43 p.172 shows the percentage representation of education categories across clusters (see also Table 6 p.173).

Fig. 44 shows, for each cluster, the magnitude of deviations from the sample percentage frequencies, for each education index category. (This is the simplified education index: see above, p. 161

This graph is derived from Table 7 (p.173).

We see that SocKx is predominated by the minimally educated (category a, 21/27).

SocKy is split into the minimally educated, those with occupationally oriented further training, and those with higher academic education. (5/15 left school at legal minimum age, (a);8/15 continued study into working life (c);2/15 studied full time at university, polytechnic, or colleges of education (e) .)

SocKz splits into the minimally educated, and those who have undergone full time tertiary education. (A bi-modal distribution). (6/10 in category a, 2/10 in category d, 1/10 in category e).[FN]

FN. One informant (XSMIT) was coded NC on Q's 8 and 9.

Fig. 44 shows that SocKx has a higher proportion of informants (than the sample) in category a (minimally educated). This is also true (to a lesser degree) of SocKz, but this cluster also contains a relatively higher proportion of more highly educated informants, (categories d and e). It is perhaps significant that SocKx is predominantly female. (All informants in SocKx left school at the legal minimum age, 6 resumed their education later, 5 in category c, 1 in d).

SocKy has relatively fewer informants in category a. Those with in-job training (c) are highly represented, as is category e.

Thus, each of the three clusters is mixed with respect to the categories of this education index, but we can say generally that the

Fig. 43,

Percentage frequencies of representation of education index categories across three social clusters.

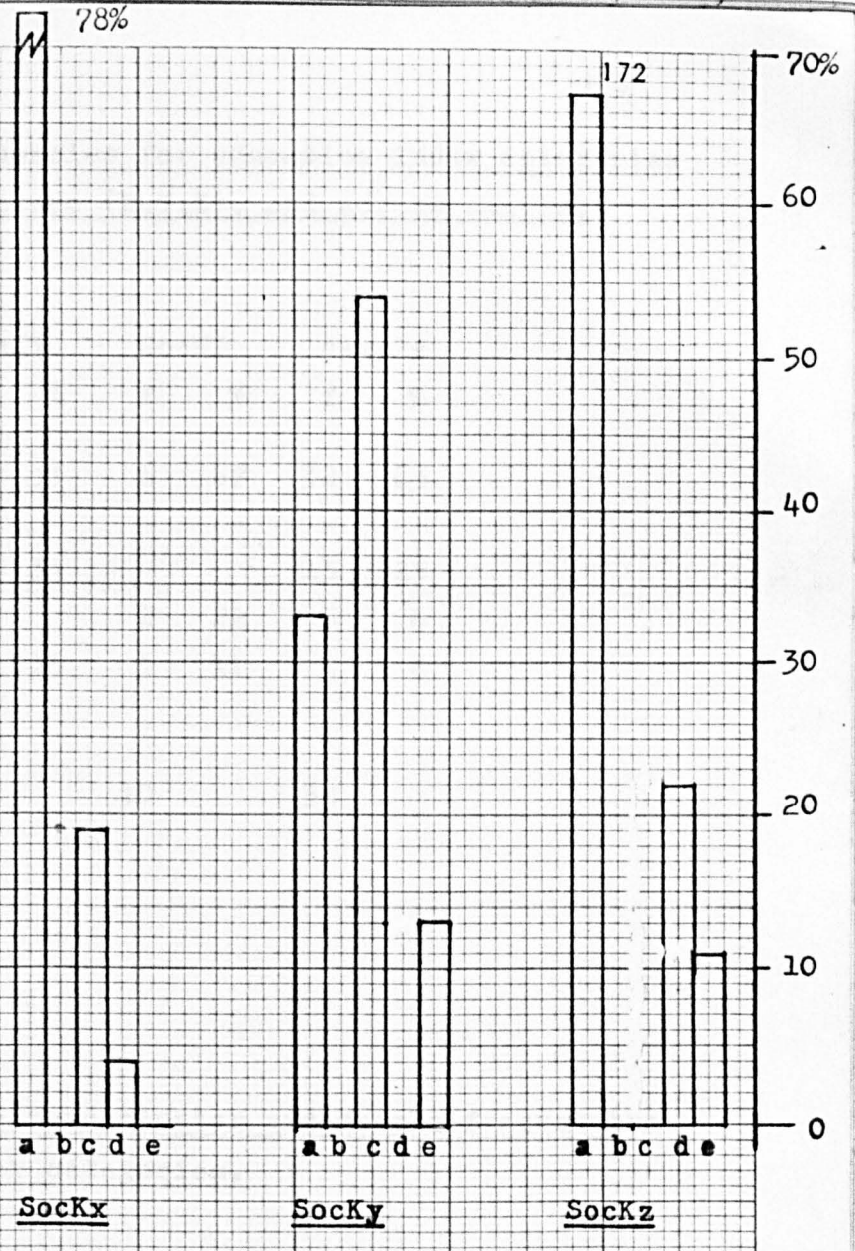


Fig. 44.

Percentage deviations from sample expectation for distribution of education index categories across three social clusters.

Table 6.
Raw and %age frequencies for education index categories across social clusters (X,Y,Z), and across sample.

Educ.	SocKX		SocKY		SocKZ		sample	
index	f	%	f	%	f	%	f	%
a	21	78	5	33	6	67	32	63
b	-	-	-	-	-	-	-	-
c	5	19	8	54	-		13	25
d	1	4	-		2	22	3	6
e	-		2	13	1	11	3	6
NC	-		-		1		1	
TOT	27		15		10		52	

Table 7.
%age difference between cluster and sample frequencies for education index categories.

Educ.	SocX	SocKY	SocKZ
index	% diff.	% diff.	% diff.
a	+15	-30	+4
b	---	---	---
c	-6	+29	-25
d	-2	-6	+16
e	-6	+7	+5

membership of SocKx tends towards the lower educated, whilst SocKy and SocKz tend towards the further educated, with full time academic training more highly represented in the latter, and vocational training in the former.

The distribution of occupation groups across clusters.

If we look at the variable, 'informant's present occupation', we find a situation analogous to that which emerged from the data on education. Consideration of Fig. 45 & Table 9 (pp.175, 176), shows that

SocKx (predominantly female, lower educated, and middle aged) has the lowest occupation group modal, (7:unskilled manual).

Note that this does not mean that there is an interference effect due to the typical career history of women in this age group: (fewer opportunities, careers disrupted by child rearing etc.) as this question (Q.29) refers to primary breadwinner's occupation.

Occupation group 5 is modal for the total sample. This is also the modal value for SocKy.

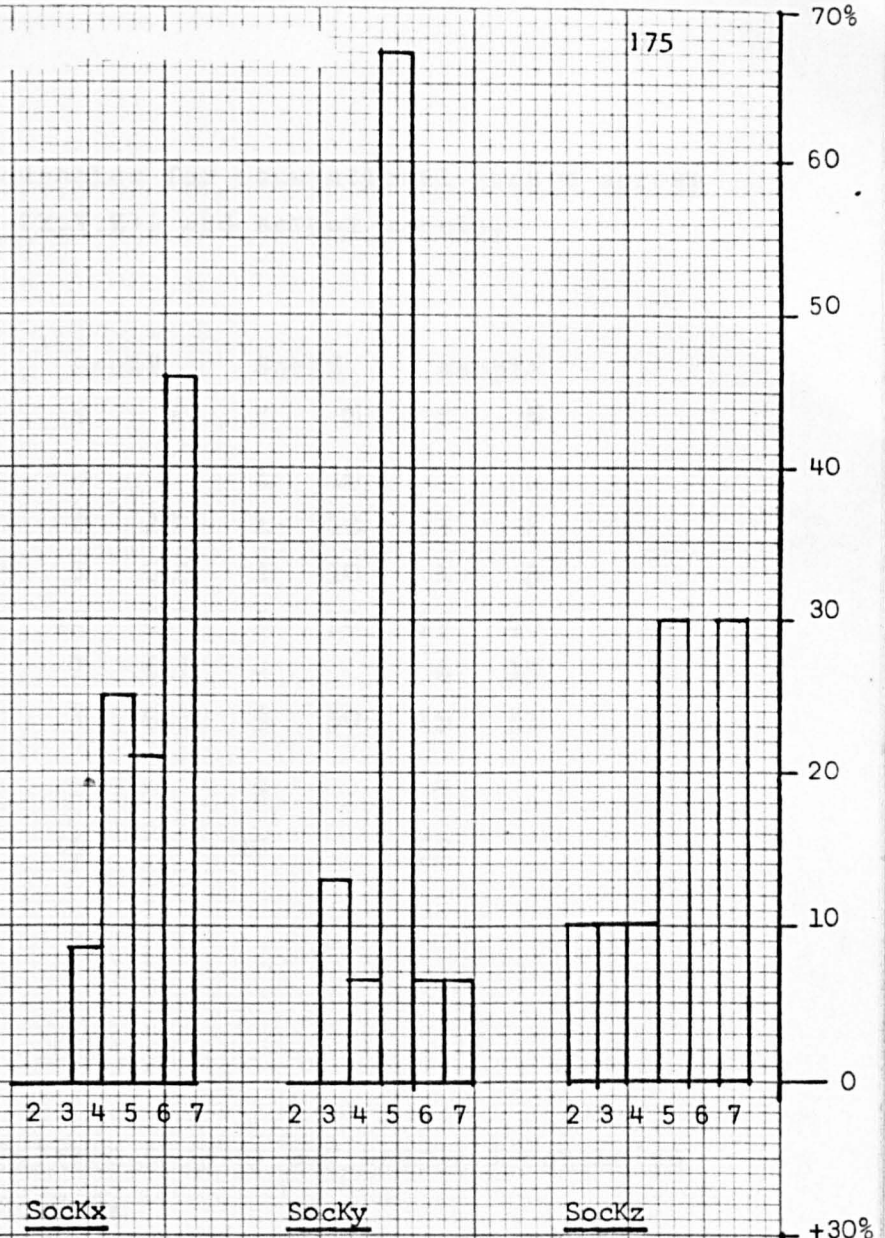
SocKz has again a bi-modal distribution, splitting between groups 5 and 7, (although the cluster membership is spread over all occupational groups except 6).

If we compare the percentage frequencies in occupation groups in each cluster, with the total sample values for each occupation group, (see Fig. 46 derived from Table 9), we find that the cluster characterised by the lower educated (SocKx) shows a trend towards lower occupational status, (occupation group 7). In contrast, SocKy, predominantly male, and highest on vocational training, tends towards occupation group 5 and higher. SocKz shows a clear trend towards higher occupational status in comparison to total sample values, but again, characteristically, is spread out across the whole span of categories.

With this variable, as well as with those already examined, we find that clusters display modal tendencies, but their membership is mixed with

Fig. 45.

Percentage frequencies of representation of occupation groups across three social clusters.



Percentage deviations from sample expectation of distribution of occupation groups across three social clusters.

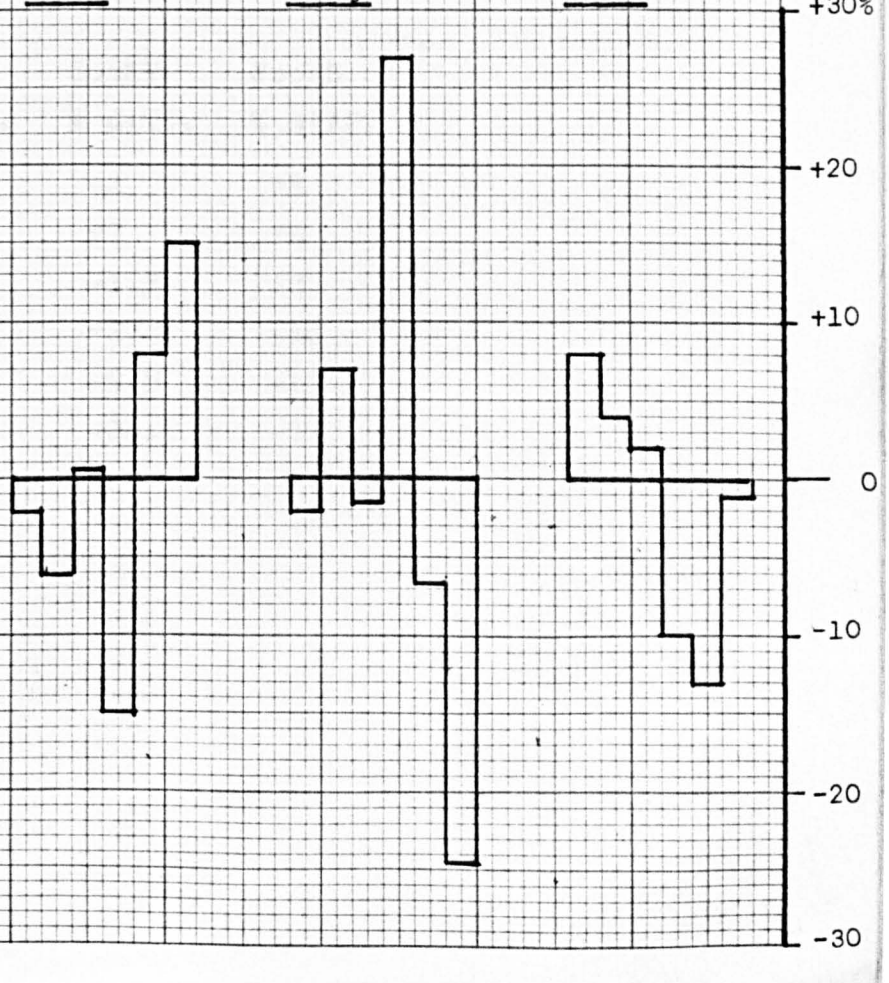


Table 8.

Raw and %age frequencies for occupational groups across social clusters (X,Y,Z), and across sample.

occ. gp.	SocKX f	SocKX %	SocKY f	SocKY %	SocKZ f	SocKZ %	sample f	sample %
2	-		-		1	10	1	2
3	-		2	13	1	10	3	6
4	2	8.5	1	6.5	1	10	4	8
5	6	25	10	67	3	30	19	40
6	5	21	1	6.5	-		6	13
7	11	46	1	6.5	3	30	15	31
NC	3		-		1		4	
TOT	27		15		10		52	

Table 9.

%age difference between cluster and sample frequencies on occupational groups.

occ. gp.	SocKX % diff.	SocKY % diff.	SocKZ % diff.
2	-2	-2	+8
3	-6	+7	+4
4	+0.5	-1.5	+2
5	-15	+27	-10
6	+8	-6.5	-13
7	+15	-24.5	-1

respect to the categories belonging to this variable. A social classification based on the entire range of attributes covered by the social coding frame, then, produces groupingswhich could not be discovered if a single social variable (or a small number of variables) were used to place informants in social strata, or classes. (See FN., p.167, above, and my remarks above (pp. 7ff.) on selectivity and atomism).

We can now summarise the overall impression of the social status of the membership of the 3 clusters given by the age, sex, education, occupation distributions.

SocKx tends towards the middle aged groups (41-60), is overwhelmingly female, tends towards low educational status, and low occupational status.

SocKy tends towards youth, is predominantly male, further education is vocational/technical, and occupation group 5 (skilled manual and routine non-manual) is relatively highly represented.

SocKz, also predominantly male, shows a mixed age distribution (with only the 71-80's not represented), and splits into the 2 educational extremes: those who left school at the legal minimum age (and did not proceed to further education) and those who attended full time college, university or polytechnic courses.

Other components of the social space, then, are the basis for intra-cluster similarity here. A closer look at the other social attributes measured will reveal which social factors bind this group together. And not only this apparently heterogeneous group (heterogenous are the basis of 4 classic sociological variables: age; sex; education; occupation),

the two other clusters, SocKx, and SocKy, though exhibiting trends, (discussed in the foregoing), are also by no means homogeneous with respect to any of these standard measures. (This can be seen from Tables 4,5 & 6.)

With this in mind we now undertake an analysis of the diagnostic statistics from the 3 clusters which were supplied by CLUSTAN.

The statistic is a 'binary percentage frequency ratio', defined as:

$$Pcj \;/\; Pj$$

where Pj is the percentage occurrence of the jth variable in the total sample input to CLUSTAN, and Pcj is the percentage occurrence of the jth variable in cluster c. (Thus if variable j is positive for 50% of the population, and 75% of cluster c have this variable, then the level of diagnosticity of variable j for cluster c will be 1.5).

The maximum possible diagnostic level is the inverse of the ratio:

$$\frac{\text{no. of cluster members}}{N}$$

where N is the number of cases (informants) input to CLUSTAN.
E.G. a population of 100, split into clusters, one of which has 50 members (cluster c), has as the highest possible diagnostic value 2. (I.e. where cluster c has a monopoly on one variable, the percentage occurrence in the cluster will be double that in the total population. Thus if total occurrences of variable j = 20 and total occurrences of variable j in cluster c = 20,

20 = 40% of 50 (50 members in cluster c)

20 = 20% of 100 (total cases in sample)

Thus

Binary frequency % age ratio on Vj = Pcj/Pj

= 40/20 = 2,

where V = variable.

This will be true whether there are <u>20</u> instances or <u>one</u> instance of the variable in question, where that variable is <u>exclusive</u> to the cluster. So actual raw frequencies of variables in the sample under investigation must be borne in mind when assessing the importance of diagnosticity levels from different variables.

A binary percentage frequency ratio of 1 means that the variable in question is represented in the cluster in the same proportion as in the whole sample. Thus this variable is <u>not</u> diagnostic for this cluster.

If the ratio is lower than 1, this means that this variable is less

frequently represented in the cluster than in the whole sample: it is thus a negative diagnostic (i.e. this cluster is characterised by the absence (or lower frequency of) this variable).

Cluster diagnostics - SocKx.

Table 10 lists the positively diagnostic variables for SocKx down to the value of 1.30 of the binary percentage frequency ratio, and the negative diagnostics from 0 to the 0.55 level. (The lowest possible diagnostic value a variable can take is zero, which means that this variable does not occur for any member of the cluster). The table shows the code of the binary variable, (which can be referred back to Table 1 (pp.142ff), the level of diagnosticity, the number of occurrences of the variable in question in the cluster, and in the total sample input to CLUSTAN (i.e. the T.L.S. sub-sample of 52 informants).

The maximum possible diagnostic level for SocKx is 52/27 = 1.93, as there are 27 cases (CLUSTAN's term for experimental entities i.e. informants) in cluster SocKx.

Eleven variables have the highest variable diagnostic value, 1.93, these variables occur exclusively in SocKx. More weight, however, attaches to those whose frequency in the total sample is higher, e.g.V66, (variable 66) distance of spouse's primary regionality < 50 miles, and > same local authority. All 8 instances of presence of this variable occur in Cluster SocKx.

Though not as significant numerically, we must note the presence of 3 occurrences of one parent of Midland regionality, and one of a Lowland mother or father, i.e. possible first general in-migrants to Tyneside.

At the next lowest level of diagnosticity (1.72) we have 8/9 of the positive responses to 'housework as a hobby', obviously connected with the concentration of women in this cluster.

V 137 is a significant social marker. 8 informants in this cluster

Table 10 High positive diagnostics - SocKx.

Var.	No. in K	No. in sample	level *	definition
128	1	1	1.93	leisure satisfaction=disgruntled
131	2	2	"	hobbies=7 (see social coding sheet)
136	2	2	"	hobbies=22
134	2	2	"	hobbies=15
3	1	1	"	citiness=market town
66	8	8	"	dist. spouse < 50m > local authority
4	2	2	"	citiness=other
9	2	2	"	parent's reg.=UK N Midland
10	1	1	"	" " =UK Midland
12	1	1	"	" " =UK Lowland
117	1	1	"	TV,radio?=predom. radio
125	8	9	1.72	housework as hobby
137	8	9	"	connection occup./voting behaviour
101	12	14	1.66	father's occup.=7
111	17	20	1.64	info's 1st occup.=7
105	5	6	1.61	" present " =6
77	20	25	1.55	'taste aspir'n'= indifferent
141	19	24	1.53	vote Labour
46	6	8	1.45	distinction educ. boys/girls
92	3	4	"	neighbours (integr) non-existent/known
72	9	12	"	mic. env. (housing), dissatisfied
64	9	12	"	sex bias of children= M
106	11	15	1.42	occupation=7
50	16	22	1.41	parental control=direct physical
51	5	7	1.38	" " =indirect physical
91	5	7	"	neighbours (integ.)=non-exist./unknown
121	7	10	1.35	TV viewing =intense,non-selective
48	7	10	"	parental control=direct verbal
90	17	25	1.31	+ Tyneside consciousness

* level of CLUSTAN diagnostic statistic 'binary percentage frequency ratio

Table 10 cont. Negative diagnostics - SocKx.

Var.	No. in K	No. in sample	level	definition
97	0	2	0	father's occup.=3
86	0	5	0	mac. env. pref.= south
102	0	3	0	occup.=3
28	0	11	0	lma+1 year **
108	0	3	0	occup.=4
33	0	2	0	university/polytechnic
30	0	2	0	lma+3 years
124	2	12	0.33	drinking as hobby
112	3	16	0.37	job preference= 'I'
36	1	5	0.39	fe=day release
41	4	16	0.49	attit. educ.= job oriented
26	7	27	0.52	sex=M

**lma=legal minimum school leaving age.

(out of 9 in the sample) claimed to approve of a connection between occupation and voting behaviour: this fact tells us something about the belief that political allegiance is part of one's class loyalty. Significantly perhaps, in this predominantly lower occupation/education group, i.e. lower working class, we find the tendency to vote Labour dominating (V 141). (19 out of the 27 informants in this cluster said they vote Labour as compared to only 5 others in the rest of this sample).

Next on the list of diagnostics are variables 101, 111 and 105, informant's father's occupation = group 7 (unskilled manual); informant's first occupation = group 7; and informant's present occupation = 6. (semi-skilled manual).

(Registrar General's Soc. Classes V and IV respectively)

Contrary to the working class stereotype, perhaps, (or possibly a consequence of rehousing schemes), 8 informants claim to have little or no contact with neighbours (3 cases of Social Integration with neightbours = non-existent/known, i.e. neighbours are known, but there is no social intercourse, V 92; and 5 cases (V 91) where informants do not even know their neighbours).

Regarding parental approach to controlling children, there were 16 positive responses to 'direct physical' measures, 5 to 'indirect physical', 7 to 'direct verbal', and only 3 to 'indirect verbal', i.e. reasoning with the child.

V 121. 7 out of the 10 in the sample whose daily exposure to radio and T.V. was coded as 'intense, non-selective' are found in this group.

V 90. Identification with the area is strongly positive in 17 of the members of cluster SocKx, out of 25 in the total sample.

Negative Diagnostics.

Of the significant negative diagnostics, V 112 (job preference = 'I', (shorthand for prospects, thinking and self-deciding - see above p. 154ff.))

occurs only 3 times in this cluster. i.e. 'R' is predominant (immediate gain, learned, supervised).

V's 124, 26. (drinking as a regular hobby, and sex = M), are untypical of this group, (2/12, and 7/27 respectively).

None of the informants expressed a desire to move to the south (5 did in the rest of the sample).

V.28. No one in this cluster stayed on at school 1 year after the legal minimum age (i.e. as this is an ordered multistate variable, this can be interpreted as: all 27 left school at the legal minimum school leaving age).

These diagnostics taken together present a fairly stereotyped lower working class group profile. However, it is evident from the actual frequencies of occurrence of variables (Table 10 and Tables 4, 5, 6) that the group is far from homogenous. There is a range of occupational groups, and level of educational attainment, and individuals themselves do not display uniformly stereotypic responses to the social questionnaire It is the summation of the social facts about, and social opinions of, one individual which place him with respect to the dimensions of the social space, and which 'fill out' his social profile in a more comprehensive way than any SES or class index can do.

Cluster diagnostics - SocKy

The maximum diagnostic level for SocKy, which has 15 cases, is 52/15 = 3.47, the minimum is zero.

Table 11 shows the positive diagnostics for this cluster, down to the 1.50 cutoff point, and a selection of the negative diagnostics.

V's 28, 29, 30 and 31 show that there is a higher average school leaving age for members of this cluster than for the whole sample.

3 stayed on 1 year,

1 stayed on 2 years,

1 stayed on 3 years,

Table 11 High positive diagnostics - SocKy.

Var.	No. in K	No. in sample	level	definition
107	1	1	3.47	info's 1st occup.=3
82	2	2	"	financial commit. (taste)=10
35	1	1	"	fe= college of education
30	2	2	"	lma+3 years
11	1	1	"	parent's reg.= UK Lowland
8	1	1	"	" " = UK E & W Ridings
31	1	1	"	lma+5 years
36	4	5	2.78	fe= day release
29	3	4	2.60	lma+ 2 years
133	3	4	"	hobbies=12
43	3	4	"	attit. educ. (children)=RRR
75	11	15	2.55	taste aspiration= good
102	2	3	2.32	occup.=3
140	4	6	"	vote=Conservative
93	3	5	2.08	neighbours (integr.) = antagonistic
143	3	5	"	voting preference=floater
86	3	5	"	mac. env. preference=south
124	7	12	2.03	drinking as hobby
37	5	9	1.93	fe= night school
49	5	9	"	parental control=indirect/verbal
28	6	11	1.90	lma+ 1 year
81	7	13	1.87	financial commit. (taste)=6-7
104	10	19	1.83	occup.=5
109	10	18	"	info's 1st occup.=5
129	1	2	1.74	hobbies=4
96	1	2	"	neighbours (integ.) = intimate
112	8	16	"	job preference ='I'
6	1	2	"	reg= UK London SE
142	1	2	"	voting preference=refusal
33	1	2	"	fe= university/polytechnic
120	1	2	"	TV/radio=non-own
62	4	8	"	sex bias of children=zero
97	1	2	"	father's occup.=3
98	2	4	"	" " =4
119	1	2	"	TV only
99	8	17	1.64	father's occup.=5

Table 11 cont.

Var.	No. in K	No. in sample	level	definition
40	7	15	1.62	attit. educ.=liberal
26	12	26	1.60	sex=M
17	5	11	1.58	no. of moves in 5 yrs. after marriage=1
123	10	22	"	TV viewing=non-intense,non-selective
139	4	9	1.55	disapprove connect. occ./vote
41	7	16	1.52	attit. educ.=job oriented
65	10	23	1.51	dist. spouse's reg.= same local auth.
84	2	7	1.49	mac. env. pref.=smaller town
73	5	12	1.45	mic. env.(housing)=satisfied ambitious
114	13	37	1.22	job satisfaction=high
67	2	6	1.16	spouse's reg. $\geq$ 50m

Negative diagnostics - SocKy.

Var.	No. in K	No. in sample	level	definition
22	1	24	0	age=40+
23	1	12	0	age=50+
90	5	25	0.70	+Tyneside consciousness

1 stayed on 5 years.

(NB. Binary variable frequencies for V.s 28 through 31 are 6, 3, 2, 1, respectively, as this is an ordered multistate variable. (See above (p.138).)

4 informants proceeded to further education in the form of day release (out of 5 in the total sample) , and 5 by attending night school.

Actual figures for tertiary education categories for this cluster are: (Q9, Tertiary and further education).

	frequency
V. 32 none	5
V. 33 full time Univ/Poly.	1
V. 34 full time nursing, secretarial, tech. coll.	0
V. 35 Coll. education	1
V. 36 Day release	4
V. 37 Night school	5

Attitudes to education

6 cases had a negative attitude to their own education, but all but one of the 15 in this cluster had a positive attitude to their children's education.

Multiple coding is permissable on Q.10 (attitude to education). 5 informants were coded on responses 3 and 4 (liberal, and job-oriented). 2 were coded 'liberal' only, and 2 'job-oriented' only.

With respect to their children's education, informants' attitudes were generally more positive, (14/15), and (for individuals) did not always correspond to the codings for attitudes to their own education.

This could mean that some parents, though retrospectively disillusioned with their own educational experience, nevertheless retain a belief in the positive value of education for their children; or, that some of these informants are unconcerned with taking their own education further now, but are concerned with the academic and/or occupational success of their children.

e.g. TEASD, who felt negatively about his own education, but positively (in utilitarian terms) about his children. (Coded 2 + 4 = RRR and job oriented, with respect to his attitude to his children's education - V's 43, 45.)

In contrast,FRENC had a liberal and job oriented attitude to his own education, but a negative attitude to the education of his children.

In SocKy, 5 were coded liberal and job oriented,

1 was coded liberal,

5 were coded job oriented,

2 were coded job oriented and RRR,

1 was coded RRR,

1 was coded negative.

Occupation groups have already been discussed, but it is worth repeating that occupation group 5 (skilled manual and routine non-manual) is the modal value for this cluster. (This group corresponds to the Registrar General's Class III).

10 members of this cluster belong in this occupational group: 10, moreover, have their first occupation coded under this group.

Regarding job preferences: 8 informants in this cluster were coded 'I', (prospects, thinking, self-deciding); 1 was coded 'R', (immediate gain, learnt, supervised) and 6 were coded NC. Here there is a strong contrast with SocKx, where 'R' predominates.

Voting behaviour[FN].

FN. As explained above (pp.140, 146) only 4 of the original 7 categories were coded positively for any members of the sample. Thus only these 4 categories are discussed.

Voting preferences, Conservative, floater, and refusal, emerge as positively diagnostic for this cluster, whilst the Labour vote is lower than the sample expectation.

In this cluster, there is a relatively high proportion of those who

disapprove of voting behaviour being tied to occupational status.

Regarding macro-environmental preference, 3 informants in this cluster expressed a preference for the south of England, and one would choose to live abroad.

13/15 in this cluster (out of 37 in the sample), claim high job satisfaction.

V 26 12/15 informants are male.

Of the negative diagnostics, V's 22 and 23 show that all informants, except one, are younger than 40.

Only 5/15 (lower than the sample proportion) claim to have a positive sense of identification with Tyneside.

Parental control is effected more often by indirect verbal means (reasoning, explanation) than in SocKx.

The overall impression of this cluster is one of predominantly upper working class group, (cf. Registrar General's class II; or Hall and Jones' (1950) occupational grade 5 (skilled workmen)). Predominantly male, and younger than 40, this group tends towards further training of the vocational or technical kind.

Job satisfaction is generally higher, and job preferences involve the freedom to take initiative, the challenge of non-routine work, and the opportunity for career advancement. (Q.31).

Political allegiance is not class-entrenched, (cf. SocKx), in that voting behaviour is spread over 2 major parties; and only one informant believe that voting behaviour should be determined by occupational status.

Attitudes to education are generally more positive than those of members of SocKx, and a higher proportion of informants in this cluster have taken up opportunities to further their education after leaving school.

Positive Tyneside consciousness is rarer than in SocKx. Three informants would prefer to move south, one to move abroad, and one to move to a smaller town.

The general trend (in comparison to SocKx), is that members of this cluster are more career oriented. Occupational status is higher (and rising rather than static); members have a slightly less conservative attitude to education, ethnicity, social class affiliation and politics.

However, it must again be stressed that modal values for social variables have only limited predictive power: there is a considerable amount of overlap with respect to social attributes across the 3 social clusters, (i.e. particular attributes tend not to be the exclusive property of one cluster), and social attributes are variable within, clusters, (i.e. very few attributes are shared by all members of a cluster). This is not unexpected: as predicted, we have not discovered any key diagnostics defining exclusively (and exhaustively) the membership of any one cluster.[FN]

FN. I.e. we have found no 'necessary and sufficient' criteria for group membership.

It may be possible to use single social dimensions to divide the population into convenient categories (e.g. occupational groups), but this achieves a one-dimensional, and severely limited, foundation for grouping a sample. The social attribute-set composing the dimensions of the social space implemented here provide, on aggregate, a fuller profile of an individual's social set, which inevitably, (and realistically), proves to be non-stereotyped.

The two clusters, SocKx, and SocKy, together include 42 of the 45 Gateshead informants. As expected, there has emerged one predominantly lower, and one upper, class group, although these two groupings are not discrete with respect to any of the single social variables discussed above.

Cluster diagnostics - SocKz

As indicated in the foregoing, those informants with an 'X' prefixed

to their mnemonics are the 7 from the Newcastle sample. SocKz encapsulates all 7 of these, together with 3 Gateshead informants: BRUCE, WILKN, MARCH.

This cluster is a more heterogeneous collection of informants than the rest of the sample; as mentioned earlier (p.170), there is a much wider range of values for the social variables analysed in detail. Or, put another way, intra-cluster distances are higher, as can be seen from the distance levels at which fusions occur within this cluster. (See Fig.39,p.164.)

Consequently the diagnostics supplied by CLUSTAN (Table 12) are generally of lower numeric significance (although the diagnosticity level may apparently be as high as 5.20: as explained earlier this is a consequence of the cluster/sample ratio in terms of numbers of cases).

e.g. V 149 - diag. level = 5.20, actual number of occurrences in cluster = 1.

The lower frequency of occurrence is not only due to the low number of cases in this cluster (10), but also to its heterogeneity.

Seven of the eight age groups, and five of the six occupation groups are represented. Values for the education index are spread across the categories, with modal tendencies at the two extremes (categories a and e).

However, these ten informants have clustered together, though somewhat more loosely than is the case with the other clusters. This cluster has no general characteristics which can be deduced from an examination of Table 12 . This is not inconsistent with the notion of a 'polythetic' class (see above, p.23), and the classificatory theory underlying the techniques implemented here allows for the possibility of obtaining well-formed clusters without any definitive, or even highly predictive, diagnostics emerging. (See above, pp.177ff.)

Pairs of individuals are deemed similar on the strength of the attributes they share: however, these need not be the same attributes across different pairs.

Having identified and, as far as possible, outlined the characteristics of the three social clusters, the next phase in the analysis involves

Table 12 High positive diagnostics - SocKz.

Var.	No. in K	No. in sample	level	definition
149	1	1	5.20	parent's reg.= UK NW
152	1	1	"	occup.=3
153	1	1	"	info's 1st occup.=3
148	2	2	"	reg.= UK NW
147	1	1	"	" = UK Midland
145	1	1	"	" =Leeds
151	1	1	"	hobbies=10
150	1	1	"	hobbies=2
144	2	2	"	citiness=Merseyside
146	2	2	"	reg.=UK E & W Ridings
71	3	4	3.90	mic. env. (housing)=neutral
87	3	4	"	mac.env.pref.=north
61	2	3	3.47	nuclear family size=6+
2	2	3	"	citiness= London
34	2	3	"	fe=tech./nursing/secretarial
108	2	3	"	info's 1st occup.=4
6	1	2	"	reg.= UK London SE
120	1	2	"	TV viewing - non-owm
15	1	2	"	no. moves before marriage 5
18	1	2	"	" " after " 2
24	3	6	"	age =60+
33	1	2	"	fe= university/polytechnic
119	1	2	"	TV only
25	1	2	"	age=70+
54	4	8	"	marital status=widow
97	1	2	"	father's occup.=3
28	5	11	2.37	1ma+1 year
110	2	5	2.08	info's 1st occup.=6
86	2	5	"	mac. env. pref.=south
60	2	5	"	nuclear family size=5+
116	4	10	"	job satisfaction=fairly low
130	1	3	1.74	hobbies=5
42	1	3	"	attit. ed. (of children)=negative
102	1	3	"	occup.=3
23	4	12	"	age=50+
68	5	15	"	mic. env. pref. (sentiment)=neutral

Table 12 cont.

Var.	no. in K	no. in sample	level	definition
14	4	12	1.74	no. moves before marriage=1-3
122	4	12	"	TV - intense/selective
41	5	16	1.63	attit. ed.= job oriented
112	5	16	"	job pref.='I'
83	3	10	1.56	mac.env. pref.=rural
22	7	24	1.52	age=40+
84	2	7	1.49	mac.env.pref.=smaller town
132	7	25	1.46	hobbies=8
63	5	18	1.45	sex bias of children=F
113	3	11	1.42	job preference='R'
115	7	26	1.40	job satisfaction=medium
26	7	26	"	sex=M
29	1	4	1.30	1ma+2 years
124	3	12	"	drinking as hobby
73	3	12	"	mic.env.pref.(housing)=satisfied ambitious
21	8	34	1.23	age=30+

Negative diagnostics - SocKz.

Var.	no. in K	no. in sample	level	definition
90	3	25	0.63	+Tyneside consciousness
57	6	40	0.78	nuclear family size=2+

the derivation of clusters, (from the same sample), on the basis of scores on linguistic variables.

CHAPTER 6

THE LINGUISTIC CLASSIFICATION

The results of classifying the sample, by cluster analysis, on the basis of linguistic variables are presented here.

The classifications described here involve the variables from the segmental phonological sub-space only.

The organisation of the linguistic coding frame, and in particular, the segmental phonological data, has been discussed above. (Chapter 2 pp.37ff.)

A complete specification of segmental variables, by their 5-digit codes, and their CLUSTAN variable numbers, can be found in Appendix X . (This is the translation table output by Program TRAN, described above, pp.131ff.)

For reasons explained above, (p.105), the 542 segmental variables (state scores expressed as within-OU percentages, see above, p.93f.) are split into three separate batches, and processed in three CLUSTAN runs. Each of the three sub-spaces of the segmental phonological space thus covers a subset of phonological variables. These three subspaces are designated %FON1, %FON2, and %FON3 respectively, and cover the states subordinate to the OU's shown here:

%FON1: iː ɪ ɛ æ a ɒ ɔː ʌ ʊ u (10 OUs, 154 states);

%FON2: eɪ əʊ aɪ aɪə aʊ ɔɪ ɜ ɪə ɛə ʊə

ə3 ə4a ə2ə1 ɪ1 ɪ2 ə4b (16 OUs, 189 states);

%FON3: p b t d k g tʃ dʒ f v θ ð s z

ʃ ʒ h m n ŋ l r j w ŋ (in bound morpheme -ing)

(25 OU's, 199 states).

Thus %FON1 covers monophthongs,

%FON2 covers diphthongs, triphthongs and reduced vowels;

and %FON3 covers consonants.

By classifying the sample on the basis of these three subspaces, taken independently, we can test whether the sample behaves differently with respect to different sub-sets of variables from the same linguistic system (segmental phonology).

If the three classifications produce similar distributions of informants across clusters, then it will be demonstrated that (at least, a large) subset of variables is an adequate basis for representing linguistic variability, and an exhaustive inclusion of variables means the inclusion of redundant information.

If, however, the sample clusters differently with respect to the three subsets of variables, we can say, with confidence, that the sub-sets of variables chosen produce only partial classifications, and cannot be taken as representing overall linguistic (segmental phonological) variability. This outcome would have important consequences in relation to the practise of selective variable sampling. (See above for a discussion of selectivity in sampling of variables, (pp. 7ff.)

We have already seen, (pp. 116ff.), that the members of the sample are ranked differently on their scores on single states: we may expect, then, that classifications based on different sub-sets of state scores will not produce similar distributions of informants across clusters.

Three cluster analyses were performed on the sample of 52 informants, using the same CLUSTAN options applied to the data from the three segmental subspaces, %FON1, %FON2, and %FON3.

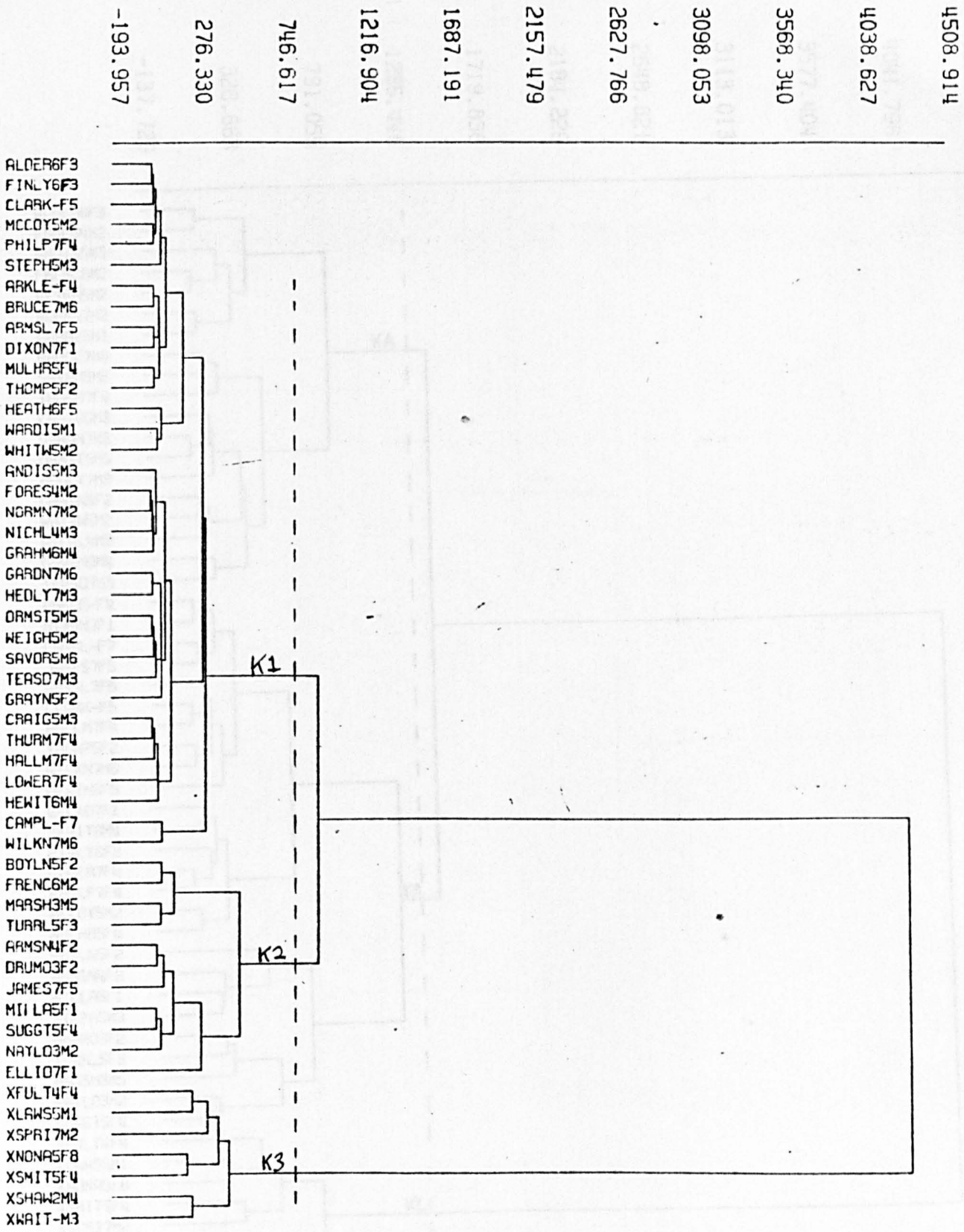
The distance coefficient used, in each case, was Squared Euclidean Distance, and the clustering algorithm used was Ward's Method (see above, pp. 106ff.)

Figs.47,48,49 show the fusion trees (dendrograms) output by CLUSTAN Program PLINK, summarising the fusion steps occurring in the three CLUSTAN runs, (%FON1, %FON2 and %FON3, respectively).

Figs. 50,51,52 show the number of clusters plotted against the value of D^2, for these 3 CLUSTAN runs. For %FON1 and %FON3, the first considerable plateau (indicating no change in the number of clusters, therefore no change in K-membership over a range of D^2 values), occurs at the 3K level. Therefore K = 3 would seem to be a useful level to take as the basis of the classification. %FON2 has a more extensive plateau at K = 4, but for the purposes of comparability between the 3 subspaces, K = 3 was taken as the significant level, especially as we have identified 3 social clusters for this sample.

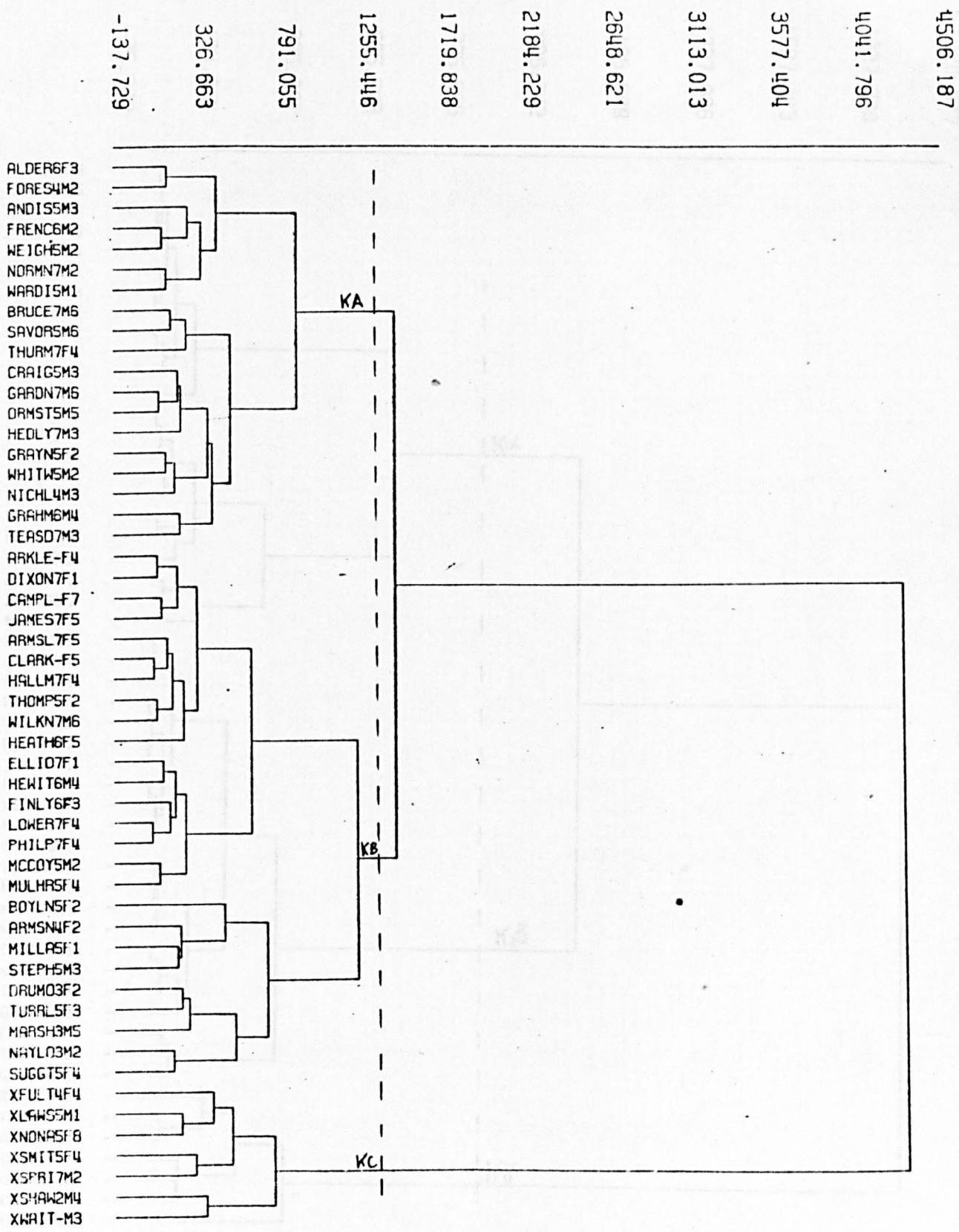
The 2-K level may also be significant: however, if for the moment we

Fig. 47. Dendrogram for % FON1 (monophthongs).



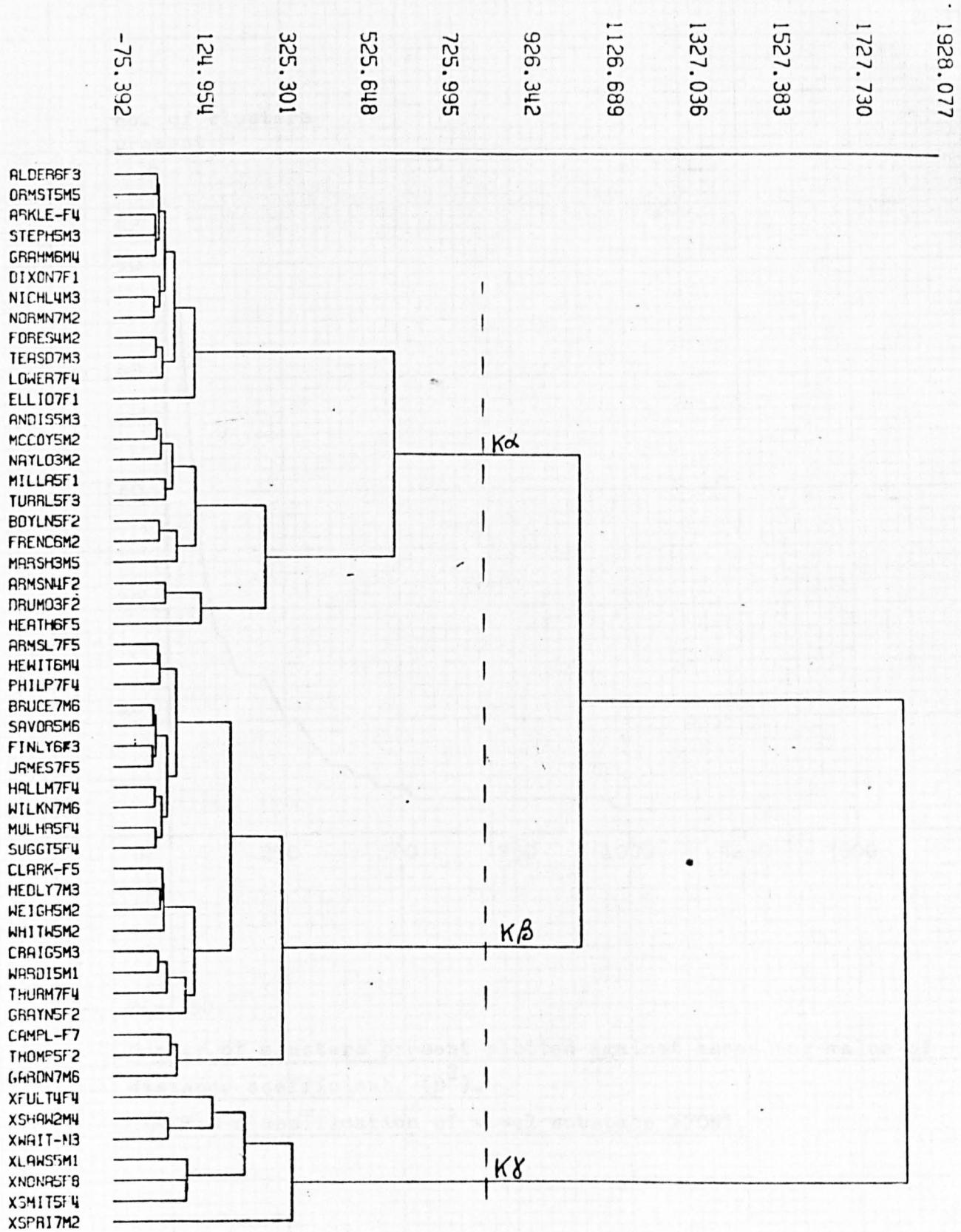
52 CASES VARS 1-154 EUC D^2 WARDS %FON1

Fig. 48. Dendrogram for % FON2 (diphthongs etc.)



52 CASES VARS 155-343 EUC D WARDS %FON2

Fig. 49. Dendrogram for % FON3 (consonants).



52 CASES VARS 344-542 EUC D^2 WARDS %FON3

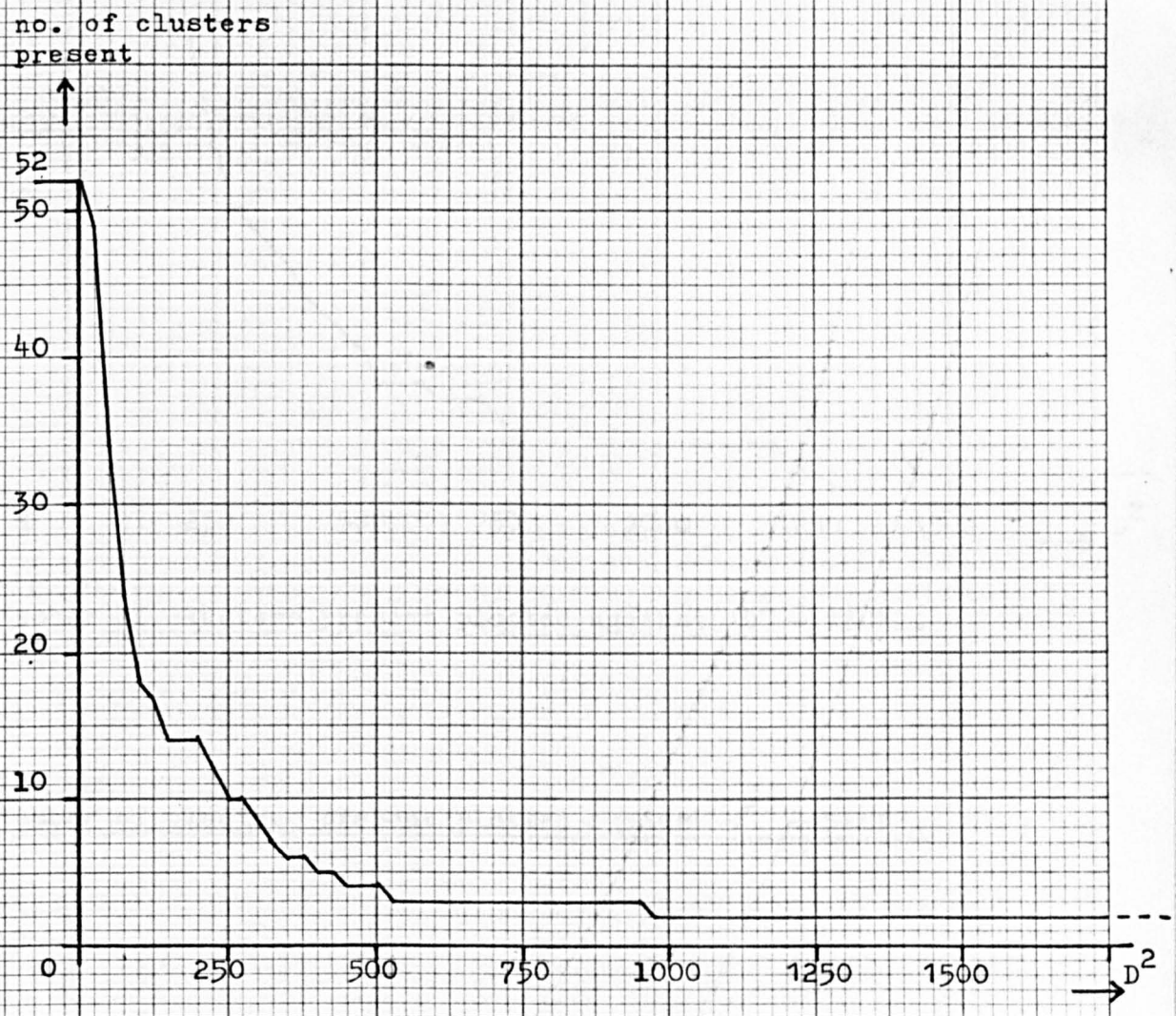


Fig. 50.

Number of clusters present plotted against ascending value of distance coefficient, (D^2).

CLUSTAN classification of vowel subspace %FON1.

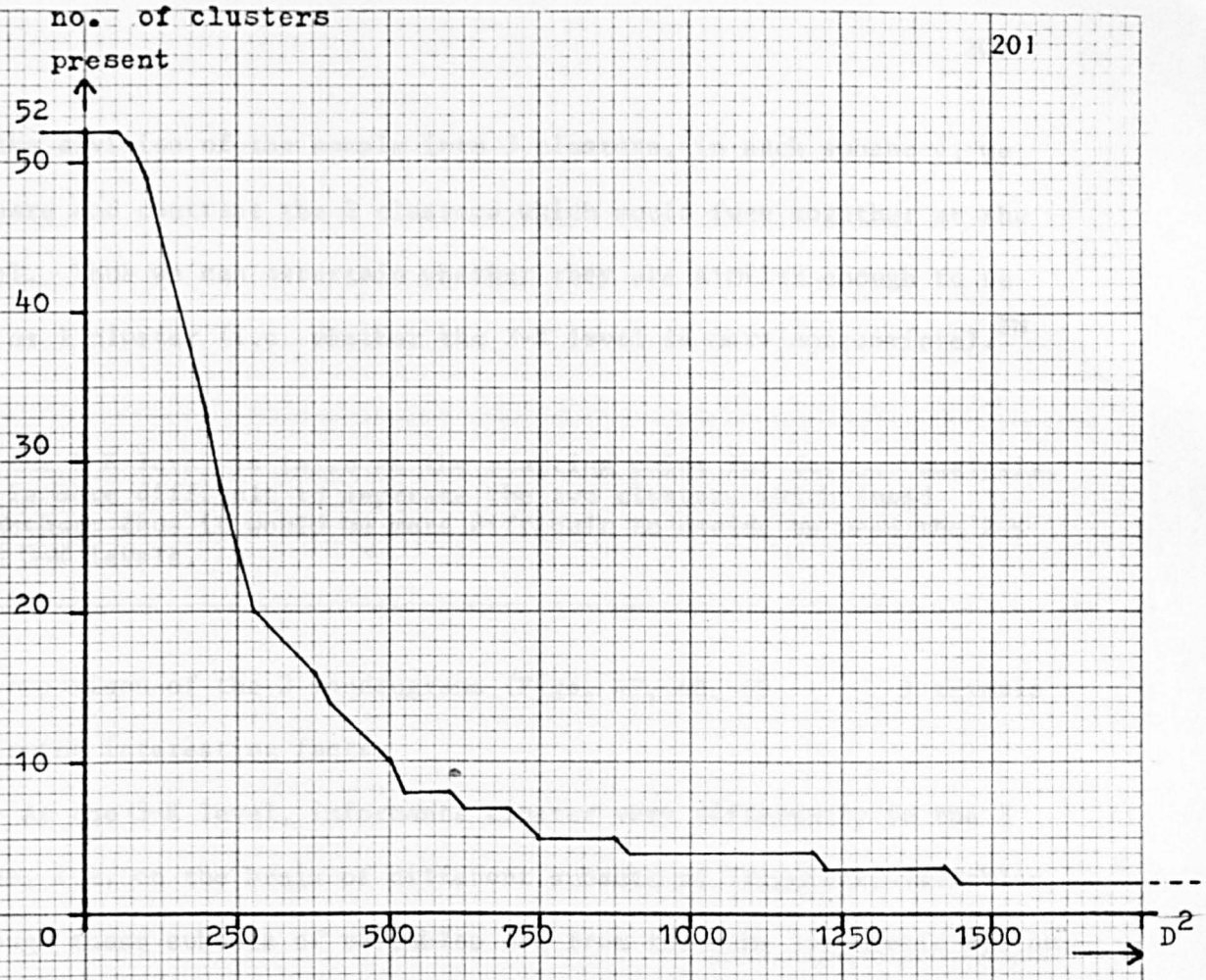
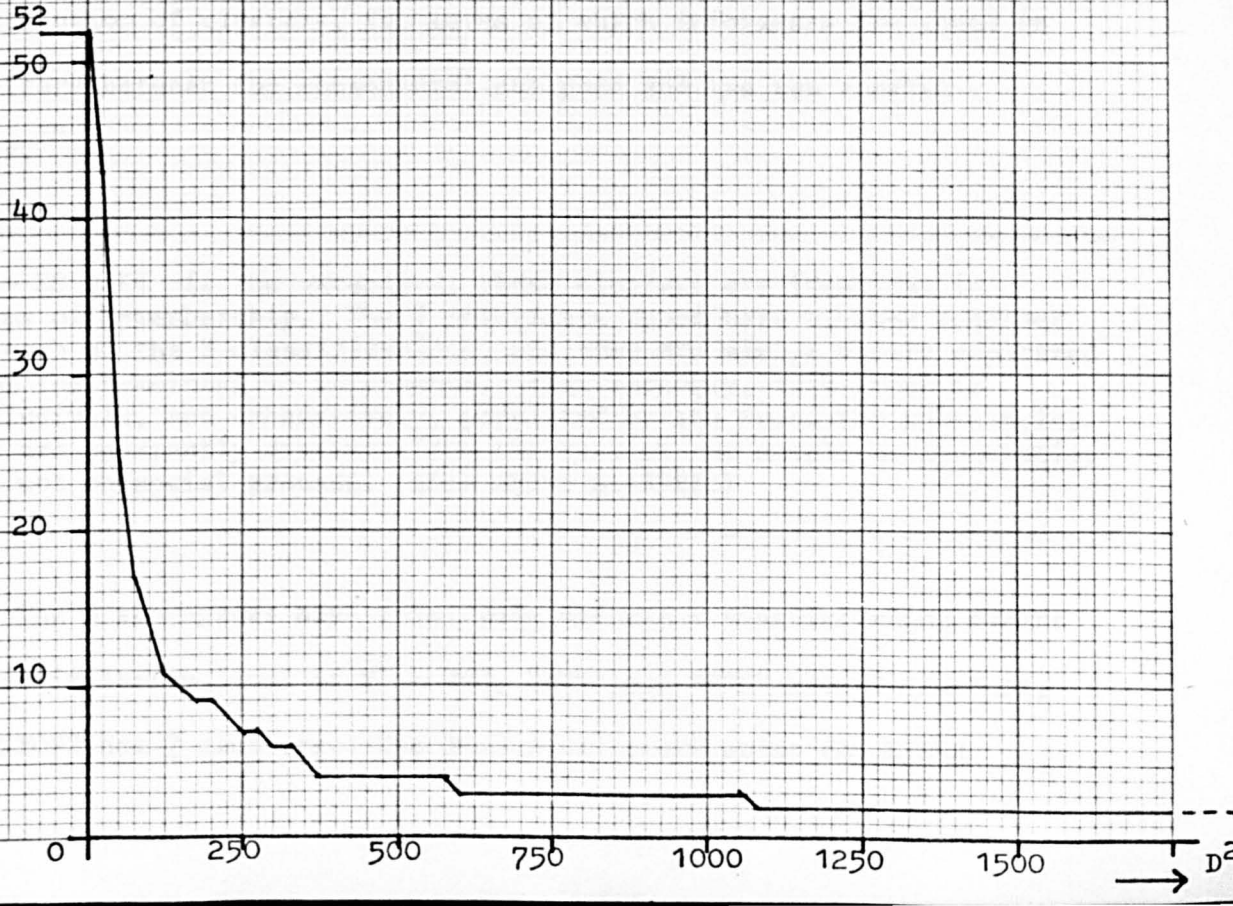


Fig. 51.
Number of clusters present plotted against D^2 - %FON2.

Fig. 52.
Number of clusters present plotted against D^2 - %FON3.



retain the division of the sample into 3 clusters, in each subspace, we can compare and contrast the 2 clusters which would fuse together at the 2-K level. Thus we can ascertain whether they are similar enough to be treated as 1 cluster (i.e. whether the 2-K level is more appropriate).[FN]

FN. If the 2-K level is taken as the starting point for further analysis, it is more difficult to separate the two clusters which fused together: thus it would be more difficult to decide between the 2-K and 3-K levels.

A comparison of the 3 dendrograms (Figs. 47, 48, 49) reveals the following interesting facts:

1. At the 3-K level, informants cluster very differently in the 3 subspaces, i.e. on the basis of different subsets of linguistic variables, even though these subsets of variables are from the same linguistic system, (segmental phonology), and are at the same analytical level, (distribution of phonetic states across OUs).

Informants cluster differently in two ways: firstly, the order of fusion of individuals differs between the three classifications; and secondly, cluster membership varies across the three classifications. So, the constitution of clusters, (in terms of which informants are found in them) differs between the consonantal subspace and the two vocalic subspaces.[FN]

FN. K3 = KC = K& is the exception: these clusters are identical in terms of K-membership. The 7 informants found here cluster together in each of the 3 classifications, and they are always widely separated from the remaining 45 informants. Interestingly, this group is exclusively, and exhaustively, populated by the Newcastle sub-sample, who are apparently distinct from the rest of the sample on linguistic, but not on social grounds. (See above p.189ff.)

---

2. The 3 sub-spaces have relatively different discriminating power: at the dissimilarity level of $D^2 = 700$, %FON1 and %FON3 show 3 clusters, whereas %FON2 has 7 clusters. The 3-cluster cutoff point for %FON2

(diphthongs, triphthongs and reduced vowels) is much higher: $D^2 = 1220$. This subspace would appear to contain more highly discriminating dimensions, as informants and clusters are more clearly separated, and individuals are distanced further from each other, on the basis of these variables.

The different overall ranges of dissimilarity levels within which the n-1 fusions occur (n = the number in the sample = 52), show, in addition, that the maximum distance between pairs in the consonantal sub-space (%FON3) is relatively lower.

	(K=n-1) first fusion level(D^2)	(K=1) last fusion level (D^2)
%FON1	19.806	4295.148
%FON2	73.335	4295.102
%FON3	15.674	1837.010

So, the outlying group, K3 = KC = K♂ , though well separated from the rest of the sample in each of the 3 subspaces, is relatively closer to the rest in the consonantal subspace than in the vocalic subspaces. (Here we see %FON2 encapsulates more variability: the closest pair are 73.335 units of distance apart).

This would appear to suggest that consonantal features are less variable than vocalic ones; also that dipthongal and triphthonal OU's display more variability than monophthongal OUs.

The differences between the 3 classifications, in terms of cluster membership, is illustrated by slides 1 to 4.(Appendix T, comprising six transparencies, numbered 'Slide 1' etc., is contained in a separate folder.)

Slide 1 shows the sample population of 52 informants represented by 3-, 4-, and 5-, character mnemonics.

Slide 2 shows the outlines of the cluster groupings obtained from the classification of %FON1 (subspace 1). If Slide 2 is superimposed over

Slide 1, the division of the sample into the 3 clusters (cf. Fig.47, p.197) is shown.[FN]

FN. The spatial locations of the informants are not plotted according to geometric relationships. This is impossible to project from 154 to 2 dimensions. And the relative spatial locations of the cases differ between the 3 subspaces. I am concerned with the topological relationships between clusters, both within, and across the 3 classifications.

Similarly, Slide 3,and Slide 4, superimposed on Slide 1 show the division of the sample at the 3K level in the classification based on %FON2 (subspace 2), and %FON3, respectively.

If all the slides are superimposed, (Slides 1 through 4), the complexity of the relationships between the 3 segmental subspaces is evident: the distribution of informants across clusters is very different in the 3 subspaces.

As observed already, one cluster maintains its identity across the 3 subspaces: K3 = KC = Kγ . This phenomenon will be descussed more fully below (p. 213, passim).

As far as the other 45 cases are concerned, the membership of K1 (Subspace 1), is split across 2 clusters (KA, KB), in subspace 2. 18 members of K1 are found in KA, the remaining 16 in KB. K2 is also split across those 2 clusters: 3 members join KA, and the remaining 8, KB.

Comparing the cluster membership of subspace 1 with subspace 3, the membership of K1 splits across Kα and Kβ in the proportion 14:20, and K2 has 9 members in Kα , and 2 in Kβ.

And the mapping for subspace 2 to subspace 3 is equally complex:

KA aplits across Kα , Kβ (9:12)

KB splits across Kα , Kβ (14:10)

The fact that the patterning within the groupings derived by the clustering process differs between the 3 segmental subspaces has important consequences on the issue of sampling of variables. (See above, pp. 7ff.,& 50ff.)

1. Selection of a sub-set of linguistic variables at the segmental level, (and therefore exclusion of the rest of the set) depletes the linguistic profiles of speakers. These results demonstrate that, by choosing different sub-sets of variables, one obtains different classifications, each of which is based on a partial linguistic profile.

Until it is known which dimensions (variables) covary significantly with social features, there can be no grounds for excluding any of these variables (or sub-sets of these variables), from a sociolinguistic classification.

2. Each of the 3 subspaces is capable of generating variety clusters (though each subspace represents a depleted and distorted domain of measurement in its selection of variables).

Each subspace, then, must contain variables which are (analytic) linguistic diagnostics, some or all of which may be sociolinguistically salient.

Although individuals are not dispersed so widely in the consonantal subspace, sub-groups are distinguished in it: consonants as well as vowels display considerable variation across the sample, and contribute to the structure of linguistic diversification within a speech community.

The fact that the distributions of speakers are non-isomorphous between the 3 subspaces strengthens the argument: if, e.g. the consonants were omitted, linguistic profiles of speakers would be correspondingly depleted and distorted, seeing that the distinct contribution made by these variables would be excluded.

If the 3 subspaces are conflated, we find that certain groups of informants belong to the same cluster as each other across the 3 classifications. For example, six informants are found together in

%FON1, K1;

and %FON2, KB;

and %FON3, K4;

(they are MCCOY, STEPH , ARKLE, DIXON, HEATH and LOWER). This group constitues a, 'derived cluster', which represents a variety cluster based on the whole segmental space. This derived cluster can be called K1B∝. If we superimpose the cluster pictures derived from the 3 classifications, we see that there are 8 such derived clusters. If slide 5 is placed over slide 1 the distribution of the 52 speakers across the whole segmented space is shown. (See Appx. T.)

The presence of these groups could not be deduced from a clustering based on any one subspace alone. The claim is demonstrated that, although a sample can be clustered on any subset of segmental variables; because different variables have different distribution patterns across a population, these partial classifications will differ. Different subsets of variables divide the sample differently, and therefore no restrictively selected sub-set of variables could produce a complete and coherent account of linguistic variation in a speech community. We would expect also, in the light of this evidence, that no subset of linguistic variables would be an adequate base for a realistic and undistorted sociolinguistic classification.

Having established the clusters in the 3 subspaces, it is now possible to examine the relative degrees of diagnosticity of the variables making up the dimensions of each subspace in terms of predicting cluster membership. Thus we can discover which variables are more powerful discriminators between variety clusters. Pellowe, Nixon & McNeany (1972), define a diagnostic feature as:

> "Some value of any given variable which is interpreted as a characteristic of a particular variety cluster (rather than an axial property of the space)" (p.5)

A variable can have a value characteristic of a cluster in at least 2 ways:

i) The cluster-mean value for the variable differs for the sample-mean value.

ii) The within-K variance of that variable differs from the sample variance.

A linguistic feature can be said to be positively characteristic of a given cluster when:

a) K-mean is higher than sample mean;

or b) within-K variance is lower than sample variance;

or, better still, where both apply.

And, to extend the argument in quantitative terms, a particular frequency of usage (or range of frequencies of usage) of a given variable distinguishes a given cluster when:

i) K-mean differs from sample mean (by being greater or less);

and ii) intra-K variance is lower than inter-K variance.

CLUSTAN provides diagnostic information on numeric (quantitative) variables, in the form of F-Ratios, and T-values, cluster means and standard deviations, for all variables used in the CLUSTAN run.

The F-Ratio is defined thus (Wishart: 1969):

$$Fj = Scj/Sj,$$

where S = standard deviation, on the jth variable,

and c = the cluster in question.

The T-value is defined thus:

$$Tj = (\bar{X}cj - \bar{X}j)/Sj,$$

where $\bar{X}cj$ is the cluster mean value for variable j.

Low F-Ratios, (< 1), indicate that the within-K variance is lower than sample variance, for the variable in question, and deviations from zero for T-values indicate that the within-K mean for that variable differs from the sample mean value. Positive deviations indicate a higher mean frequency within the cluster; negative T-Values indicate lower within-K mean values.

Variables with positive T-Values are positive diagnostics in the sense that they occur with relatively higher mean frequency: i.e. they are the

states which are preferred relatively more often by members of this cluster, than by the rest of the sample. High positive T-Values, when co-occurring with low F-Ratios, indicate that the actual value of the variable is fairly stable for members of the cluster, as well as being used relatively more frequenty by members of this cluster.

E.G. Table 15 (p.210), VAR 17; (VAR = variable);

K3 has a cluster mean frequency for VAR 17 of 82.6% Sample mean for this state is 11.6, T-Value = 2.5155.

Members of this cluster, on average, for 82.6% of instances of the lexical set associated with the phonological entity I, realised the sequent in question as i (e.g. as in 'fit').

Not only is this state the major partition for OU I , for this cluster; we know also from the F-ratio that the within-K variance for this variable is low, (F-Ratio = 0.1465)., and therefore, that this particular value of the variable (around 80%) is a stable characteristic of the cluster (relative to the whole sample).

Table 13,14,15(pp.209,210) show a selection[FN] of the cluster diagnostics produced by CLUSTAN, and the phonetic transcription of the states which the CLUSTAN variables represent. The superordinate PDV and OU are also shown for each state. Only those variables with positive T-Values, and F-Ratios less than 1 are shown, and these are listed by descending value of T-Value.

FN. CLUSTAN produces diagnostic statistics for all variables, i.e. a total of 542 sets of statistics for the 3 classifications: hence the need to select variables with the most significant F-Ratios and T-Values, as cluster diagnostics. The criteria for this selection imply one definition of diagnosticity.

Table 13.

Linguistic diagnostics, %FON1, K1.

CLU VAR	F RATIO	T VALUE	St.	5-dig code	OU	PDV	state	K-mean %	sample mean %
134	.1465	.6259	134	00843	ʊ	ʊ	ʊˈ	87.8	65.6
114	.2493	.5948	114	00742	ʌ	ʊ	o̞ˈ	61.9	46.0
123	.7253	.4846	123	00801	ʌ	I	ɨ	17.5	12.4
64	.4400	.4217	64	00421	ɑ	ɑ	ɑ̟	79.0	66.3
21	.7042	.4131	21	00161	I	ɐ	ə	24.7	19.0
14	.7333	.4044	14	00122	i:	Ii	ii	31.2	25.1
33	.6042	.3855	33	00242	ɛ	ɛ	ɛ̝	36.4	30.7
146	.7693	.3600	146	00904	u	u	ɣ	14.7	10.9
3	.3438	.3543	3	00023	i:	i:	i̞	59.3	50.7
53	.2396	.3514	53	00342	æ	æ	a̟	73.5	65.0
56	.8096	.3041	56	00345	æ	æ	ɐ̟	11.5	9.5
77	.8638	.2920	77	00504	ɒ	ɒ	ɒˈ	22.8	19.2
101	.4033	.2714	101	00626	ɔ:	ɔ	ɔ̟	57.9	50.9
19	.2762	.2722	19	00144	I	I	i̟	60.3	53.8
37	.4667	.2723	37	00246	ɛ	ɛ	ɛ	26.5	22.8
74	.5232	.2196	74	00501	ɒ	ɒ	ɒ̟	45.9	41.4
148	.8725	.1842	148	00906	u	u	ə͡u	34.9	31.1
32	.8897	.1393	32	00241	ɛ	ɛ	ɛ̟	19.0	17.4
143	.8511	.1283	143	00901	u	u	u̟	23.2	21.5

Table 14.

Linguistic diagnostics, %FON1, K2.

CLU VAR	F RATIO	T VALUE	St.	5-dig code	OU	PDV	state	K-mean %	sample mean %
19	.2771	.5549	19	00144	I	I	i	67.1	53.8
74	.4500	.4997	74	00501	D	D	D	51.6	41.4
32	.6659	.3668	32	00241	ɛ	ɛ	ɛ	21.6	17.4
148	.4592	.3534	148	00906	u	u	əʊ	38.5	31.1
30	.6637	.3394	30	00203	I	ɜ:	ɛ	13.7	10.8
76	.3014	.2561	76	00503	D	D	Dɔ	24.5	21.1
66	.8146	.2449	66	00423	ɑ	ɑ	ɑ	13.5	8.9
143	.6678	.1812	143	00901	u	u	u	23.8	21.5
3	.9956	.1022	3	00023	i:	i:	i	53.2	50.7
101	.8430	.0961	101	00626	ɔ:	ɔ	ɔ	53.4	50.9
55	.5709	.0731	55	00344	æ	æ	æ	1.4	1.2
97	.6799	.0506	97	00622	ɔ:	ɔ	ɔ	15.9	15.1

Linguistic diagnostics, %FON1, K3.

CLU VAR	F RATIO	T VALUE	St.	5-dig code	OU	PDV	state	K-mean %	sample mean %
17	.1031	2.516	17	00142	I	I	i	82.6	11.6
75	.1753	2.449	75	00502	D	D	Dɔ	77.7	14.7

%FON1 - Cluster Diagnostics

Table 13 shows the cluster diagnostics for K1 in the first subspace (%FON1).

The first column shows the number assigned to the variable by CLUSTAN; this is followed by F-RATIO, T-VALUE, and the identifier of the variable by STATE (subscript). (This is the same as the CLUSTAN variable number for %FON1, but not for %FON2 and %FON3). Cf. the remarks above on variable names, (Ch.4 & Appx. X TRAN output).

The 5-digit code with which the state is recorded for input is in the next column (for ease of reference in the specification list.

The next three columns show the OU, PDV and the state of that PDV which the variable represents:

thus VAR 134 (STATE (134)), represents the third state of the 1st PDV of OU9, ʊ, which is [ʊˈ].

K1 has a mean value for this state of 87.8%: i.e. on average, 87.8% of instances belonging to the lexical set subsumed by ʊ (in non-localised English) are realised, by members of this cluster, by [ʊˈ], (open spread). The mean for the whole sample is 65.6%.

Although all OU's are represented in Table 13 , and for each OU, the major partition(the variant state with the highest frequency, e.g. STATE (134) for OU9, ʊ) appears as a diagnostic, neither of these facts will necessarily be the case in all such tables. Thus, OUs ʌ and ɒ are not represented at all in the diagnostic list for K2. And the major partition of ɑ for this cluster, which is STATE(64), [ɑ+.] (as in, e.g. 'father'), (62.8%) does not appear in the list. This is because the sample mean frequency for this state is 66.3%, and therefore this state has a negative T-value.

The definition of diagnosticity becomes more difficult if we consider variables such as this, (STATE(64)), for K2.

This state is the majority partition of OU ɑ, for K2; thus in one sense it could be said to be a positive cluster diagnostic for K2, despite

its negative T-Value. If we select as significant only those variables with positive T-Values, then, in cases such as this, we ignore the major phonetic variant of the phonological entity in question.

In order then to characterise the linguistic profiles of clusters, as well as focus on those states which show positive deviations from sample frequencies, we need to examine variables with negative, as well as positive, T-Values.

By the definition of diagnosticity which depends on relatively lower intra- than inter-cluster variance, together with higher cluster than sample mean frequency, we find that the third cluster only has 2 cluster diagnostics.

Table 15 (K3) only has 2 entries, as there are only 2 states for K3 with F-ratios < 1, i.e. with lower intra- than inter- cluster variance, which also have positive T-values. This does not give a very full picture of the linguistic characteristics of the clusters.

I have therefore supplemented the information in tables 13, 14, 15 with tables showing, for each OU, the distribution of states realised with high frequency by members of each cluster.[FN] (Tables 16-25.)

FN. Abbreviated to "K1 use" etc.

('High frequency' is arbitrarily defined as state scores of 10%, or more.)

Where more than one PDV is represented, I have totalled the mean state scores by PDV: thus Table 16 shows not only which states are used, and in what proportions, by members of the 3 clusters, but also the proportions in which the 2 PDV's which are used are used. (Frequencies below 10% are not shown).

OU1 [i:][NL] (See Table 16 p.215)

PDV i: is the major partition for all 3 clusters, PDV Ii accounting for most other realisations. These 2 PDV's account for 91%, 89%, 87% of all realisations for the three clusters respectively. Clusters are distinct

at PDV level only by the proportions in which the PDV's are used. PDV ratios i:/Ii are approximately 6:3, 7:2, 5:3 respectively for K1, K2, K3.

Clusters are distinguished qualitatively at state level, as well as quantitatively.

K1's realisations are distributed across state 3 [i], (e.g. 'field') and state 14 [ii], (as in 'see'). K2 - state 3, state 4 [i] and state 14. K3 - state 2 [i], state 4 [i] and state 13 [ii]. Where clusters share the same state, the proportional representation for that state differs. State 3:K1 (60%), K2 (53%), K3 (0%).

K3 is more distinct from the other 2 clusters than they are from each other, in terms of states used (1 match with K2 = state 4).

It is known that this cluster is relatively distant from the other 2 clusters: (Fig.47p.197) shows a large gap in distance level before K3 fuses with the rest of the sample).

This relationship:

also holds for this set of variables (OU1), at state level, in that

K1:K2 have 2 matches and 1 mismatch in terms of states used

K1:states 3,14

K2:states 3,4,14.

K2:K3 have 1 match and 4 mismatches:

K2 : 3,4,14

K3 : 2,4,13.

K1:K3 have 5 mismatches:

K1 : 3, 14

K3 : 2, 4, 13

However, the relationships between the 3 clusters are different at PDV level for this OU:

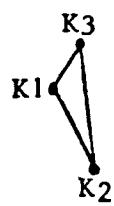
total percentage representation of PDV i:

ranks	K2 highest	69%
	K1 middle	60%
	K3 lowest	54%

with (K1 : K3) (K2 : K1) (K2 : K3).

At this level of representation, we have

PDV i:



(This ranking is reversed for PDV Ii as these 2 PDVs function as alternants for this sample).

So, even taking one speech sound, i:, we find that the level of representation (degree of delicacy of linguistic analysis) crucially affects the relationships between clusters. The PDV level produces a different classification of the clusters than does the state level, and it is therefore likely that a clustering based on PDV scores throughout the whole segmental space would structure the sample very differently.

Table 16.

The most frequent PDVs , states of OU1 i: used by members of the three clusters, K1, K2 and K3, (cluster mean frequencies.)

state	K1	K2	K3	PDV	K1	K2	K3
2 i‹			44				
3 i	60	53		i:	60	69	54
4 i‹		16	10				
13 ii			33	Ii	31	20	33
14 ii‹	31	20					

Table 17.

Cluster mean frequencies of states and PDVs used by members of K1,K2, and K3. OU2 I .

state	K1	K2	K3	PDV	K1	K2	K3
17 i«			83	I	60	67	83
19 i«	60	67					
21 ə	25	14		ə	25	14	
30 ɛ:		14		ɜ:		14	

Table 18.
OU3 ɛ

state		K1	K2	K3	PDV	K1	K2	K3
32	ɛ̝	19	21					
33	ɛ̞	36	28					
34	ɛ̞ə			24	ɛ	82	83	81
35	ɛ̝̞		14	57				
37	ɛ	27	26					

Table 19.
OU4 æ˕

state		K1	K2	K3	PDV	K1	K2	K3
52	a̞		30					
53	a̝	74	51	46				
54	a̟			38	æ˕	86	81	84
56	ɐ̽	12						

Table 20.
OU5 ɑ˕

state		K1	K2	K3	PDV	K1	K2	K3
64	a̟	79	63	10				
65	a̟ː	13		59	ɑ	92	77	84
66	a̟̟		14	15				

Table 21.

OU6 ɒ L

state	K1	K2	K3	PDV	K1	K2	K3
74 ɒ	46	52					
75 ɒː			78				
76 ɒˤ	23	25	9	ɒ	92	95	87
77 ɒˤː	23	18					

Table 22.

OU7 ɔː

state	K1	K2	K3	PDV	K1	K2	K3
97 ɔ̤	12	16	27				
98 ɔˑ			15				
100 ɔ̟			9	ɔ	70	69	84
101 ɔ̠	58	53	13				
102 ʔˤ			20				

Table 23.

OU8 ʌ

state	K1	K2	K3	PDV	K1	K2	K3
113 uˤ			26				
114 ɵˤ	62	26		ʊ	62	35	26
115 oˤ		9					
123 ɨ	18			I	18		
128 ə			24				
130 ə̈		39		ə		39	45
131 ʌ̈			21				

Table 24.

OU9 ʊ

state		K1	K2	K3	PDV	K1	K2	K3
132	ʊ			28				
134	ʊ	88	35		ʊ	88	64	46
135	u		29	18				
136	u			12	u			27
137	u			15				

Table 25.

OU10 u

state		K1	K2	K3	PDV	K1	K2	K3
143	u	23	24	10	u	89	85	90
144	u		14	11				
145	u			26				
146	y	15						
147	ɣ>u	16	8	43				
148	ə₂	35	39					

OU2 [I] NL (Table 17 (p.215))

K1, K2, are distinguished from K3 by their use of central vowels as variants of I. (PDV ɐ state ə (K1, K2) and PDV ɜ: state ɛ̝̈ (K2)).

K1, K2 are distinct from each other in that K2 uses ɛ̝̈ and K1 does not, and also by the ratio of usage of states.

OU3 [ɛ] NL (Table 18) , ('head', 'bread', ...)

All clusters use variant state of PDV ɛ predominantly (82%, 83%, 81% respectively).

K3 alone used the centralised diphthong ɛ̱ə. Other distinctions depend on precise articulatory position: (e.g. retraction, lowering): ɛ̱ , ɛ̝̈ , ɛ̝ , ɛ.

OU4 [ae] L (Table 19 (p.216).)

All clusters have state 53 a̝ as the most frequent phonetic state. (E.G. as in, 'path', 'grass' ....)

PDV æ is the main partition for all clusters.

OU5 [ɑ] L ('father', 'barn' ....) (Table 20 (p.216).)

Table 20 shows again that clusters are distinguished, as far as this OU (ɑ) is concerned) by fine phonetic differentiation within

the major PDV, PDV ɑ , and by the proportions with which the states of this PDV are used. Distinguishing parameters are degree of lip-rounding, degree of fronting, and degree of raising from cardinal vowel 5 position.

OU6 [ɒ]^{L} ('off', 'because' ...), OU7 [ɔ:]^{NL}, ('war', 'talk' ...) (similar to OU5). (Tables 21,22 (p.217).)

Table 23 OU7 [ʌ]^{NL} is an interesting case: the phoneme /ʌ/ as in 'mother, cup, love...' is generally not used by localised northern speakers. The lexical set associated with the phonological entity ʌ is split up by this sample, and distributed across 3 PDVs, ʊ , I, ə.

K1 use predominantly an advanced version of cardinal vowel 7 (o) with lip spreading, (62%), and also the centralised high vowel ɨ (18%).

K2 and K3 use a lower proportion of mid-high and high back vowels, and a higher proportion of central vowels (ɵ̈ (K2), and ə, ʌ̈ (K3)) than K1. K2 and K3's central vowels are lower than K1's.

OU9 [ʊ]^{NL} (Table 24) ('pull', 'book'...) K1, K2 use PDV ʊ predominantly, whilst K3 uses PDV u for 27% of realisations.

OU10 [u]^{NL} (Table 25), ('moon', 'beau tiful'...) All states used by each of the 3 clusters are state of PDV u.

Within this, there is a range of phonetic distinctions, covering the high back vowel region (u) (states 143, 144, 145); and centralised or fronted variants (state 140 is retracted variant of secondary cardinal vowel 1, ʏ , used by K1 (15%); state 147 is a high back offglide from advanced variants of secondary cardinal vowel 7(ɤ) and primary cardinal vowel 8 (u) with less lip spreading, state 147 is used by K1 (16%), K2 (8%), K3 (43%).).

K1 and K2 use also a laxer variant of this state (state 148) 35%, 39% respectively.

To summarise, then, the monophthongic vowel subspace (%FON1) tends towards variability at state level more than at PDV level. (Cf. %FON2, see below, pp.221ff.)

%FON1 - Key Diagnostics

The major diagnostics are the variants of OU8, Λ. This OU distinguishes the three clusters on the basis of their choice between PDV's ʊ , I and ə.

OU9 ʊ shows K3 to be distinct in the use of variants of PDV u.

OU2 I distinguishes all three clusters on the basis of their distributions across PDVs I , ʃ , ɜː .

%FON2 Cluster Diagnostics

Tables 26 through 28 show diagnostic lists similar to those given for the first subspace (%FON1), derived from the CLUSTAN diagnostics, for each of the 3 clusters in this space. These are called KA , KB , KC respectively. These clusters are NOT, of course, the same 3 clusters as in the previous subspace, except for KC = K3, the Newcastle subsample. The (working class) Gateshead subsample represented in K1, K2 is variably distributed across KA and KB.

Once again we see that these diagnostics (selected on the basis of F-Ratio < 1 and T-Value $> \emptyset$, arranged in descending T-Value) do not provide an adequate picture of the distributional characteristics of state values for clusters. This is partially due to the fact that state scores are treated as independent and unrelated variables in the classification procedures (i.e. the structuring of the coding frame cannot be reflected in the definition of variables for CLUSTAN).

In addition, however, the definition of 'diagnosticity' on the basis of which these statistics (F-Ratio and T-Value) are based are not the only possibly useful ones.

Once again this information is supplemented by tables showing distributions of cluster-mean scores across PDV's and states of OU's in this subspace. (See Tables 29-41, pp.226-229.)

The most immediately striking feature of these tables concerns the

Table 26.

Linguistic diagnostics , %FON2, KA.

CLU VAR	F RATIO	T VALUE	St.	5-dig code	OU	PDV	state	K-mean %	sample mean %
86	.9616	.9635	240	14601	aʊ	ɔɑ	ɔɑ	34.8	14.6
129	.9065	.7613	283	01688	ɜ	ɔ	ɵ	15.2	7.9
16	.7677	.6743	170	11202	eɪ	ɪə	eə	30.3	16.7
178	.4799	.5259	332	02081	ɪ,	red.(ə)	ə,	77.3	61.0
118	.7750	.3500	272	01641	ɜ	ɛ(ə)	ɛ(ə)	19.6	14,6
149	.1731	.3122	303	01822	ɛə	ɛ	ɛ	81.6	72.9
185	.0288	.3077	339	02123	I_2		i	97.6	91.7
173	.2925	.2502	327	02021	$ə_2ə_1$	red.	ə,	77.6	70.5
34	.4102	.1607	188	01204	əʊ	u:	o˕ə	8.5	6.4
176	.7915	.0630	330	02042	$ə_2ə_1$	non-red.	ɪ	7.1	6.6
54	.2897	.0011	208	01304	ɑɪ	ɑ:	ɑ	51.7	51.6

Table 27.

Linguistic diagnostics , %FON2, KB.

CLU VAR	F RATIO	T VALUE	St.	5-dig code	OU	PDV	state	K-mean %	sample mean %
27	.3902	.7360	181	01181	əʊ	ɔ:	o̞ˑ	73.2	47.9
163	.5865	.6098	317	01961	ə3	non-red.	ɛ̲	63.9	46.8
117	.6967	.5552	271	01623	3	ø	θˑ	55.7	39.7
13	.6106	.5131	167	01123	eɪ	i:	i̲	77.9	61.8
78	.9271	.4746	232	01143	aʊ	ɛʊ	ɛ̲ʊ	56.0	39.9
173	.3336	.3111	327	02021	ə2ə1	red.	ə1	79.3	70.5
54	.5115	.2874	208	01304	aɪ	ɑ:	ɑ̟	56.2	51.6
142	.9306	.2543	296	17801	ɪə	iɛ	iɛ	54.4	47.0
149	.5634	.2200	303	01822	ɛə	ɛ	ɛ̲	79.0	72.9
103	.9361	.1396	257	15201	ɔɪ	ɒɪ	ɒˑɪ	44.8	39.6
185	.1162	.1390	339	02123	I2		i̤	94.4	91.7
180	.5567	.1122	334	02083	I1	red(ə)	I	14.1	13.0
178	.5451	.1037	332	02081	I1	red(ə)	ə1	64.2	61.0
169	.8733	.0760	323	02001	ə4a	non-red.	ɛ̲	40.8	37.9
175	.7584	.0036	329	02041	ə2ə1	non-red.	I	10.0	9.7

Linguistic diagnostics, %FON2. KC.

CLU VAR	F RATIO	T VALUE	St.	5-dig code	OU	PDV	state	K-mean %	sample mean %
182	.6906	2.362	336	02102	I1	non-red.(ɪ)	i:	68.6	15.0
161	.0320	2.274	315	01941	ə3	red.	ə3	92.1	22.0
167	.0120	2.632	321	01981	ə4a	red.	ə3	95.9	21.2
6	.2884	.2011	160	01062	eɪ	ɛ	e̲	4.6	2.3
62	.7328	.1934	216	01362	aɪə	aɪə	a̟ɪə	9.6	5.6
74	.8464	.1574	228	01424	aʊ	aʊ	a̟ʊ	16.7	12.8

range of PDV values taken up by the sample from each OU. In the monophthongal subspace, cluster mean values tend to be highly concentrated within fewer PDVs for each OU.

Realisations of OU i: are distributed mainly across 2 PDVs, OU's I , ʌ , and ʊ have realisations spread over 3 PDVs each, but for the remaining 6 OU's state realisations are heavily concentrated on one PDV only.

In the second subspace (%FON2), however, all diphthongal OU's have relatively high mean frequencies spread over 3 or 4 PDV's. The reduced vowels, whose variants are structured differently, (into reduced and non-reduced realisations), all display dispersions across these 2 categories.

In other words, majority partitions into PDV's of the lexical sets associated with OUs in the first subspace tend more towards uniformity for the whole sample, whereas majority partitions in the second subspace tend to be cluster based, and spread across different PDV's.

In other words, clusters in the second subspace are distinguished more often by variability at PDV level of representation. This means that for this sample of informants, diphthongal and reduced vowels carry variability at a structurally higher level than monophthongs (phoneme-like distinctions, as opposed to fine phonetic distinctions within one PDV). Hence, perhaps, the higher initial fusion level of this subspace: 73.335 (%FON1 = 19.806, %FON3 = 15.676). (D^2).

Tables 29 to 41 (p.235ff) show distributions of cluster mean state scores across PDV's and states.

I shall limit my discussion of these tables to those OU's which display the more striking distributional patterns, and which apparently carry relatively more diagnostic power. (Here I apply a stronger operational definition of 'key diagnostic' than could be applied to %FON1).

The choice of this selection of OU's and their variants is made on the basis of 2 considerations:

i) that variants of a OU (states) satisfy the criteria implied by the CLUSTAN T-Value and F-Ratio statistics; namely cluster mean differs from sample mean (in this case by exceeding it), and cluster variance is lower than sample variance.

ii) that variants (either state scores or PDV scores) of an OU discriminate all three clusters from each other.

Working with these 2 criteria, a short-list of 6 OU's from this subspace is arrived at. These abstract phonological entities display patternings of variant realisations across this sample which demonstrate these items to be salient linguistic variables (in terms of the definition of the space).

The OU's in question are:

eɪ aɪ əʊ 3 ə4a ə3 .

Each of these OU's appears in the lists of diagnostics derived from the CLUSTAN output, and each of them subsumes at least one dimension (either of PDV or state scores) which discriminates all 3 clusters from each other, in terms of cluster mean frequencies.

OU [eɪ] (See Table 29) ('eight', 'railway'...) (As in the previous section, these tables show mean state frequencies by clusters: the left-most column shows the STATE( ) subscript for reference by TRAN table (Appendix X), next right is the phonetic description of the state. Totals for states belonging to the same PDV are shown.

PDV eI was predicted as the majority partition of this lexical set for non-localised speakers (Pellowe, Nixon & McNeany: 1972, p.12f.).

PDVs i: and iə were postulated as majority partitions for localised Tyneside speakers.

The K-mean scores for i: and iə show that KA has realisations heavily concentrated on the localised variants, KB also, but with a few realisations (5%) as the state ɛ̝I, a variant of the NL (non-localised) PDV, (PDV eI).

(Text resumes on p.230.)

Tables 29 - 41.

Cluster mean frequencies of states and PDVs used (high frequencies only) by members of KA, KB, and KC, (%FON2).

Table 29.

OU11, eI.

state	KA	KB	KC	PDV	KA	KB	KC
3 ɛ̝I		5	28	eɪ		5	28
13 i̝	61	78	5	i:	61	78	5
16 e̞ə	30	4	26	iə	30	4	26
20 ɛ̝̈I			16	ɛɪ			16

Table 30.

OU12, əʊ

state	KA	KB	KC	PDV	KA	KB	KC
26 o̞ʊ	1	6	67	əʊ	1	6	67
27 o̞	29	73	4	ɔ:	29	73	4
34 o̞ə	9	1	20	u:	24	5	20
35 o̞	15	4					
36 a̝	17	1		ɑ:	38	4	
39 a̞	21	3					

Table 31.

OU13, aI

state	KA	KB	KC	PDV	KA	KB	KC
48 ä̝I		3	15	aɪ	5	12	42
49 a̝̽I	5	9	27				
53 a̞			9	ɑ:	52	56	43
54 ä̞	52	56	34				
58 ɛ̝i̞	40	25	3	eɪ	40	25	3

Table 32.
OU15, aʊ

state	KA	KB	KC	PDV	KA	KB	KC
73 a̲ʊ	1	3	15	aʊ	4	22	32
74 a̲̲ʊ	3	19	17				
77 ɛ̤ʊ	1	4	35	ɛʊ	25	60	58
78 ɛ̲ʊ	24	56	23				
86 ɔɷ	35	4		ɔɷ	51	6	
87 ʌɷ	16	2					

Table 33.
OU16, ɔɪ

state	KA	KB	KC	PDV	KA	KB	KC
100 ɔ̣ɪ	1	9		ɔɪ	1	20	7
102 ɔ̲ɪ		11	7				
103 ɒˈɪ	44	45	7	ɒɪ	50	48	21
104 oˈɪ	6	3	14				
105 əɪ	10	6		əɪ	25	9	
106 ə̈ɪ	15	3					

Table 34.
OU17, ɜ

state	KA	KB	KC	PDV	KA	KB	KC
111 ə̲ˀ	1	7	65	ɜː	1	7	65
116 θˈ	6	7	13	ɵ	39	63	13
117 θ̲ˈ	33	56					
118 ɛ̲(ə)	20	12	12	ɛ(ə)	20	12	12
129 θ̲	15	5		ɔ	15	5	

Table 35.
OU18, Iə.

state	KA	KB	KC	PDV	KA	KB	KC
132 i̤ə	1	9	29	Iə	1	9	29
138 i̤:	24	24	3	i̤:	24	24	3
142 i̤ɛ	44	54	27				
143 l̤iɛ	3	2	32	iɛ	71	57	59
144 l̤ä	24	1					

Table 36.
OU19, ɛə

state	KA	KB	KC	PDV	KA	KB	KC
146 ɛ̤ə	1	1	38				
147 ɛ̲			17	ɛə	1	1	55
148 e̤	7	13	3				
149 ɛ̲	82	79	26	ɛ	97	97	29
150 ɛ̤	8	5					
152 ɜ̤			12	3:			12

Table 37.
OU21, ə3 final open.

state	KA	KB	KC	PDV	KA	KB	KC
161 ə3	11	11	92				
162 ə4	10	15	5	red.	21	26	97
163 ɛ̲	40	64	1				
166 a	36	7		non-red.	76	71	1

Table 38.

OU22, ∂_{4a} / ___ C# (V#)
[fortis] ___ (r)# C,V...

state	KA	KB	KC	PDV	KA	KB	KC
167 ∂_3	7	12	96	red.	12	40	99
168 ∂_4	5	28	3				
169 ε	48	41	1	non-red.	65	43	1
170 a	17	2					

Table 39
OU23, $\partial_2 \partial_1$ /[non-fortis] ____ #

state	KA	KB	KC	PDV	KA	KB	KC
173 ∂_1	7	79	18	red.	78	79	18
175 I	13	10		non-red.	20	18	
176 I:	7	8					

Table 40.
OU24, I_1 /___ C

state	KA	KB	KC	PDV	KA	KB	KC
178 ∂_1	77	64	5				
179 ∂_4	8	12	2	red.(∂)	93	90	29
180 I	8	14	22				
182 i:	5	8	69	non-red.(I)	5	8	69

Table 41.
OU25, I_2 / ____

state	KA	KB	KC
183 I	2	3	16
185 i	98	94	66

KC shows realisations spread over 4 PDVs, with higher frequencies for PDV's eI and iə , a NL, and a L, variant respectively.

KA and KB, then, fairly consistently use localised variants, with KA using them slightly more frequently (91%:82%). KB, however, favours the localised PDV i:, a monophthongised variant of this OU, while KA also has a fairly high percentage (30%) for PDV iə , the centralising dipthong. KC shows heterogeneous tendencies.

The monophthongal variant, state $\underline{\underline{i}}$ of PDV i: discriminates all 3 clusters, having 61, 78 and 5% as their respective K-means.

OU əʊ Table 30 ('phone', 'go'...) 4 PDV's are represented here, the NL PDV, əʊ , and 2 majority partitions for L varieties, PDV's ɔ: and u: (Pellowe, Nixon and McNeany: 1972, p.13).

PDV ɑ: also appears (localised realisation for lexical items such as 'old', 'know', 'no', 'cold').

In this case KC has realisations of this item concentrated heavily on the NL PDV (67%), (cf. the spread of KC's distribution across the variants of the previous OU's.) For this OU, KA shows a heterogeneous spread of realisations, across 6 states, with state $\overset{o}{\underset{c}{}}$ represented slightly more frequently. However, at PDV level, KA is seen to be slightly more concentrated on the L PDV ɑ:, (38%) while KB has the localised PDV ɔ:, ('so', 'smoke'...) as the overwhelming majority partition (73%).

Summed state frequencies for PDV ɔ: discriminate the 3 clusters from each other: 29, 73 and 4% respectively.

OU aI (Table 31), ('side', 'five'...)

Once again KC displays a heterogeneous distribution across states of the NL PDV, aI, and the L PDV ɑ:. ('I', 'five'...)

KA divides its realisations across 2 L PDVs, ɑ: and eɪ (e.g. 'mine') (52%, 40% respectively), whilst KB has ɑ: as the major partition, (but also has eI as a large minor partition - 25%).

The state $\underset{c}{\dot{\varepsilon}\dot{i}}$, of PDV eI discriminates all 3 clusters, with 40,

25 and 3% respectively.

OU [3]^{NL} Table 34 ('bird', 'earth' ...). PDV's highly represented are 3:, ø, ɛ$^{(ə)}$ and ɔ.

PDV's 3: and ø, (e.g. 'year'), are both postulated as majority partitions for some NL varieties, the latter also being predicted as a majority partition of the lexical set for some L varieties. KC favours PDV 3: state ə$^{ɹ}_{+}$, (65%) a centralised variant of this NL PDV.

KB has a high concentration of realisations under PDV ø, 63%, and KB a slightly lower concentration, 39%. All clusters have a fronted variant, ɛ$^{(ə)}$ with a slight schwa colouration - 20, 12, 12% respectively.

KA and KB use a variant of PDV ɔ, (e.g. 'birth') (θ) 15, and 5% respectively, whilst KC do not use this state at all.

The 3 clusters are discriminated at the PDV level, for PDV ø, 39, 63, 13% respectively. However, as mentioned above, this PDV is a majority partition for some L and some NL varieties, so the interpretation of this distribution pattern is tricky, in relation to L and NL categories. Empirically, however, we can say that, for this sample of informants, frequencies of realisations from the lexical set associated with OU 3 as variants of PDV ø do have discriminatory power for the clusters obtained.

Unstressed Centralised Vowels

"These criteria cover the vowel of unstressed syllables" (Pellowe, Nixon and McNeany,: 1972, p.18), and represent positions in the articulatory space as shown below, (Fig. 53 (taken from Pellowe,Nixon & McNeany, 1972).

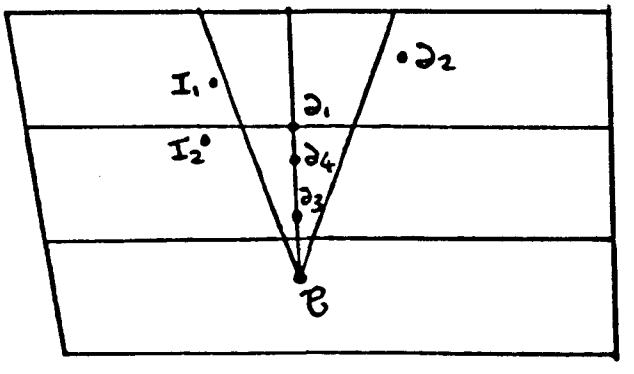


Fig. 53

State realisations for central vowels (unstressed syllables) are divided into reduced and non-reduced, as this distinction has been shown to be significant for individual Tyneside speakers.

"PDV's are defined in terms of the dichotomy reduced/unreduced, since these seem to represent the major distinctions between S's. [speakers] Evidence for this was found in a pilot study (McNeany: unpub.) in which we found that whilst intra-speaker mixing of what we represent as states was possible, intra-speaker mixing of reduced and unreduced forms was rare " (Pellowe, Nixon and McNeany: 1972, p.19).

For example ə3 final open, (Daniel Jones' notation) can be realised with ə3 or ə4 , or as one of the following unreduced vowels:

ɛ̲ ɪ ɐ ɑ ,

e.g. (china).

Such realisations of unstressed syllable vowels as non-reduced vowels feature frequently in Tyneside varieties.

For this particular OU, (Table 37), we find KC favouring the reduced form, state ə3, for 92% of realisations, whereas KA and KB favour the non-reduced states ɛ̲ and ɑ.

At the level of the reduced/ unreduced distinction, we find KC almost exclusively using reduced (NL) forms, and KA and KB using very high frequencies of non-reduced forms.

	KA	KB	KC
red.	21	26	97
non-red.	76	71	1

KA shows a slightly greater tendency towards the localised (non-reduced) forms, with similar proportions for states ɛ̲ and ɑ (40 and 36%), with KB concentrated mainly on state ɛ̲ (64%).

KC is distinguished sharply from the rest of the sample then, on this OU, in terms of vowel reduction. The three clusters are mutually distinct in terms of mean frequencies of usage only at the state level; for state ɛ̲, 40, 64 and 1% respectively.

An analogous situation holds from the OU ə4a , (e.g. 'standard', 'interview' ...).

Table 38 shows that KC once again favours reduced forms (99%), whilst KA and KB use both reduced and non-reduced forms, but use the latter with higher mean frequencies.

Once again KA realises 'central' unstressed vowels with non-reduced forms ɛ and ɑ , whilst KB favours ɛ .

The 3 clusters, however, are not mutually distinct at any state level, but at the level of distinction reduced / non-reduced. KA favours non-reduced forms (12%:65%), whilst KB uses both in approx. equal proportions, (40:43%) and KC uses reduced variants (99%).

On this OU, we have a ranking of clusters from L to NL,

(see text) L ← KA — KB — KC → NL

It appears that members of KC have stability of realisation according to the distinctions reduced/non-reduced for unstressed vowels: these categories, then, are powerful diagnostics for this cluster. (Cf. the heterogeneous distributions displayed by this group for dipthongal OU's).

If it is true that "intra-speaker mixing of reduced and unreduced forms [is] rare", (Pellowe, Nixon and McNeany: 1972, p.19), then the question must be posed, why do we find the phenomenon (e.g. in ə3 and ə4a) of cluster mean frequencies being spread across both categories for KA and KB . Either individuals do use a mixture of reduced and unreduced forms for one OU, or each of these clusters contains 2 sub-groups, one of which habitually uses reduced forms, the other consistently using non-reduced variants.

In this case, we have clusters which are heterogenous with respect to the reduced vowel OUs; in other words, scores for states of the dipthongal OUs in this subspace have overwhelmed the unstressed vowel OUs. This is very probable, given the number of variables (states), yielded by these 2 categories respectively,

(dipthongs = 160, central vowels = 29).

This can be tested by checking the realisations of individual informants with respect to the unstressed vowel OU's only. Since the vowel reduction phenomena may constitute useful diagnostic criteria, there may be grounds for either:

a) assigning these features more weight in the classification;

or b) reducing the number of dimensions associated with the other OU's in this subspace.

b) could be achieved by clustering informants on PDV scores rather than state scores. Interestingly, the distributional picture is reversed for the unstressed vowel I_1. KA, KB use predominantly the central reduced form, (e.g. for houses, stupid), ∂_1, whereas KC use (69%) state i: (non-reduced), most frequently.

There appears, then, to be a reversal operating between (KA + KB) : KC, with respect to reduction of unstressed vowels. For KA and KB, schwa type OU's tend to be fronted to [ɛ] , or lowered and retracted to [ɑ] , whilst KC uses central variants. But for OU I_1 (I unstressed followed by consonant), KA and KB centralise, whilst KC favours the raised, fronted variant [i:]

To summarise, then, the second subspace (%FON2), covering diphthongs and unstressed vowel OU's, appears to contain more variability than subspace 1. This is attested by the fact that clusters are distinguished in some instances by frequency of state realisations, but very often by distributions of cluster mean scores between different PDV's of a given OU. Level of representation obviously crucially affects the resultant classification.

Although an opposition emerges between (KA + KB) and KC in many of the variables discussed, this relation is not reflected for all variables; e.g. OU [əʊ] (Table 32 p.227.),

state ɛʊ, where KA and KC are similar, and differ from KB, with 24, 56 and 23% respectively.

However, at the PDV level (the PDV of which this state is a variant),

KB and KC are opposed to KA.

PDV ɛʊ of OU əʊ has frequencies 25, 59 and 58% for KA, KB, KC respectively.

Detailed analysis of these distributions shows then that:

1. Different variables do not diplay isomorphous distribution patterns across the 3 clusters;
2. different distinctions between clusters exist at different levels of representation;
3. classifications of a given sample are dependent on the level of representation (fineness of analysis) selected as the basis of definition of variables.

Although I have examined cluster mean scores at state, and PDV level in the foregoing, it must be noted that the classification which produced these clusters was performed using variables at state level.

It is hypothesised that a recomputation of clusters using PDV frequencies will produce a different classification of this sub-sample. This exercise is proposed as a future extension to the present research, in order to test the hypothesis that the analytic level of representation selected as a basis for measurement and comparison of the speech output of informants partially determines the nature of the results obtained.

The phonological entities which emerge from this classification as the more significant diagnostic dimensions are the following:

eɪ realised by PDV i:
or iə

aɪ realised by PDV eɪ

əʊ realised by PDV ɔ:
or u:

3 realised by PDV ø

ə4a and ə3 realised with reduced or non-reduced forms.

%FON3 - Consonantal Subspace - Diagnostic Features

As for the 2 other subspaces, selected diagnostics derived from the CLUSTAN output are tabulated (Tables 42,43,44), by clusters (K α, Kβ, K γ,

This information is supplemented in tables 45 through 57 which display distributions of scores across variants of OU's, by clusters.

NB. OU's p, b, t, d, k, g are divided at 'PDV' level, into positional variants, initial, medial and final.[FN] (Tables 42-57 appear below pp.237-242.)

FN. Where 'initial' includes appearance in a prevocalic cluster, 'medial' means intervocalic, or appearing before a syllabic consonant. 'final' means appearance in a post-vocalic cluster. These distinctions apply to the syllable, except for 'medial', "which applies in the absence of free morpheme boundary " (Pellowe, Nixon and McNeany: 1972, p.22ff).

Percentage representations of states are computed, (for each individual), as for other OU's, on the basis of state score as a percentage of the sum of all state scores coded under the OU in question.

Assuming stable distribution of lexical sub-sets defined by d-init, d-medial, d-final, across the lexical set associated with d, then these 'PDV' level distributions reflect accurately relative frequencies of usage of variant states of a sound feature in a given position.

Plosive and Stop Consonants

OU [d] ('doctor', 'elder', 'mad') Table 45 (p.239) displays an interesting opposition between Kα + Kβ , and Kγ.

Kα and Kβ tend to use predominantly a weak voiceless plosive for d in initial position (27, 28%); Kγ use this variant more often in final position 24% (also using [d] with nearly as high frequency, 19%).

For d-init, Kγ use an advanced d in the majority of instances, (21%).

Kα and Kβ also use a weak, voiceless plosive for g in initial position, but the major realisation for each is [g] (Table 47).

(Text resumes on p.243.)

Table 42.
Linguistic diagnostics, %FON3, Kα.

CLU VAR	F RATIO	T VALUE	St.	5-dig code	OU	PDV	state	K-mean %	sample mean %
134	.5057	.9894	477	02681	ɜ		ɜ	84.3	37.3
173	.5114	.4853	516	02861	r		ɻ	90.7	84.8
154	.3746	.4052	497	02781	l	/(v)-{v j}	l	40.0	33.9
143	.0268	.3272	486	02741	n		n	99.6	98.0
130	.3243	.3190	473	02641	z		z	81.3	73.2
79	.7129	.2634	422	02421	k	final	k̚	36.2	33.2
136	.4318	.2570	479	02701	h		h	92.6	89.5
192	.0031	.2516	535	02941	w		w	99.1	95.3
159	.4527	.2442	502	02801	l	/( )c_v	l	13.2	11.8
53	.2208	.2430	396	02322	d	initial	d(d̥)	26.5	24.2
63	.4355	.2262	406	02363	d	final	d	40.0	37.4
44	.5250	.1863	387	02301	t	final	t	33.9	32.0
112	.3833	.1305	455	02561	v		v	98.0	97.5
69	.4295	.1097	412	02383	k	initial	kʰ	27.4	26.2
107	.3283	.1073	450	02541	f		f	97.2	96.6
186	.6254	.0844	529	02881	j	initial	j	94.2	93.7
115	.5956	.0710	458	02581	θ		θ	83.4	82.3
126	.6778	.0631	469	02621	s		s	95.0	94.2
99	.7393	.0484	442	02501	tʃ		tʃ	94.7	94.1
93	.4816	.0300	436	02461	g	medial	g	14.5	14.3
152	.4527	.0112	495	02766	ŋ		ŋᵏ	3.6	3.5

Table 43.

Linguistic diagnostics - %FON3, Kβ.

CLU VAR	F RATIO	T VALUE	St.	5-dig code	OU	PDV	state	K-mean %	sample mean %
199	.0548	.5925	542	27603	-ing (bound)		n	94.9	73.9
164	.5167	.5328	507	02822	l	/V_{#c}	ɳ	19.3	14.8
167	.6320	.4827	510	02825	l	/V_{#c}	e	17.5	13.6
53	.2898	.4208	396	02322	d	init.	d(d$_o$)	28.1	24.2
69	.6780	.4003	412	02383	k	init.	k^h	30.7	26.2
159	.3642	.3898	502	02801	l	/()c_V	l	14.0	11.8
44	.5290	.3768	387	02301	t	final	t	35.9	32.0
143	.0068	.3720	486	02741	n		n	99.8	98.1
19	.3760	.3438	362	02201	b	init.	b	68.8	63.8
112	.5539	.2857	455	02561	v		v	98.6	97.5
130	.3029	.2855	473	02641	z		z	80.5	73.2
192	.0030	.2816	535	02941	w		w	99.6	95.3
79	.4525	.2760	422	02421	k	fin.	k	36.3	33.2
63	.5163	.2747	406	02363	d	fin.	d	40.5	37.4
154	.3848	.2327	497	02781	l	/(v)_{j}	l	37.4	33.9
2	.7559	.1632	345	02142	p	init.	p	49.4	47.3
10	.8239	.1166	353	02164	p	med.	p	10.2	9.2
104	.6292	.1137	447	2521	dʒ		dʒ	98.5	98.0
132	.0479	.1121	475	02661	ʃ		ʃ(ʃ$_+$)	99.6	98.1
126	.8147	.1003	469	02621	s		s	95.5	94.2
136	.8210	.0902	479	02701	h		h	90.6	87.5
55	.7149	.0764	398	02341	d	med.	d	14.6	14.1
140	.5447	.0450	483	02721	m		m	99.8	99.8
33	.3529	.0433	376	02263	t	init.	t^h	14.4	14.2
89	.6292	.0362	432	02441	g	init.	g	62.9	62.3

Table 44.

Linguistic diagnostics - %FON3, Kγ.

CLU VAR	F RATIO	T VALUE	St.	5-dig code	OU	PDV	state	K-mean %	sample mean %
165	.3725	2.466	508	02823	l	/V_{#c}	ɬ	32.4	4.7
47	.4189	2.444	390	02304	t	fin.	tˀ	27.1	4.5
174	.8935	2.019	517	02862	r		ɾ	34.3	10.4
121	.3627	1.509	464	02601	ʃ		ʃ	94.7	62.3
89	.5758	1.245	432	02441	g	init.	g	83.0	62.3
107	.0594	.4997	450	02541	f		f	99.4	96.6
140	.0000	.3199	483	02721	m		m	100.0	99.8
132	.0000	.1808	475	02661	ʃ		ʃ(ʃ)	100.0	98.9

Tables 45-57.

Cluster mean frequencies of states and PDVs used (high frequencies only) by members of Kα, Kβ, and Kγ. (%FON3).

Table 45.
OU30 d.

state	Kα	Kβ	Kγ	
52 d+	2	1	21	init.
53 d0	27	28	4	
63 d	40	41	19	fin.
64 d0	9	9	24	

Table 46.
OU31, k.

state		Kα	Kβ	Kγ	
67	k	6	6	31	initial
69	k^h	27	31	8	
79	k	36	36	13	final
84	k'	1		12	

Table 47.
OU32, g.

state		Kα	Kβ	Kγ	
89	g	55	63	83	initial
90	$\dot{g}$	27	16	1	

Table 48.
OU38, δ

state		Kα	Kβ	Kγ
121	δ	61	53	95
123	$d\delta$	8	13	4
124	$\overset{n}{n}$	6	7	
125	$\tilde{\delta}$	22	24	1

Table 49.
OU39, s.

state		Kα	Kβ	Kγ
126	s	95	96	88
127	s^s			11
128	$\underset{,}{s}$	4	4	1

Table 50.
OU40, z.

state		Kα	Kβ	Kγ
130	z	81	80	24
131	s($\underset{\vee}{s}$)	19	20	76

Table 51.
OU43, h.

state		Kα	Kβ	Kγ
136	h	93	91	76
139	Ø	7	9	24

Table 52.
OU46, $\mathcal{J}$ (free).

state		Kα	Kβ	Kγ
148	$\mathcal{J}$	92	93	92
151	n	3		6
152	ng	4	5	
153	k	1	2	

Table 53.
OU47, 1.

state		Kα	Kβ	Kγ	PDV
154	1	40	37	3	
155	$\underline{1}$			29	$1/(V)\frac{\{V\}}{\{j\}}$
156	$\underline{\underline{1}}$			14	
159	1	13	14		
160	$\underline{1}$			8	$1/(\ )c\_V$
161	$\underline{\underline{1}}$			6	

Table 54.
OU47, 1 (cont.)

state		Kα	Kβ	Kγ	PDV
164	ɧ	15	19		
165	t	1		32	1/V__ c
167	e	14	18		
168	t	3		7	
170	l	5	4		1/( )
171	ɧ	4	4		

Table 55.
OU48, r.

state		Kα	Kβ	Kγ
173	ɹ	91	85	65
174	ɾ	5	9	34

Table 56.
OU50, w.

state		Kα	Kβ	Kγ
192	w	99	100	69
194	ʍ			27

Table 57.
OU51, -ing (bound).

state		Kα	Kβ	Kγ
197	ŋ	32	5	56
199	n	68	95	28

OU [k]

Table 46 (p.240) . ('classic', 'anchor', 'make'). Kα and Kβ favour an aspirated k in initial position; Kγ have almost equal proportions of k and k' (ejective), in final position.

Fricatives and Sibilants

OU [ð] Table 48 ('there', 'that').

Kγ uses ð predominantly; Kα and Kβ have this state as their most frequent realisation (61, 53%), but also nasalise, ð̃ , 22 and 24% of realisations respectively, and use the affricate dð (8 and 13% respectively).

A dentalised n is used by Kα and Kβ (6 and 7% respectively). e.g. in'them'.

OU [z] Table 50 ('zoo', 'ooze' ...)

All 3 clusters have 2 variants, the states z and s(s^{s}_{v}), but in reversed proportions for K γ : (Kα + Kβ).

Kα , Kβ use z most frequently:81, 80%; Kγ use S(S^{s}_{v})- 76%.

Liquids, Semivowels, and Miscellaneous

OU [h] ('hopeful', 'happy' ...)(Table 51)

Kα , Kβ use the NL variant h for 93% and 91% of realisations, and Kγ slightly less often (76%). Kγ delete h more often than the rest of the sample (24%).

OU's ŋ (free) ('sang') and -ŋ (bound) ('eating') Tables 52 & 57

For free morphemes, all clusters use ŋ with very high frequency (92, 93, 92%). In bound morphemes, however, /ŋ/ is realised almost exclusively as n by Kβ (95%), whereas Kα uses ŋ and n in the ratio 32:68, and Kγ in the ratio 56:28.

OU [l] (Table 53 (p.241). ('leaf','pull' ...)

4 PDV's are defined for this OU on the basis of phonetic environment. (Pellowe, Nixon and McNeany :1972, p.28). These are:

A) /(v)- $\left\{\begin{matrix} v \\ j \end{matrix}\right\}$ .... = /l/ followed by vowel, or /j/, and optionally preceded by a vowel. e.g. like, filling

B) /( ) Co - V .. = l following a consonant in a syllable-initial cluster, and followed by a vowel (e.g. 'cloud'.)

These two cover the cases (in English) of phonetic environment typically producing 'clear' l. The other 2 PDV's cover environments typically conditioning l as 'dark' l.

C) /V - $\left\{\begin{matrix} \# \\ c \end{matrix}\right\}$ = l preceded by vowel, followed by consonant or word boundary, (e.g. 'old')

D) /( ) - # = word final syllabic l (e.g. 'bottle').

For AandB, (intervocalic l and syllable initial consonant cluster with l followed by vowel), Kα and Kβ use l , whereas Kγ shows a marked tendency to retraction l or even l.

For C and D, Kγ uses dark l, ɫ , exclusively (post-vocalic l, followed by consonant or word boundary, or syllabic l), whereas Kα and Kβ use ɫ and e for C, and l and ɫ for D. In addition, Kα uses ɬ for D.

Apparently, for some Tynesiders, in this case the Gateshead subsample, clear l can be used in an environment (word-final syllabic l), which is supposed, for English, to produce dark l. The allophonic status of the clear/dark ldistinction may not then exist for Tynesiders, or it may involve different rules of contextual conditioning.

Use of the voiced fricative ɮ , for Kα and Kβ (the Gateshead subsample) is also interesting in the 'dark l' contexts of C and D.

OU [r] Table 55 ('reek', 'Harry', 'tribe'). All 3 clusters use states ɹ (lingual frictionless r) and ɾ (flapped r), for most realisations, with ɹ as the major, and ɾ as the minor variant. However, the proportions vary, thus:

	α	β	γ
ɹ:ɾ	18:1	9:1	2:1 (approx.)

OU [w] Table 56 ('wind', 'when', 'away').

Here again we have a contrast between the Newcastle and Gateshead subsamples. The latter, ($K\alpha$ + $K\beta$), use w predominantly, while the former ($K\gamma$) use ʍ, and also the voiceless fricative ɱ ('RP' realisation for orthographic 'wh', e.g. 'which', (Strang: 1969, p.49).) roughly in the proportion 7:3.

In the consonantal subspace, there are six states only which satisfy the stronger definition of diagnosticity (one sense of 'key' diagnostic) proposed earlier, (p. 225) namely, that, in addition to the CLUSTAN defined statistics, the 3 clusters are each distinguished from the other 2 on the basis of their scores for these variables. (In other words, given the mean score of a cluster for one of those 'key' diagnostics, one could identify the cluster as $K\alpha$, $K\beta$ or $K\gamma$).

These 6 states are the following

OU	PDV	State	α	β	γ
[g]	g init	g	55	63	83
		g̊	27	16	1
[r]		ɻ	91	85	65
		ɾ	5	9	34
[-ing] (bound)		ŋ	32	5	56
		n	68	95	28

Summary

The 3 classifications of the subsample of 52 Tyneside informants have been examined, at the 3-K level, in terms of their cluster-mean state scores, as distributed across OU's, and PDV's of OU's.

It has been demonstrated that each of these subspaces contains dimensions of measurement which discriminate either:

(a) one cluster from the rest of the sample

or (b) all 3 clusters from each other.

Some problems in the definition of the diagnosticity of variables, in

terms of cluster identification, have been discussed.

It has also been demonstrated that the level of representation of variation significantly affects the diagnostic status of variants of an OU. Some OU's have variants which discriminate between clusters only at state level, some at PDV level. It may be concluded, then, that a very different classification of a given sample will be arrived at if the level of representation is changed.

An extension to the present research will involve a classification of the same subsample, under the same clustering conditions, at the level of PDV scores.

Variability of realisation, in terms of which states are used, and with what relative frequency, is evidenced for all OU's, in each subspace.

However, this subsample apparently displays more variation with respect to %FON2, (dipthongal and unstressed vowel OU's), than with respect to the other 2 subspaces. For this sample, then, we can tentatively suggest that the abstract phonological entities subsumed by %FON2 carry relatively more realisational variation than the contents of the 2 other subspaces, and some of these items, therefore, may be more salient sociolinguistic markers.

This is not to suggest that the subspaces %FON1, %FON3 do not also encapsulate items carrying sociolinguistic salience.

Given the scope of the present research, however, it is not possible to investigate the social correlates of all clusters from the 3 subspaces. In view of the remarks above, the following chapter will examine the relationship between the 2nd linguistic subspace, and the social classification, in terms of a comparison of the membership of the social, and linguistic, clusters, and also in terms of the diagnosticity of dimensions of the social space for this set of linguistic clusters (KA, KB, KC).

CHAPTER 7

THE RELATIONSHIPS BETWEEN THE SOCIAL SPACE AND THE LINGUISTIC SPACE : AN ANALYSIS OF THE SOCIAL CHARACTERISTICS OF LINGUISTIC CLUSTERS.

As demonstrated above, (in chapter 6), the relationships between the 3 linguistic subspaces, in terms of the manner in which the sample under investigation is variably dispersed through them, is far from simple.

In this chapter, I shall examine the relationships between the social classification of this sample (see chapter 5) and the linguistic classifications, (with special reference to that based on %FON2, which covers diphthongal vowel, and reduced vowel OUs).

Several approaches would be possible: two are taken here:

i) the constitution of clusters (in terms of members) in the Social Space (SocSp) is compared with the constitution of clusters across the 3 linguistic subspaces;

ii) the social diagnostics of the linguistic clusters are examined (for the %FON2 clusters only).

The second approach is made feasible by the possibility of running 'mixed mode' CLUSTAN jobs. (see above, ch.4) With mixed mode data, 2 types of data are input to CLUSTAN, (i.e. numeric, and binary data).

In the CLUSTAN run described here, each informant's data consists of a file containing:

a) linguistic data (the same data as was used in the %FON2 run described in the chapter on the linguistic classifications, ch.6), which is of numeric type, (i.e. continuous quantitative values),

and b) social data (the same data which was used in the social classification (see ch. 5)), which is in binary form, (i.e. informants are coded for presence or absence of social attributes).

One or other data mode must be masked from the clustering process, but diagnostic statistics can be derived for all variables from both data modes.

In this run, the sample was classified on the linguistic data (numeric

mode) (%FON2 subspace), whilst the social data (binary mode) was masked out from the computation of the similarity matrix, and from the clustering process. Thus the clustering obtained is identical to that described above (ch.6 subspace 2, %FON2) so that the 3 clusters KA, KB and KC are obtained from a classification based on diphthongal vowel OU's and reduced vowel OU's. However, in this CLUSTAN run, diagnostic statistics are produced for all the social variables, in respect of the linguistic clusters. Thus the social diagnostics of linguistic clusters are obtained direct.[FN]

FN. It would also be possible to mask out the linguistic data, cluster on the social data, and obtain linguistic diagnostics for social clusters. I plan to implement this procedure at a later date.

Firstly, however, the distribution of informants across the SocSp is compared with their distribution across the 3 linguistic subspaces, (the first approach, i. - see p.247 above).

i) The Relationships between the Classifications based on the Linguistic Subspaces and the Social Space.

Chapter 4,5 described the definition of dimensions of the social space, and presented the results of a CLUSTAN run based on the social variables constituting this space.

Three social clusters were obtained, which were designated SocKx, SocKy, SocKz. Slide 6 superimposed on slide 1 shows the distribution of the sample (52 informants) across these 3 social clusters. SocKx is shaded red, SocKy, green, and SocKz, brown. They have 27, 15 and 10 members respectively. (The slides are in Appx. T.)

If slide 2 is now placed on top of slides 6 and 1, the distribution of members of the 3 social clusters across the linguistic clusters based on %FON1 is shown.

It is obvious that there is no simple relationship between the social classification and this linguistic classification.

K1 and K2 are both made up of members of each of the 3 social clusters.

K3 is made up exclusively of members of SocKz, but there are 3 members of SocKz not contained by the linguistic cluster K3. 2 are found in K1, 1 in K2.

Table 58 shows a breakdown of the overlap in cluster membership from %FON1 to the social classification.

Members of K1 are split between SocKx and SocKy in the ratio of 2:1 (approx), and K2 across SocKx, SocKy in the ratio of exactly 1:1.

SocKx is split across K1, K2 in the ratio 4:1 (approx);

SocKy is split across K1, K2 in the ratio 2:1.

K3 is contained by SocKz, but the reverse is not true.

The most marked relationship existing, then is between SocKz and K3, but even here there is some overlap.

If slide 3 is placed over slides 1 and 6, it is evident that a similar situation holds for the 2nd linguistic subspace. Table 59 shows a breakdown of the figures.

Correspondences in K-membership between the social, and linguistic spaces.

Table 58.
%FON1:SocSp.

K	SocKx	SocKy	SocKz	Tot.
1	22	10	2	34
2	5	5	1	11
3	-	-	7	7
Tot.	27	15	10	52

Table 59.
%FON2:SocSp.

K	SocKx	SocKy	SocKz	Tot.
A	11	9	1	21
B	16	6	2	24
C	-	-	7	7
Tot.	27	15	10	52

Table 60.
%FON3:SocSp.

K	SocKx	SocKy	SocKz	Tot.
α	11	11	1	23
β	16	4	2	22
γ	-	-	7	7
Tot.	27	15	10	52

Table 61.
'Derived' clusters:SocSp.

K	SocKx	SocKy	SocKz	Tot.
1Aα	4	4	-	8
1Aβ	5	4	1	10
1Bα	4	2	-	6
1Bβ	9	-	1	10
2Aα	-	1	-	1
2Aβ	2	-	-	2
2Bα	3	4	1	8
3Cγ	-	-	7	7
Tot.	27	15	10	52

Slide 4 placed over slides 1 and 6 shows a similar situation yet again, for the consonantal subspace (%FON3). (See Table 60.)

In all 3 cases, SocKz contains the 3rd linguistic cluster (K3 = KC = Kδ), but is not contained by it. The 3rd cluster in each of the linguistic subspaces, as mentioned earlier (chapters 5,6), consists exclusively of the Newcastle subsample. This group is very distinct from the rest of the sample in all 3 linguistic subspaces: however, they are joined by 3 informants from the Gateshead subsample on the social classification.

Despite this partial equivalence relationship between the linguistic clusters K3 = KC = Kδ , and the social cluster SocKz, the mapping from the social space to any one of the linguistic subspaces is complex, and non-discrete in terms of cluster membership. However, it must be remembered that each of the 3 linguistic classifications is a partial classification, based on a subset of phonological variables. It is possible that a superspace containing all the variables from the 3 subspaces would produce an overall (segmental phonological) linguistic classifiation of this sample which would map onto the social classification in some less complex way.

For reasons given above, (chapters 4, 6), it is not possible to cluster the sample on all segmental variables at once (at least at state level).

However, I have isolated groupings ('derived' clusters) which are maintained across all 3 segmental subspaces, and which, therefore, represent in some sense, clusters of informants derived on the basis of the whole segmental space.

If slide 5 is placed over slides 1 and 6, it is apparent that the relationship between these 8 'derived' linguistic clusters, and the social classification is no more straightforward than the relationship between the social classification and any one of the linguistic classifications based on one subspace.

All linguistic clusters (except 3Cδ , and those with 2 or less members), overlap with two, or all three, of the social clusters. The figures for membership of 'derived' linguistic clusters by social cluster are shown in Table 61 (p.250).

There are many possible explanations for the complexity of the results obtained here.

Four issues are discussed here, which must be borne in mind in assessing the significance of these results.

1. The first issue concerns the operational definition of the sociolinguistic measurement space.

(a) Classificatory bias may have been introduced by the presence of redundant (correlated) variables. For instance, in the SocSp, we have seen that clusters tend to be sex differentiated. Only one binary variable (Sex = M = 1, sex = F = 0) is overtly concerned with sex: the effect on the clustering of the value of this one variable is very small. Thus we can conclude that other dimensions of the SocSp are sex differentiated. In other words, the social variable sex is empirically highly correlated with other social attributes. (This is hardly surprising).

Some researchers, (e.g. Brennan: 1972, ch.2) suggest that empirically correlated variables skew the classification, and should be omitted, or replaced by a factor.

However, it can be argued that empirically (as apposed to logically) correlated variables reflect real structural properties of the social orientation of members of the sample. In other words, if responses to other social questions are conditioned (or influenced) by the sex of the informant, then sex is a valid key diagnostic, and represents a fact about the social organisation and beliefs of the community under investigation. This is the kind of information we are seeking, and should not be eliminated.

Logically correlated variables are a different matter. For instance, if one variable represents marital status, and another, number of children, there will be a partial association between the two, which will introduce bias. This is why 'number of children' is not included as a category on the social questionnaire: nuclear family size is included, which tells us how many children live at home when the informant is a parent, but which does not cause bias due to correlation of variables.

Logical correlation of variables is more difficult to anticipate where the linguistic space is concerned. For instance, allophonic variation may produce a negative association between states which function as allophonic variants for individuals. (See comments on vowel reduction above, ch. 6.) Moreover, certain states may show a degree of association due to phonetic contextualisation.

(b) Certain variables may be subject to effective weighting, as a consequence of the design of the coding frame. This may be true in the SocSp, where the magnitude of the effect of a social attribute on the classification depends on the numebr of binary variables which are needed in order to express the range of responses to the question. This is particularly true of ordered multistate variables such as age. (See above, ch. 5.)

(c) Irrelevant categories produce noise in the classification, and tend to blur the relevant distinctions which are inherent in the data. Here we have a paradox: we do not know, a priori, which categories are relevant, and which are not. One of the purposes of classifying is to discover this. However, the classification itself may be skewed if many irrelevant variables are included. This is a non-trivial problem.

(d) The level of analytic representation of variables, as we have

seen, radically affects the classification. (See above, ch. 6). It has been demonstrated that, for single OU's, the sample is grouped differently, depending on whether the PDV level, or the state level, is examined. Although these two levels may produce equally valid classifications, we may want to decide which is more useful for our purposes, in terms of discovering patterns of interaction between components of the linguistic, and social spaces.

The output from Program COLLAPSE (see Appx. X), i.e. a reduction of the state scores to PDV scores, will be used as a basis for further classifications, and these two levels of analytic representation can then be compared.

An incidental consequence of coding speech at the level of the state introduces the problem of sparse matrices. Where the paradigm of variables is so great (542 states), in order to accommodate very fine phonetic distinctions, we find that more cells in the case x state matrix contain zeros than contain positive values. Hence cluster gorupings may be determined as much by shared zero state scores between individuals, as by similar scores on states which are used, and used with similar frequency, by pairs of individuals. This gives a further reason for classifying at PDV level, where fewer zero scores will be found in the matrix.

2. Inadequate data.

The problems outlined above (under 1) concern the definition of the measurement space. This problem (inadequacy of data) concerns the availability of data in relation to the definition of this measurement space. CLUSTAN 1A has no facilities for compensating for the biassing effects of missing data (e.g. NC scores in SocSp).[FN]

FN. CLUSTAN 1C, which is not yet available at NUMAC, can handle missing data. Other improvements to this version of CLUSTAN include a higher maximum number of variables, which would be useful for the T.L.S., as the whole segmental subspace could be processed in one run, at state level, instead of being split into batches not exceeding 200 variables.

As far as the social classification is concerned, the problems under 2, and 1 (above), regarding the definition of the space, and the adequacy of the data collected, can be to some extent bypassed. Seeing we can obtain socially diagnostic information direct for linguistic clusters, any bias (e.g. spurious weighting) resulting from the implementation of the social space per se becomes irrelevant, as this classification (the mixed mode run) depends only on the linguistic space for the generation of cluster groupings.

3. Assessing the validity of the classification, in terms of the effects of the methods used.

As remarked above (chs. 3,4), different clustering methods produce different results on the same set of data. Everritt (1974, pp.8-18) discusses the differential effects of various hierarchical clustering algorithms, in terms of the kinds of fusion patterns, and morphology of clusters, which they typically generate. The question must be asked, how far does any combination of clustering methods produce results which reflect one of a range of valid (i.e. reflecting some kind of reality) structurings of the data. Or, are the clusters obtained merely an artifact of the mathematical procedures applied.

Internal evaluation of the validity of a classification must rely on similar mathematical criteria to those used to generate the clusters.[FN]

FN. e.g. various optimisation techniques improve the configuration of points in the space with respect to some measure of intra-cluster homogeneity, such as within-group Error Sum of Squares.

Cross validation (reclassifying using different clustering methods) works well with artificial data, where discrete and well-defined groupings are built into the data, but this is not the case with the TLS data. (See above, pp. 110ff., where different distance coefficients and clustering algorithms were tested on one set of linguistic data). We must keep

open the possibility that the chaining effect produced by single linkage, for example, represents an equally (or more) valid structuring of the sample.

4. The fourth explanation for the complexity of the results obtained is that the cluster patterns across the 3 linguistic spaces, and the social space, are related in no less complex a way than are the social and linguistic behaviour of the sample. The expectation of discovering simple sociolinguistic phenomena, in terms of the relationships between social and linguistic features, and groups, is less tenable than the position adopted here: viz. that these relationships, though possibly systematic and regular, are likely to be extremely complex. The results presented so far indicate that this is, in fact, the case.

Having discussed the correspondences (and non-correspondences) between the social and linguistic spaces in terms of the distribution of members of the sample across clusters, I now turn to an analysis of the social diagnostics derived for the linguistic clusters using the CLUSTAN mixed mode facility.

ii) The Social Characteristics of Linguistic Clusters.

For the reasons given in the foregoing, (ch.6 summary), the subspace %FON2 was selected for analysis in terms of all social variables.

The procedure adopted for analysing the social characteristics of the three linguistic clusters in this segmental subspace is the same as that found in ch. 5 That is, cluster frequencies for age, occupation, and education index categories are tabulated, and presented in histogram form, and then the list of CLUSTAN diagnostics is analysed. The difference is, of course, that the clusters analysed in ch.5 are social clusters: the analysis in that chapter was an investigation of the properties of the social space, and the characteristics of the social clusters generated by it. Here I am taking clusters generated from linguistic data, and applying similar methods of analysis in terms of social attributes. The clusters themselves,KA, KB, and KC, are those described in ch.6 - the clusters generated by the 2nd linguistic subspace, %FON2; covering the following OU's:

eɪ əʊ ɑɪ ɑɪə aʊ ɔɪ ɜ ɪə ɛə ʊə.

and (in unstressed syllables) four types of schwa (in different phonetic contexts), and two types of the vowel I.

These clusters are depicted on the transparencies: (Slide 1 plus slide 3 Appx. T.)

Cluster sizes are:

KA - 19 informants

KB - 26 "

KC - 7 "

SEX

KA and KB are strongly polarised on sex: 16/19 in KA are male, 20/26 in KB are female. KC has 4 male, 3 female.

AGE

Table 62 shows raw, and percentage frequencies for age categories, by cluster. The sample frequencies are also shown, and the differences

between sample and cluster percentages for each age group. Fig. 54 shows percent representation of age groups for each cluster, and Fig. 55 shows % differences between cluster frequencies and sample frequencies for each age group.

Age groups are defined as follows:

1 - 17-21 yrs.

2 - 21-30 yrs.

3 - 31-40 yrs.

4 - 41-50 yrs.

5 - 51-60 yrs.

6 - 61-70 yrs.

7 - 71-80 yrs.

8 - 80+ yrs.

Figs. 54 & 55 and Table 62 show that:

63% of KA are between 21 and 40 yrs. Age groups 2, 3 and 6, (21-40, 61-70), are more highly represented in this K than in the whole sample.

KB. spans the whole age range (with the exception of 71-80), with frequencies for ages 17-30, 41-60, exceeding the sample expectation.

KC is too small (7 members) for statistical generalisations to be made about it.

EDUCATION INDEX

This is the index applied in ch.5 to the social clusters: it represents a reduction of the information on educational background stored in the social files.

The categories mean:

a) left school at legal minimum age, no f.e.,

b) extended secondary education, no f.e.,

c) education continued into working life (night school, day release),

d) full time technical college/nursing/secretarial training,

Tables 62, 63, 64.
The social characteristics of linguistic clusters KA, KB, KC.

Table 62.
The distribution of age groups across the clusters.

Age	sample		KA		KB		KC		% differences (K-sample)		
gp.	f	%	f	%	f	%	f	%	KA	KB	KC
1	5	10	1	5	3	12	1	14	-5	+2	+4
2	13	25	6	32	6	23	1	14	+7	+2	-11
3	10	19	6	32	3	12	1	14	+13	-7	-5
4	12	23	2	11	7	27	3	43	-12	+4	+20
5	6	12	1	5	5	19	-		-7	+7	-12
6	4	8	3	16	1	4	-		+8	-4	-8
7	1	2	-		-		-		-2	-2	-2
8	1	2	-		1	4	1	14	-2	+2	+12
NC	-		-		-		-				
Tot.	52		19		26		7				

Age group:

1 = 17-20

2 = 21-30

3 = 31-40

4 = 41-50

5 = 51-60

6 = 61-70

7 = 71-80

8 = 80+

Table 63.
The distribution of education index categories across clusters,

(sample frequencies and %differences (cluster - sample) shown.)

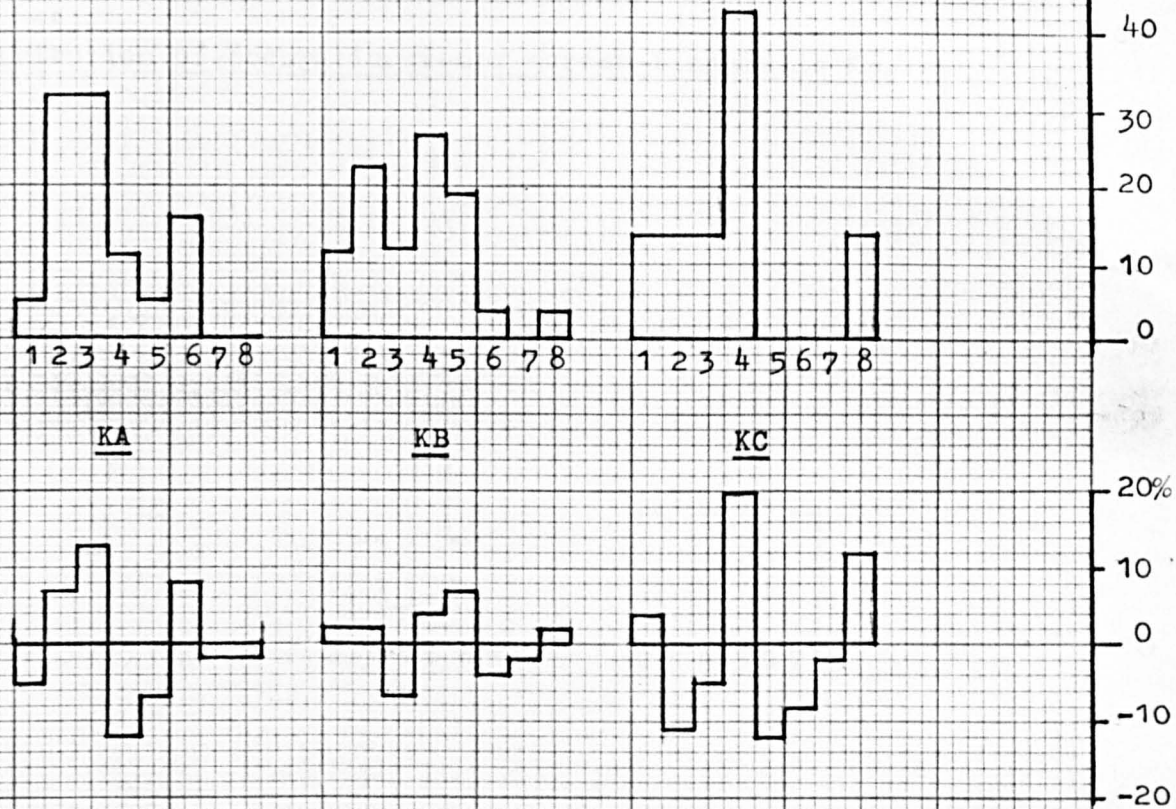
educ. index	sample		KA		KB		KC		%diffs.		
	f	%	f	%	f	%	f	%	KA	KB	KC
a	32	63	13	68	16	62	3	50	+5	-1	-13
b	-		-		-		-				
c	13	25	6	32	7	27	-		+7	+2	-25
d	3	6	-		1	4	2	23	-6	-2	+27
e	3	6	-		2	8	1	17	-6	+2	+11
NC	1		-		-		1				
Tot.	52		19		26		7				

a = left school 1ma. no fe.
b = extended secondary educ., no fe.
c = education continued into working life (not full time.)
d = full time technical/secretarial college / nursing.
e = College of education / university/polytechnic

Table 64.
The distribution of occupation groups across clusters, and sample.

occ. gp.	sample		KA		KB		KC		%diffs.		
	f	%	f	%	f	%	f	%	KA	KB	KC
2	1	2	-		-		1	17	-2	-2	+15
3	3	6	-		3	13	-		-6	+7	-6
4	4	8	2	11	1	4	1	17	+3	-4	+9
5	19	40	8	42	8	35	3	50	+2	-5	+10
6	6	13	3	16	3	13	-		+3	0	-13
7	15	31	6	32	8	35	1	17	+1	+4	-14
NC	4		-		3		1				
Tot	52		19		26		7				

Fig. 54.
Distribution of age groups across linguistic clusters.



Percentage differences between cluster and sample means.
Fig. 55.

Fig. 56.
Distribution of education index categories across linguistic Ks.

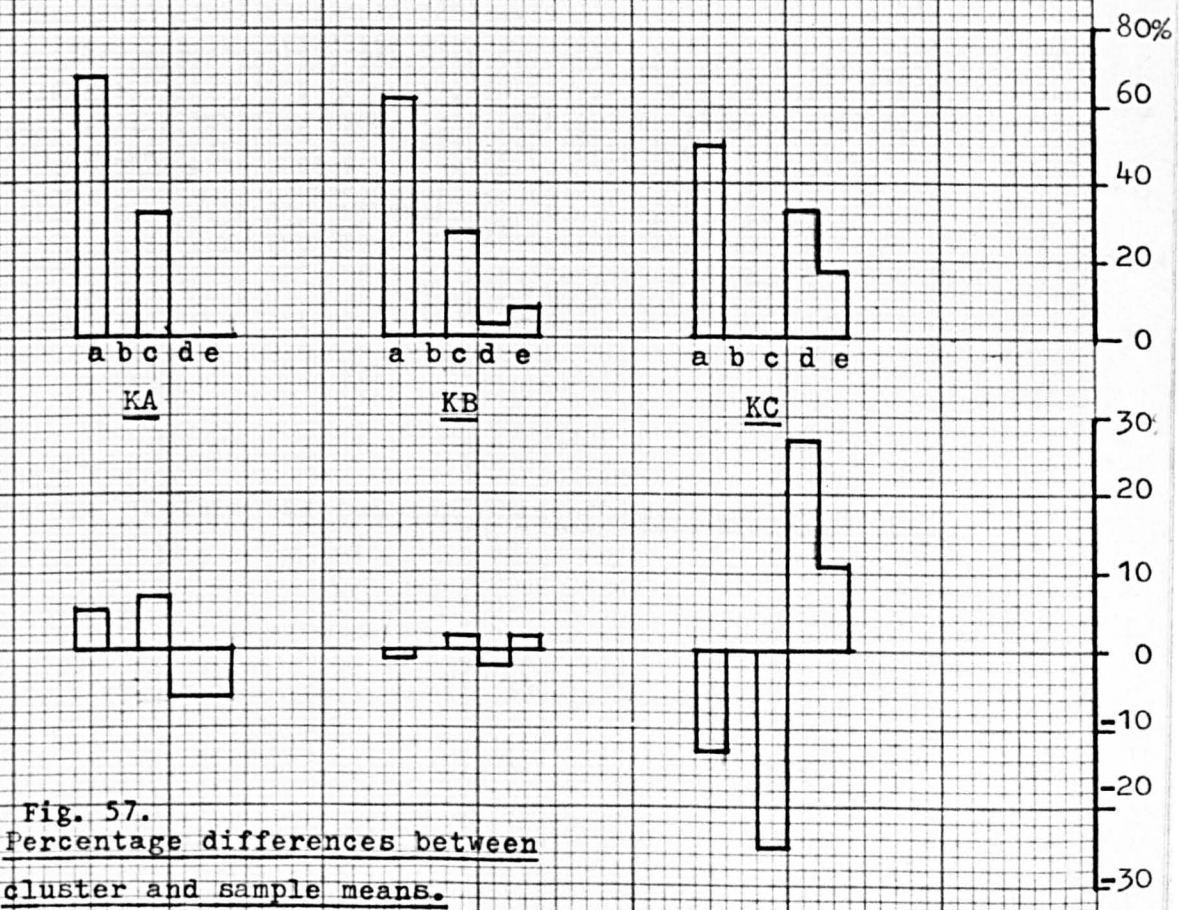
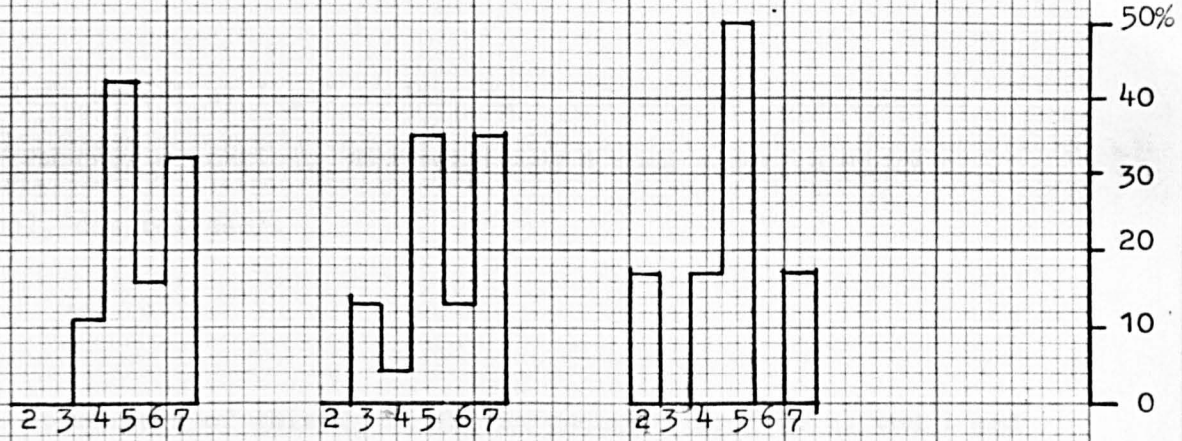


Fig. 57.
Percentage differences between cluster and sample means.

Fig. 58.
Distribution of occupation groups across linguistic Ks.



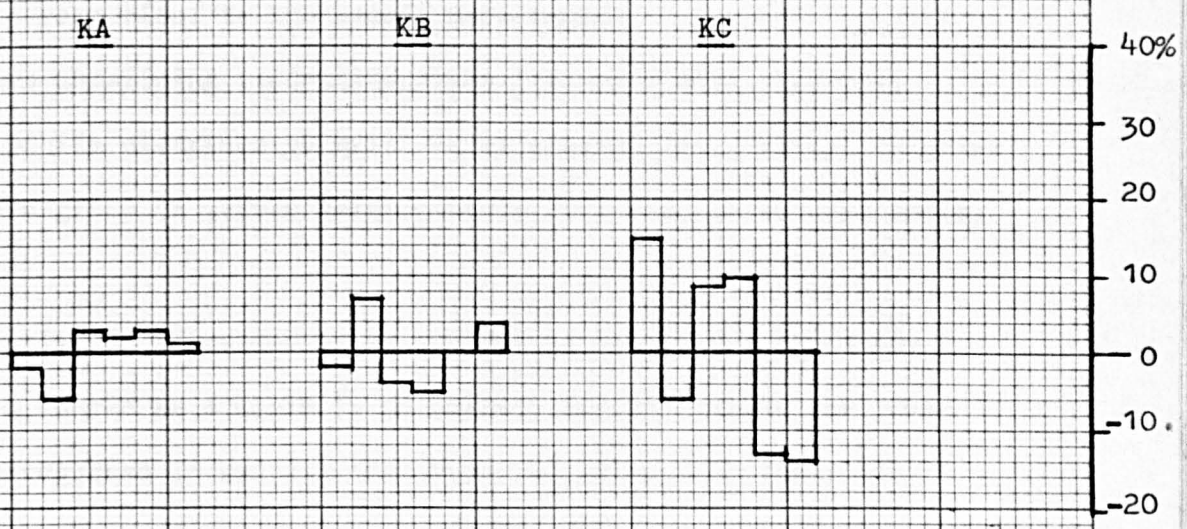


Fig. 59.
Percentage differences between cluster and sample means.

e) College of Education, University, Polytechnic (full time).

Table 63 , and Figs.56, 57 show the raw and percent frequencies for each education index category by cluster.

KA

13/19 left school at legal minimum age, and have not proceeded to any kind of further education. The remaining 6 come under category c (night school, day release).

KB

Of the 26 members of this K, 16 are in category a, 7 in c, and 1 and 2 in d and e respectively. This cluster includes 3 of the 6 cases in the sample who have undergone full time teriary education. However, the percentages in the categories a and c (b is a null category for this sample), are very close to the sample expectation.

KC shows polarising tendencies: given its small size, however, discussion of its distribution will not be entered into, except to point out that this group is rather heterogeneous with respect to most social attributes.

Apparently, KA is marked by an absence of the higher educated: however, the sample as a whole is so heavily skewed towards the lower education categories (a and c), (See Figs. 43, 44 p.172), that the rarity of instances of the higher categories makes it impossible to make any positive assertions about these distributions. (If the situation had arisen, for example, where one cluster was exclusively composed of category a, or categories d and e cases, then this index would have been more useful). However, it can be stated that KA differs from KB with respect to this index, by KA exceeding example expectation on categories a, c and having no instances of any other categories, while KB exceeds sample expectation on categories c and e.

OCCUPATION

A parallel trend is evident when education index is compared to occupation groups. Cf. Table 64 (p.260).

Members of KA all belong to occupation categories 4 down to 7, with sample expectation exceeded for all these categories, but especially for groups 5 and 7.

Members of KB span the categories 3 to 7. This cluster contains the only 3 cases in the sample who belong to occupation group 3. There is also a higher proportion of occupation group 7 in this K, than in the whole sample. These linguistic groupings, then, do not covary simply with level of education, or occupational status. There are large areas of overlap (KA and KB share education categories a and c and occupation groups 4-7), and even within single clusters we find polar tendencies. (KB - exceeds sample expectation for frequencies for occupation groups 3 and 7).

KC, (= K3 = Kδ), the most distinct grouping in each linguistic subspace, shows great heterogeneity with respect to each of the social measures discussed so far. (Sex, age, education, occupation). It is important to bear in mind that these are precisely the kind of measures on which many sociolinguistic investigations focus. (See above, ch.1.)

The only social correlate of this linguistic classification which has emerged as a discriminating factor so far is sex: this feature distinguishes KA and KB, (but not KC). And even here the distinction is one of relative frequency, not an absolute and definitive characteristic. (Sex was also a major discriminating factor in the social classification (ch. 5).

This can be explained in terms of many values of variables in the SocSp covarying with sex. Apparently, some of the dimensions of this linguistic space also covary with sex: a very different, and socio-linguistically more interesting phenomenon .

Having examined the distribution of some classic sociological indices across the linguistic clusters, I now turn to the diagnostic statistics generated by CLUSTAN, for all variables in the social classification,

applied to the linguistic clusters.

I take first the most rigorous definition of a diagnostic; viz. a value of a variable which is uniform for, and exclusive to, a given cluster. That value of that variable will be, then, an absolute predictor of group membership for the sample under study. (The possibility must, however, be borne in mind that the inclusion of new data, i.e. new cases, may alter the diagnostic status of any variable, e.g. to take the simplest instance, the addition of one new case may not alter the cluster groupings (except that the new case joins an existing cluster) but this new case may not score positively on a variable which previously was positive for all members of that cluster. Thus that variable is no longer a 100% certain predictor of membership of that cluster.)

Table 65 shows all variables which are either a) uniform for a given cluster (i.e. all members of a given cluster share that attribute: shown in CLUSTAN output as, 'Percentage Occurrence for Binary Variable' = 100,) or b) exclusive to a given cluster. (Possession of that attribute is unique to 1 cluster).

Table 65 shows that one binary variable satisfies condition a), and 21 binary variables satisfy condition b), but no binary variables satisfy both a) and b).

(Value = 1 for a binary variable means that an individual has a certain value for a certain social feature).

Thus, a value of 1 for Bin.Var. 5 means that the response to the question concerning the informant's regionality was coded under the category 'U.K. Northern'.

This is the only binary variable which satisfies condition a). However, this variable is uniform for both KA and KB (100% of informants in each K were coded as 'U.K. Northern' regionality). Thus these 2 clusters are not discriminated by this variable.

Even if we take the 2K level, (where KA and KB fuse, and are in

Table 65.

Social diagnostics for %FON2, KA, KB, KC, which are uniform for, or exclusive to, a given cluster.

Variables which are uniform for a cluster are marked * , the rest are exclusive to a given cluster.

VAR	KA		KB		KC		sample	definition	
	f	%	f	%	f	%	f		
5*	19	100	26	100	6	86	51	reg.=UK Northern	
12	1	5					1	parent's reg.=UK Lowland	KA
102			3	12			3	occ.=3	KB
3			1	4			4	citiness=market town	
35			1	4			1	college of education	
31			1	4			1	lma+5 years	
10			1	4			1	parent's reg.=UK Midland	
11			1	4			1	" " =London SE	
8			1	4			1	" " =E and W Ridings	
9			2	8			2	" " =N Midland	
30			2	8			2	lma+3 years	
107			1	4			1	informant's 1st occup.=3	
152					1	14	1	occup.=2	K
153					1	14	1	1st occup.=2	
150					1	14	1	hobbies=1 (golf)	
151					1	14	1	" =10 (bridge).	
147					1	14	1	reg.=UK Midland	
145					1	14	1	citiness=Leeds	
144					2	18	2	" =Merseyside	
149					1	14	1	parent's reg.=UK NW	
148					2	28	2	reg.=UK NW	
146					2	28	2	" =UK E and W Ridings	

opposition to KC) we find that 86% of KC (6/7) are also coded positive in this variable.

In other words, the only social variable which satisfies condition a), does so because it is very nearly uniform for the whole sample.

Concerning criterion b);

1 binary variable is exclusive to KA

10 " variables are exclusive to KB

10 " " " " " KC

However, when we examine the frequency of occurrence of these binary variables we find that each of them is positive for only a small minority of members of the given cluster; thus cannot be claimed to typify the goup.

For instance, the only Bin.Var. exclusive to KA is Var. 102: parents' regionality = U.K. Lowland, and this accounts for the only positive instance of this category in the whole sample. By CLUSTAN's Binary Percentage Frequency Ratio (BPFR) (% in K/% in sample), this feature has the maximum level of diagnosticity for this cluster. (This value will be equal to the ratio:total number of cases in sample/ total number of cases in K, when a bin.var. is exclusive to a K).

The BPFR has a maximum value for all variables which are exclusive to a cluster, regardless of the actual frequency, which may range from 1 to n, where n is the number of cases in the cluster.

Thus, the BPFR statistic does not provide a satisfactory definition of cluster diagnostics, taken alone. It must be considered together with the within-cluster percentage occurrence of a given bin.var.

However, though actual (and percentage) frequencies for all these (single) binary variables which are exclusive to any one cluster are very low, there are groups of bin.vars. which are related, and which may together be informative about the extra-linguistic characteristics of these linguistic clusters.

KB has 5 instances of informants with a parent of regionality other

than Tyneside (3 Midland, 1 S.E. England, 1 E and W. Ridings). KC has one instance of parent's regionality = UK N.W.

There are also other cases, e.g. where

informant's regionality = UK Midland (1)

" = UK N.W. (2)

" = UK E and W. Ridings (2)

And 3 cases of citiness being other than Tyneside:

Citiness = Leeds (1)

" = Merseyside (2)

Collating this information, it is found that KA consists of informants who are all indigenous Tynesiders (i.e. have never lived anywhere other than on Tyneside (for 5 years or longer); and whose regionality is U.K. Northern (2 year criterion) - see Q's 1 and 2 on social space coding sheets Appx. B).

Moreover all members of KA, with the exception of one case, have both parents coded as Regionality = U.K. Northern.

Thus KA is made up exclusively of informants having lived all their lives on Tyneside, and whose parents were also northerners.

All members of KB are also indigenous Tynesiders: however, there are 4 instances of parents regionality being other than U.K. Northern. (The frequencies number 5: this is due to the fact that one informant's parents were both coded under regionality other than U.K. Northern: multiple coding is permitted here to account for both parents, also to allow for the possibility of one parent having lived in more than one region).

So KB contains exclusively Tyneside informants, but some have non-Tyneside (and non-Northern) parents. This fact may well have influenced the language of the informants in question.

KC contains 3 (out of 7) cases who(amongst them) account for the instances of Var's. 144-149 shown on Table 65 (p.266).

These variables include 1 instance of parents' regionality in U.K. N.W.

In addition, these 3 cases have non-Tyneside backgrounds: one has lived on Merseyside, one on Merseyside and in London, and one in Teesside and Leeds, though all were resident on Tyneside when the survey was conducted.

So, KA consists of indigenous Tynesiders, born of indigenous northerners;

KB consists of indigenous Tynesiders, some of whom had non-northern parents;

KC contains some indigenous Tynesiders, and some in-migrants from other parts of U.K.

The other cluster-exclusive variables which are of interest have already been discussed (p.264, occupation). KB contains the only informants in occupation group 3 (3 cases), and KC contains the only informant in occupation group 2. (The highest group represented in this sample).

This informant XSHAW, is extraordinary (in relation to this sample) in that not only does he belong to occupation group 2 (managerial and executive) but his <u>first</u> occupation was also of this category.

(This group also contains one informant XFULT , whose hobbies include playing golf and bridge).

As indicated in the foregoing, this group is socially heterogeneous and spans a wide range of occupation, education and age groups, and is not sex-differentiated as are the other 2 clusters. We must consider the possibility that one of the strongest influences on language behaviour is that of geography: this cluster (K3 = KC = K♂) is differentiated from the remainder of the sample in all 3 linguistic subspaces, but not in the social space. The one non-linguistic factor which unites this group is the fact that all members live <u>north</u> of the Tyne, in Newcastle, whilst the rest of the sample are resident south of the Tyne, in Gateshead.

Tables 66 through 68 list the positive social diagnostics for the 3 linguistic clusters computed by CLUSTAN (Program RESULT), in descending order of BPFR levels.

The first column shows the index number of the binary variable, the 2nd and 3rd show the raw frequency for presence of each variable in the K, and in the whole sample respectively. The 4th column shows the BPFR value for the binary variable in question. The 5th column gives the definition of the variable.

KA - Diagnosticity of Binary (Social) Variables

Table 66 shows the diagnostics for KA. (Pp.271-2.)

To summarise information given above, this is a predominantly male cluster, (16/19), with 63% of members between 21-40 years of age. No members underwent Tertiary or further education, and occupation groups 2 and 3 are not represented at all. All informants are indigenous Tynesiders, whose parents were northerners (except 1 case, having a U.K. Lowland parent).

The more significant diagnostics are discussed (i.e. those with BPFR values of more than 1, and relatively high within-K percentage frequency).

VAR.124. The highest BPFR level, and a high within K percentage (53%):

'Drinking as a hobby.' 12 informants in the sample responded positively, 10 of those 12 are found in KA. This is evidently related to the sex distribution, but not completely accounted for by this.

(62.5% of the males in this K responded positively, whilst only 20% of the males in the rest of the sample did).

Regarding education: 11 members of KA have a negative attitude to their own education; 2 have a negative attitude to their children's education; 3 regard the education of their children as the acquisition of basic skills (RRR), 4 consider there should be a distinction between the type of education received by their male and female children, 5 think not. 3 have furthered their training through day release, 4 by attending night school. None underwent tertiary (full-time) education (VAR.32).

VAR 26 Sex = male: 16/19 (in K) are male, out of 26 males in the sample.

Table 66.

Cluster diagnostics (social) for linguistic cluster KA.

VAR	no. in K	no. in sample	BPFR	definition
124	10	12	2.29	drinking as hobby
93	4	5	2.19	social integr. with neighbours=antagonistic
43	3	4	2.06	attit. to children's ed.=RRR
84	5	7	1.96	mac. env. pref.=smaller town
42	2	3	1.83	attit. to children's ed.=negative
130	2	3	1.83	hobbies=5
89	2	3	"	mac. env. pref.=abroad
76	2	3	"	taste aspir='bad'
26	16	26	1.69	sex=M
36	3	5	1.65	fe by day release
109	11	19	1.59	1st occup.=5
49	5	9	1.53	parental control=indirect verbal
115	14	26	1.48	job satisfaction=medium
121	5	10	1.37	TV (intense, non-selective)
24	3	6	"	age=60+
133	2	4	"	hobbies=12
18	1	2	"	2+ moves before marriage
97	1	2	"	father's occup.=3
120	1	2	"	non-own (TV,radio)
105	3	6	"	occup.=6
129	1	2	"	hobbies=4
15	1	2	"	5+ moves before marriage
103	2	4	"	occup.=4
82	1	2	"	financial commit. =high
136	1	2	"	hobbies=2
112	8	16	"	job pref.=I
131	1	2	"	hobbies=7

Table 66 (cont.).

VAR	No. in K	no. in sample	BPFR	definition
46	4	8	1.37	distinction in education boys/girls
95	4	8	"	soc. integr. with neighbours=cordial
96	1	2	"	" " " " =intimate
4	1	2	"	citiness=other
142	1	2	"	voting preference=refusal
92	2	8	"	soc. integr. with neighbours=non-existent/known
69	9	18	"	mic. env. pref.=dissatisfied
72	6	12	"	" " " (housing)=dissatisfied
38	11	23		Attit. to educ.(self)=negative
99	8	17	1.29	father's occup.=5
75	7	15	1.28	taste aspir.='good'
141	11	24	1.26	vote=Labour
137	4	9	1.22	approve connection occup./voting behaviour
37	4	9	"	fe=night school
65	10	23	1.19	distance spouse's reg.=same local authority
51	3	7	1.18	parental control=indirect physical
91	3	7	"	soc. integr. with neighbours=non-existent/unknown
101	6	14	"	father's occup.=7
52	16	38	1.16	married
104	8	19	"	occup.=5
64	5	12	1.15	sex distrib. of children=M bias
32	13	32	1.12	no tertiary educ.
48	4	10	1.10	parental control=direct verbal
60	2	5	"	nuclear family size=5+

Environmental Preference

VAR 84, 89. 5 (out of 7 in total in the sample) would prefer to live in a smaller town, 2 would like to move abroad. VARs 69, 72. 6 informants are dissatisfied with their micro-environment in terms of housing, and 9 in terms of sentiment. (cf. responses to Social Integration with neighbours: VARS 91 - 96).

VAR 109. For 11 of the 19 members of this cluster, their first occupation after leaving school was in category 5 (skilled manual/routine non-manual). (This group is one of the 2 modal values for present occupation, the other being group 7).

VAR 105. Occupation group 6 (for informant's present occupation) has a frequency in KA higher than expectation based on the whole sample.

VAR 103, 104. Similarly, occupation groups 4 and 5 are more highly represented in this K, than in the sample.

VAR 112. This is the 'I' coding of the composite index for job preference. (Q31 on the social coding sheets).

'I' represents a combination of preferences involving:

a) "prospects", as opposed to "immediate gain";

b) "thinking (new elements)" as opposed to "learnt";

and c) "self deciding" as opposed to "supervised".

(See ch. 5.)

8 members of KA are coded 'I' under job preference. 3 are coded 'R' (the reverse of the preferences above), and 8 are coded N.C.

The frequency of 'I' codings is higher for this K than for the total sample.

Voting Behaviour

11/19 in KA vote Labour, a higher proportion than for the whole sample (24/52). (1 votes CONS).

4 cases approve of voting preference being connected with occupation:

i.e. political allegiance should be linked to occupational (class?) status.

Leisure pastimes

Hobbies (classified by types 4, 5, 7, 12, 22 - see social coding sheets) emerge as diagnostic of this cluster according to BPFR, but actual in-K frequencies are low for all of these (1 or 2).

T.V. viewing habits = intense, non-selective (VAR 121), has a frequency of 5 in this cluster, a higher proportion than in the total sample.

VAR 124. 'drinking as a hobby' emerges as the only variant concerning leisure activities which is diagnostic for this K both in terms of raw frequency (10/19), and BPFR.

To summarise the characteristics of KA as revealed by the foregoing analysis;

KA is very predominantly MALE, has a relatively high concentration of 21-40 year old members;

18/19 left school at legal minimum age; (the other 1 stayed on 1 year);

and the 6/19 from this K who proceeded to further training did so during their working life, through day release (vocational training), or night school.

This K then, is characterised by a lower level of academic education, with attitudes to education correspondingly negative (though less so with respect to aspirations for their children's educational lives). KA as a cluster shows a tendency towards occupation groups 5 and 7; 17/19 are in groups 5, 6 and 7.

Viz: skilled manual and routine non-manual,

semi-skilled manual,

unskilled manual.

However, 8/19 have job preferences which contradict a working class stereotype. (Coded 'I' rather than 'R').

A high proportion vote Labour.

Drinking, and non-discriminating and intense T.V. viewing emerge as the most frequent leisure activities for this K.

(N.B. the low frequency with which other hobbies are respresented may be a consequence of the way in which the coding categories of Q36 (Soc.sp.) are set up).

KB - Diagnosticity of Binary (Social) Variables

KB, predominantly female, with a mixed age distribution, has a high concentration of informants having left school at l.m.a., and who had no further education (16/26). Of the other 10, 7 attended courses of further training/education through day release/night school, (so far, KB is quite similar to KA); the other 3 underwent full-time tertiary education, (here there is a divergence from KA's profile). Occupation group 3 is exclusive to this K, though the highest group (2) is represented elsewhere (KC has (only) 1 instance of occupation group 2).

Occupational trends are mixed for this K - groups 3 and 7 are represented more highly than expected on overall sample distribution.

Table 67 shows the diagnosticity levels of dimensions of the social space, for this linguistic cluster (KB)(pp.276-7).

Occupation has already been discussed: these diagnostics show additional information concerning, e.g. comparisons between informant's present occupation and informant's 1st occupation; and also with informant's father's occupation.Parallel trends exist for these 3 criteria, tending in each case towards low occupational status (6 and 7), and relatively high (3 and 4); for informant's first occupation, and informant's father's occupation, as well as informant's present occupation.

In addition, it is clear that some (but not all) informants in this K are upwardly mobile with respect to occupational status: 13 started their working life in group 7, only 8 were in this group at the time the

Table 67.

Social diagnostic for linguistic cluster KB.

VAR	no. in K	no. in sample	BPFR	definition
102	3	3	2.00	occup.=3
3	1	1	"	citiness=market town
35	1	1	"	fe=college of education
128	1	1	"	leisure satisfaction=disgruntled
134	2	2	"	hobbies=15
117	1	1	"	viewing habits=predominantly radio
31	1	1	"	lma+5 yrs.
10	1	1	"	parent's reg.=UK Midland
11	1	1	"	" " =London,SE
8	1	1	"	" " =E and W Ridings
9	2	2	"	" " =UK N Midland
30	2	2	"	lma+3 yrs.
107	1	1	"	1st occup.=3
125	8	9	1.78	housework as hobby
140	5	6	1.67	vote=Conservative
98	3	4	1.50	father's occup.=4
39	3	4	"	attit. to educ.=RRR
29	3	4	"	lma+ 2 yrs.
44	14	19	1.48	attit. to ed. of children=liberal
94	14	19	"	soc. integr. with neighbours=minimal,pleasant
50	15	22	1.37	parental control=direct, physical
138	6	9	1.34	accept connection occup./voting behaviour
100	6	9	"	father's occup.=6
40	10	15	"	attit. to ed. =liberal
111	13	20	1.30	1st occup.=7
47	9	14	1.29	no distinction ed. boys/girls
80	23	36	1.28	financial commit.=4-5
123	14	22	"	TV non-intense, non-selective

Table 67 (cont.).

VAR	no. in K	no. in sample	BPFR	definition
54	5	8	1.25	widowed
66	5	8	"	spouse's reg.< 50m > local authority
63	11	18	1.23	sex distrib. of children=F bias
77	15	25	1.20	taste aspir.='indifferent'
48	6	10	"	parental control=direct verbal
110	3	5	"	1st occup.=6
90	15	25	"	+ positive Tyneside consciousness
143	3	5	"	vote=floater
85	17	29	1.18	mac. env. pref.= town of same size
23	7	12	1.17	age=50+
22	14	24	"	age=40+
45	22	38	1.16	attit. to children's ed.=job oriented
135	22	38	"	hobbies=16
51	4	7	1.15	parental control=indirect physical
101	8	14	"	father's occup.=7
88	20	35	"	mac.env. pref.=nowhere else
79	23	41	"	financial commit. =low + (OM)
78	24	43	1.12	" " =low (OM)
37	5	9	"	fe=night school
139	5	9	"	disapprove connection occup./voting behaviour
137	5	9	"	approve " " " "
57	22	40	1.10	nuclear family size=2+
74	11	20	"	mic. env. pref. (housing)=satisfied stable
28	6	11	"	lma+1 yr.
59	6	11	"	nuclear family size=4+

interviews were conducted: evidently there has been upward movement for some.

Regarding attitudes to education (informant's own education), 10 have a liberal attitude, whilst 3 regarded their education as utilitarian (acquisition of basic skills, RRR).

With respect to the education of their children, 14 have a liberal attitude, and 22 think their children's education should be job-oriented. (NB. These figures, summed, exceed the number of cases in the K- this is because MC is possible for this set of features - Q11).

4 agree with the notion of giving a different kind of education to their sons than their daughters, whilst 9 disagree. (The remainder are neutral with respect to this Q.).

In KA, almost equal proportions (4:5) claimed that there should, and should not be (respectively) a difference in the way boys and girls are educated.

In KB, the ratio is 4:9. i.e. in KB, the belief that girls and boys should be given equal education opportunities is held by twice as many than is the reverse opinion. Attitudes to education tend to be more positive in this K, than in KA.

VAR 125. 20/26 of KB are female (this fact is not evidenced in the part of the CLUSTAN diagnostic list shown in Table 67 , as sex is treated as a binary variable: 1 = Male, Ø = Female: thus it appears only as a negative diagnostic).

Var 125 - 'housework as a hobby' is evidently related to the sex distribution across the clusters. Like KA's high frequency for 'drinking as a hobby', however, the sex distribution only partially accounts for the high frequency.

40% of women in KB responded positively to 'housework as a hobby', whilst only 35% of women in the total sample did so.

(The difference between within-K: sample frequencies, however, is

much greater for the former example of the sex-correlated variable, 'drinking as a hobby', than for this variable).

Regarding satisfaction with environment, KA showed a high rate of dissatisfaction: in contrast, KB is characterised by relatively higher frequencies for binary variables denoting satisfaction with both macro- and micro-environment.

VAR 85. (Macro-environemntal preference (type/size)) means a town of the same size as Tyneside is preferred.) (17).

VAR 88. Macro-environmental preference (location) - nowhere else. (20).

VAR 74. Micro-environmental preference (housing) - satisfied stable. (11). Moreover,

VAR 90. 'Positive Tyneside consciousness' emerges as diagnostic for this K. 15/26 (in K), out of a total of 25 positive responses in the sample.

Voting behaviour

In KB, 5 informants (out of a total of 6 in the sample) vote Conservative, and 13 Labour. (As opposed to 1:11 in KA).

5 informants from KB disapprove of the notion that there should be a connection between occupation and voting behaviour: whilst 11 accept, or approve of this.

Hobbies

T.V. viewing is cited as a leisure pastime by members of this K: Var 123 emerging as diagnostic:

i.e. viewing habits = non-intense and non-selective (cf. KA's intense / non-selective frequency).

To summarise, the membership of KB is distinct from that of KA mainly in sex distribution. KB also contains individuals of higher occupational and education status than KA, though there is overlap between

the 2 K's in terms of most of these categories.

KB generally has greater tendencies towards social stability: some members of KA are ambitious to move elsewhere (on a local, or larger, scale); whilst members of KB are apparently more satisfied with their living environment, and apparently more integrated into the local (micro) community. (See KB's diagnostics for social integration with neighbours). KB also has a higher frequency of + Positive Tyneside consciousness.

KC - Diagnosticity of Binary (Social) Variables

(NB. The magnitude of the values for BPFR's is a consequence of the small size of this K in relation to the size of the whole sample; 7:52).

Table 68 lists the positive CLUSTAN diagnostics (in descending BPFR order, to the level of 1.10).

KC is made up of 4 males, 3 females.

3 informants in KC have backgrounds of mixed regionality. (Only 1 other informant, STEPH, (KB), has regionality other than (as well as) UK. Northern. STEPH has also lived in UK London/SE (for at least 2 years).

In KC, XSHAW has lived in UK. N.W., as well as UK. Northern (the norm for this sample).

XSPRIG has UK. N.W., and UK. London/SE,

XWAIT has UK. E and W. Ridings, and UK. Midland, as well as UK. Northern. (See also the variables concerning number of moves per 5 year period before/after marraige, for an indication of a higher rate of geographical mobility in this K. See also those variables relating to distance of spouse's primary regionality).

This cluster, as well as including informants with mixed geographical background, shows tendencies to heterogeneity with respect to the social indices (education, occupation, age, sex) discussed in the foregoing.

Environmental Preferences

From the CLUSTAN diagnostics shown in Table 68 , we see that KC

Table 68.

Social diagnostics for linguistic cluster KC.

VAR	no. in K	no. in sample	BPFR	definition
153	1	1	7.43	1st occup.=2
152	1	1	"	occup.=2
151	1	1	"	hobbies=10 (bridge)
150	1	1	"	" =1 (golf)
147	1	1	"	reg.=Midland
145	1	1	"	citiness=Leeds
144	2	2	"	" =Merseyside
149	1	1	"	parent's reg.=NW
148	2	2	"	reg.=NW
146	2	2	"	" =E and W Ridings
87	3	4	5.58	mac. env. pref.=North
61	2	3	4.96	nuclear family size=6+
34	2	3	"	fe= tech/secretarial college, nursing
2	2	3	"	citiness=London
108	2	3	"	1st occup.=4
15	1	2	3.72	no. moves before marriage=5+
18	1	2	"	" " after " =2+
6	1	2	"	reg.=London SE
119	1	2	"	TV only
71	2	4	"	mic. env. pref.(housing)=neutral
97	1	2	"	father's occup.=3
25	1	2	"	age=70+
33	1	2	"	fe=university/polytechnic
60	2	5	2.98	nuclear family size=5+
28	4	11	2.71	lma+1 yr.
130	1	3	2.48	hobbies=5
42	1	3	"	attit. ed. of children=negative
83	3	10	2.23	mac. env. pref.=rural

Table 68 (cont.).

VAR	No. in K	no. in sample	BPFR	definition
113	3	11	2.03	job preference= R
68	4	15	1.99	mic. env. pref. (sentiment)=neutral
98	1	4	1.86	father's occup.=4
103	1	4	"	occup.=4
29	1	4	"	lma+2 yrs.
54	2	8	"	widowed
122	3	12	"	TV intense/ selective
73	3	12	"	mic. env. pref. (housing)=satisfied ambitious
116	2	10	"	job satisfaction=fairly low (OM)
110	1	5	"	1st occup.=6
132	5	25	"	hobbies=8
59	2	11	1.36	nuclear family size=4+
17	2	11	"	no. moves after marriage=1+
127	4	23	1.30	lma
64	2	12	1.24	sex distrib. of children=M bias
22	4	24	"	age=40+
58	4	24	"	nuclear family size=3+
14	2	12	"	no. moves before marriage=1-3
24	1	6	"	age=60+
67	1	6	"	distance spouse's reg.=50m+
53	1	6	"	marital status=single
104	3	19	1.18	occup.=5
115	4	26	1.15	job satisfaction=med um (OM)
26	4	26	"	sex=M
21	5	34	1.10	age=30+

has 3 members definitely preferring to live in the North (VAR 87), 1 out of a total of 4 in the whole sample).

3 informants would prefer rural to urban living (VAR 83).

3 are coded as 'satisfied ambitious', and 2 'neutral', with respect to micro-environmental preference in terms of housing;

4 are coded neutral with respect to micro-environmental preference in terms of sentiment.

Education

4/7 informants stayed on at school at least one extra year.

3 underwent full-time tertiary education, one at university or polytechnic.

1 informant has a negative attitude to his children's education.

Occupation

Groups 2, 4, 5 and 7 are represented in KC. 3 of the 7 informants are coded 'R' in job preference criteria. (Combinations of "immediate gain", "learnt/no new elements", and "supervised").

2 are coded 'I'.

On T.V. viewing habits, 3 are coded "intense/selective", (Cf. KA: intense/non-selective, and KB: non-intense/non-selective).

Summary(KC)

In view of the heterogeneous nature of this cluster, with respect to the dimensions of SocSp, it is difficult to make any generalisations concerning the social characteristics of this group as a group. This exercise is rendered even less useful by the small size of KC, (7 cases).

However, it can be said that there is a greater affinity between KB and KC than KA and KC with respect to age distribution, and education index, ((KB + KC) account for all 6 informants who have undergone full-time teriary education), and on distributions across occupation groups.

KA consists of informants drawn from the lower educational and occupational groups, whilst KB and KC overlap with KA, but also include the (few) informants with higher educational attainment, and occupational status.

The discreteness of KC as a group, in each of the 3 linguistic spaces, must be accountable for (if accountable at all according to non-linguistic parameters), by geographic criteria:ie.the major division of the population in linguistic space corresponds to the geographical divide of the River Tyne.

The question arises then , is the 2K level a more appropriate cutoff point for the sociolinguistic classification of this subsample?

Several considerations make this possibility reasonable:

1. For the 3 phonological classifications, the 2K level extends across a wide range of values for D^2. (See Figs. 50, 51, 52 (pp.200-201) which plot the number of K's present by D^2 values). This is evidenced on the graphs by the extension of the plateaus at K = 2. (%FON2 also shows a plateau at K = 4). It is clear also from the dendrograms, (Figs.50-52) that the 2-K level may be significant.

2. This division of the sample at the 2K level shows stability of K-membership across all 3 linguistic subspaces. (K1 + K2) = (KA + KB) = ($K\alpha$ + $K\beta$) and, K3 = KC = $K\gamma$.

If we call K1 + K2
KA + KB
$K\alpha$ + $K\beta$ } LK1 (linguistic K1)

and K3 = KC = $K\gamma$ LK2,

then LK1:LK2 corresponds to the social criteria: residence south, and north, of the Tyne respectively.

We have, then, one non-linguistic factor correlating perfectly with the division of the sample in each of the 3 linguistic subspaces.

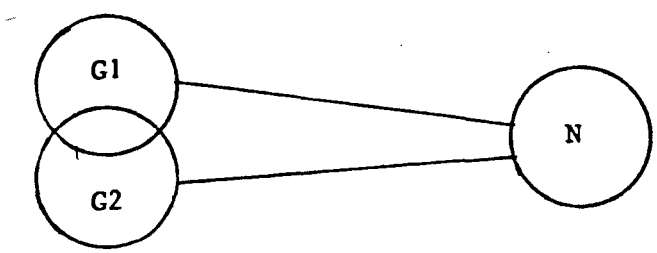
There are, however, several cogent reasons for preferring the 3K to

the 2K level.

1. For reasons enumerated above, in the description of the sample under investigation, (ch. 5), the 3-K level is preferable. Briefly: from our knowledge of the makeup of the Gateshead subsample (KA + KB etc.), we can predict that there are 2 social strata represented within this component of the sample: we also know that the 2 are quite closely related socially (in terms of the social measure "rateable value per dwelling by polling district"). Therefore it is to be expected that there may be overlap between the 2 strata with respect to non-linguistic indices.

Nor is it surprising that we find overlap, and interchange of K-membership, between these 2 strata across the 3 linguistic subspaces. This is precisely the kind of variability that is of interest to the sociolinguist.

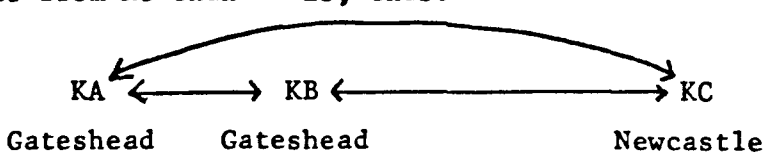
Given our knowledge of the <u>closer</u> relationship between members of these 2 strata, than between the 2 taken together in comparison to the Newcastle component, than we expect to find 3 clusters related thus:



G1, G2 – 2 Gateshead strata

N – Newcastle subsample

2. For the linguistic subspace examined, (%FON2) gradations in frequency of social variables have been demonstrated between the 3 linguistic K's, which show KA and KB to be closer to each other than to KC, and KA to be further from KC than KB is, thus:



(E.G. with respect to occupation groups, and level of education). These

gradations would be submerged at the 2K level.

More importantly, the sex trends discovered between KA, and KB would be lost, if these 2 groups were fused into one.

Regarding the social clusters, the 2K level was discussed in ch. 5 and a similar situation was shown to hold for the social classification. At the 2K level, there are significant distinctions between the 3 clusters (SocKx, SocKy, and SocKz), which are levelled at the 2K level. (SocKx: (SocKy + SocKz)).

3. In the social classification, at the 2K level, the sample is not divided into Newcastle and Gateshead residents. (At 2-K, SocKz joins SocKy. SocKz contains the Newcastle subsample).

Therefore, to take the 2K level does not actually simplify the mapping from the linguistic spaces to the social space.

The 2-K level only apparently simplifies the picture of overlap of cluster membership: this is because the number of clusters is smaller, therefore the combinations are fewer.

If n is the number of clusters (2K or 3K),

and x is the number of classifications (1 soc. + 3 ling. = 4),

then n^x gives the maximum number of 'derived' clusters of the type SocKx, KlA , SocKx,KlAβ , etc. (A measure of degree of overlap between classifications.) For 3 clusters per classification, (3K level),

$$n^x = 71$$

For 2 clusters per classification (2K level),

$$n^x = 16.$$

Slide 1 plus slide 5 and slide 6 show that the four classifications superimposed at the 3K level gives 15 combinations (derived clusters) out of a possible 71.

Slide 1 plus slide 7 show that the four classifications superimposed at the 2K level generates 3 combinations, out of a possible 16.

The ratios 3/16 (2K) and 15/71 (3K) are quite similar in magnitude;

(0.1875 and 0.2113 respectively). Thus, in terms of the potential number of overlaps between classifications, the situation at the 2K level is approximately as complex as that at the 3K level.

At the 2K level, the apparent simplicity in the distribution of informants across clusters in the 4 classifications (SocSp, and 3 phonological subspaces), cannot be adduced as an argument for taking the 2K level as the most useful division of the population under study.

The notion of finding a single social variable which accounts for linguistic groupings of the sample is an attractively simple one, however, for the reasons given above the argument for taking the 2K rather than the 3K level is not sufficiently strong. It may be the case that the Newcastle/Gateshead distinction in the linguistic spaces has a purely geographical sociolinguistic cause. Or, the different sampling methods by which the Gateshead and Newcastle subsamples were drawn may have been influential.

As far as the Gateshead subsample is concerned: it must be stressed again that social and behavioural reality within a community is fluid, complex and non-discrete (see Trudgill's (1974) remarks on the nature of sociolinguistic diversity, quoted above, ch. 1 , and see also the discussion of polythetism (ch. 2) and my remarks on well-formed clusters which do not display definitive characteristics

In the light of these considerations, we do not expect to find discrete groupings, characterised by definitive social and linguistic features, within an urban community, unless we operate at the level of the stereotype.

CHAPTER 8

CONCLUSION AND SUMMARY

Some shortcomings inherent in the methodology of some first generation sociolinguistic investigations were indicated in the first chapter. It is important for the discipline that these shortcomings are understood, and that alternative approaches, which avoid these problems, are developed.

The T.L.S. provides one alternative approach to modelling sociolinguistic variability in the urban setting. The T.L.S. objectives, which include exhaustive representation of speech varieties and adequate characterisation of social networks within the sample population were stated. The strategies whereby these objectives are fulfilled were described (T.L.S. model: the coding frame, multivariate analysis).

The T.L.S. approach, however, raises some new problems for sociolinguistics: these were discussed, and so was the issue of how far the T.L.S. model succeeds in fulfilling the aims stated by Pellowe et al (1972). Some of these problems are specific to the T.L.S. (relating to the coding frame design itself, and the computational difficulties which were consequent on its implementation). Some are general problems in linguistics (e.g. analyst variability), and some problems are specifically concerned with classification theory.

The computational processes which were applied to the data are described (search programs, standardisation procedures, reorganisation of data collections, cluster analysis).

The results obtained from the clustering procedures are discussed in detail. Classifications of the sample based on linguistic data, and on social data are described, and the linguistic, and social, diagnostics of the groupings derived are examined. The relationships between these classifications are discussed, in terms of the distribution of the sample across clusters, and in terms of the power of diagnostic features to discriminate clusters.

Some problems in defining a key diagnostic feature were encountered:

several definitions were discussed, in relation to the diagnostic statistics derived on social variables.

In addition to deriving a social classification, I arranged for social diagnostics to be generated directly for linguistic clusters. This procedure bypasses any distortion inherent in the definition of the dimensions of the social space (e.g. caused by inclusion of irrelevant variables, or by variables being effectively weighted because of the number of binary variables needed to express the range of variants).

The findings described have important theoretical and practical consequences for sociolinguistics.

Several points are worth emphasising again:

1. A single phonological variable is often-realised by several qualitatively different states, or PDVs. This is true for single informants as well as across the sample of Tynesiders. There is therefore a multiplicity of localised variants of phonological entities. This variability cannot be adequately represented on a unidimensional prestige to stigmatised scale.

2. The scatterplots discussed above (ch.4) plot the relative frequency of usage of states (as variants of OUs) across the sample. For some states, the sample is continuously distributed across the whole range of values from 0 to 100%, with no breaks in the curve. With respect to some phonological variants, then, the sample displays continuity of variability, and it is not possible to divide the sample into groups, discrete or otherwise, on the basis of those variants. A large class of states display this kind of continuity of variability.

3. The scatterplots show, in addition, that there are several kinds of distribution pattern: that mentioned under 2 above is one type. However, even for scatterplot curves with similar forms, the ranking of informants with respect to relative frequencies of usage of states differs. Even where groupings emerge from the scatterplots, those groupings are differently

constituted (in terms of members) for different states.

This fact suggests that isomorphous distribution patterns (cf. Trudgill's (1974, p.95) "typical" patterns) are the exception rather than the rule: certainly at the level of analysis of the state.

4. No diagnostics emerged, either for social, or linguistic, clusters, which have the power to assign individuals to groups with a success probability of 1. That is, there were no features whatsoever which were exclusive to, and uniform for, a given cluster.

In other words, the groupings obtained are polythetic classes: there are no necessary and sufficient criteria for group membership. Clusters are, however, bound together by mutual similarity between members across the range of variables included.

5. By partitioning off 3 different subsets of the phonological variables, and processing them separately, some interesting results were obtained.

(a) a selected sub-set of variables determines the classification obtained. Classifications of the same sample vary across the 3 segmental phonological subspaces, in terms of cluster membership. Each of these sub-sets of variables produces a partial classification. This must be taken into account if an investigator chooses to restrict his study to only a few variables.

(b) The different range of D^2 values (distancing of clusters, and individuals) varied across the 3 subspaces.

Thus we can conclude that for this sample population, dipthongal and triphthongal vowel OUs (%FON2) are more variable than monophthongal vowel OUs (%FON1). These in turn are more variable than consonantal OUs (%FON3).

However, all three subspaces demonstrate variability: all 51 OUs are shown to be variable features, and therefore potentially sociolinguistically salient.

(c) Those diagnostics which did emerge show that groupings are

discriminated by different phonological variables at different depths of analytic delicacy.

Clusters are discriminated at PDV level by OUs

ʌ I eɪ əʊ ə4a ə3

and at state level by OUs

ʊ aɪ 3 g r ŋ (bound morpheme)

It seems, then, that the level of delicacy of linguistic analysis influences the groupings into which the sample falls. This too is an important issue for linguists and sociolinguists.

6. A comparison of the linguistic clusters and the social clusters shows that linguistic variety clusters are not co-extensive with social clusters. This is true of the 3 subspaces (sub-sets of segmental variables) and of the 'derived' clusters, which represent the whole segmental subspace. The relationships between linguistic behaviour and social group membership are not simple. (This was found to be true at the 3K and the 2K level). Therefore to seek relationships between single linguistic features and single social factors (or social indices based on a small number of sociological variables) bypasses completely the complexity of sociolinguistic differentiation.

It must be remembered that the clustering methods applied represent only one combination of the statistical techniques available. Different classifications could be generated by using different distance measures and clustering algorithms. However, although this would produce a different set of clusters from the same data, it is surmised that the general conclusions drawn here would not be contradicted by the use of different statistical measures. It is hypothesised that the findings outlined above are unaffected by how far this classification approaches optimality.

One extension to the present research will involve testing a variety of similarity coefficients, combined with different clustering algorithms, in order to test this hypothesis.

Another extension to this research has already been mentioned (ch.4). As indicated above, the present findings show that some phonological features are variable at state level, and some at PDV level, for this sample. The sample will be re-classified on the basis of PDV scores rather than state scores. It will then be possible to discover whether linguistic clusters based on scores for variants at a higher level of analytic representation are more simply correlated with social group membership.

Furthermore, we can also discover whether clearer linguistic diagnostics emerge at the PDV level, than at the state level.

When the new version of CLUSTAN(CLUSTAN 1C) becomes available, it will also be possible to run the R-analysis which was planned. Hence we can discover dependencies between variables.

These strategies will provide useful indications as to how the measurement space can be refined, (e.g. we can select the most significant analytic level at which to code realisations of a given OU).

Another opportunity for further research is the possibility of investigating analyst variability. This can be achieved by analyst (b) duplicating the analyses made by analyst (a), and by clustering the two versions of the same set of informants as if they were different individuals. If case (1a) and case (1b) (the same case analysed by the two researchers) do not occupy the same place in the classificatory space, then we can calibrate analyst differences with respect to all variables (i.e. all dimensions of the space).

The fact that the sequence of 5-digit codes in the raw data files corresponds to the sequence of speech segments realised through time in the interview makes several other lines of research possible. Two are mentioned here.

Firstly, we can look at the effects of linear phonetic/emic context on phonetic realisations, by classifying PDVS (or states) according to preceding and following context. Secondly, we can test whether the speech of informants changes throughout the interview. (I.e. without the inter-

vention of the interviewer provoking style changes).

This can be achieved by taking successive sections of each informant's raw data file, and computing relative frequencies of usage of variants of each OU for each section. If the relative frequencies of variants do not change across successive sections of the raw data file, we will know that the informant in question maintained his speech variety consistently throughout the interview.

This piece of research deals only with a sub-sample of the T.L.S. informants, investigated on segmental phonological data and social data. Research will continue along two major lines: firstly, the whole T.L.S. sample will be analysed, and, secondly, the data from other linguistic systems (2-alpha and 3-alpha data) will be incorporated into the investigation.

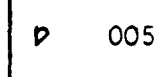
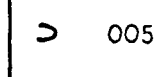
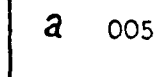
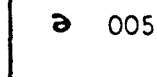
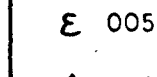
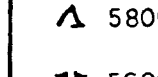
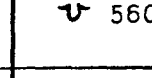
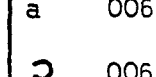
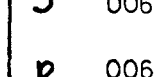
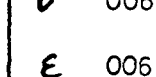
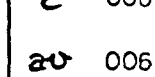
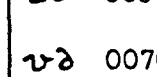
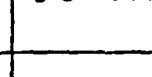
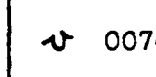
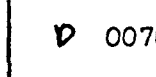
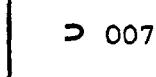
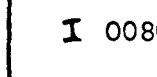
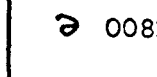
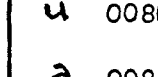
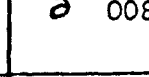
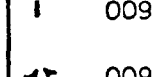
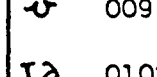
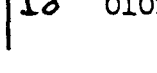
When the T.L.S. model is fully implemented and refined, we will have established a comprehensive, and empirically determined, model of sociolinguistic variation on Tyneside.

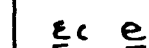
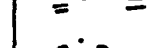
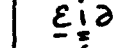
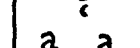
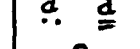
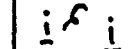
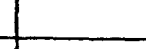
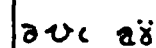
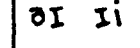
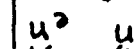
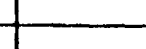
APPENDIX A

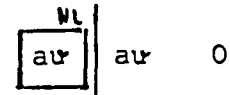
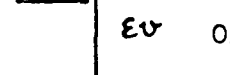
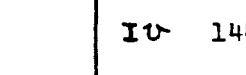
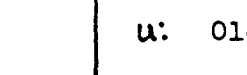
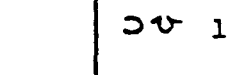
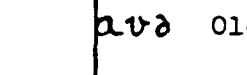
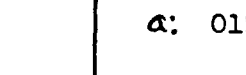
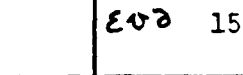
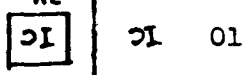
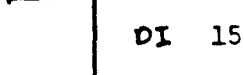
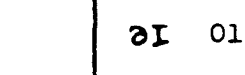
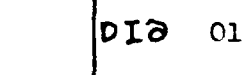
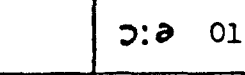
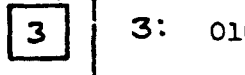
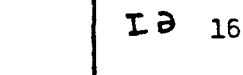
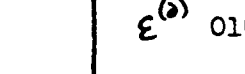
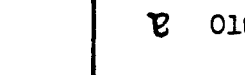
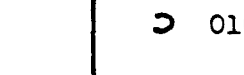
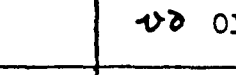
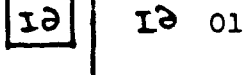
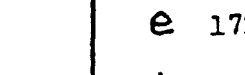
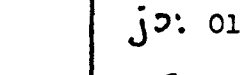
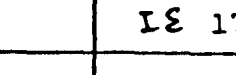
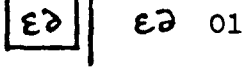
Appendix (A)
Specification of segmental phonological variables.

OU	PDV	(code)	states	lexical examples
1 [i:] NL	i:	0002	i i i i ɨ i	week, treat, see
	I	0004	i i i i ɨ	week, relief
	ɛ	0006	e ɛ e ɛ	beat
	eI	0008	ɛi əi ei ɛi	see
	Iə	0010	iɛ iɛ iə	feed
	Ii	0012	ii(back) ii(low) i	we, see
2 [I] NL	I	0014	i i i i ɨ	fit, big, till
	ɐ	0016	ə ə ə ə ɜ ɜ	shilling
	Iə	0018	iə iə iə	did
	ɜ:	0020	ɜ ɜ ɛ	shilling
	ɛə	0022	eɨ ɛɨ	miss, big
3 [ɛ] NL	ɛ	0024	ɛ ɛ ɛə ɛ e ɛ	well, men
	i:	0026	əi i i ɨ iə	head, bread
	I	0028	ɨ i i i i	centre, never
	ə	0030	ə əi əi ə ə	well, many
	ɛə	0032	ɛə e(ə)	men, embassy
4 [æ] L	æ	0034	a a a a ɑ̃	path,grass
	ɛ	0036	ə ə ɛ ɛ	have, after
	ɑ	0038	ɑ ɑi ɑ ɑ ɑ	path, grass
	ɔ	0040	ɔ ɒ ɒ ɔ ɔə ɔ ɔ	alsation
5 [ɑ] L	ɑ	0042	ɑ ɑi ɑ ɑ ɑ	father, card, half
	ɔ	0044	ɔ ɒ ɒ ɔ ɔə ɔ ɔ	farm, card
	æ	0046	a a a a	half, rather

NL=non-localised, L=localised.

6 [ɒ] L	ɒ	0050		off, one, along
	ɔ	0052		often, involved
	a	0054		swan, holiday
	ə	0056		because
	ɛ	0058		wash, long
	ʌ	5800		one, none
	ʊ	5600		once, because
7 [ɔ] NL	a	0060		all, talk
	ɔ	0062		or, four
	ɒ	0064		auction
	ɛ	0066		more, sore
	əʊ	0068		four, more
	ʊə	0070		door, course
8 [ʌ] NL	ʌ	0072		cup, onion
	ʊ	0074		pub, cup
	ɒ	0076		hurry, onion
	ɔ	0078		pub
	ɪ	0080		mother, just
	ə	0082		cup, onion
9 [ʊ] NL	ʊ	0084		pull, put
	ʉ	0086		book, cook
	ə	0088		good, butcher
10 [u] NL	u	0090		moon, two, suit
	i	0092		do, you, who
	ʊ	0094		boot, school
	ɪə	0102		tune

11 eI	eI	0104		eight, great
	ɛ	0106		take, make
	ɛɪə	0108		shape, railway
	a	0110		take, halfpenny
	i	0112		great, brains
	iə	1120		great, brains
	ɛɪ	0114		eight, straight
	ə	1140		Monday, holiday
12 əʊ NL	əʊ	0116		so, phone, nose
	əɪ	1160		so, no
	ɔ:	0118		so, smoke
	u:	0120		go, nose
	ɑ:	0122		old, know, no, cold
	ɪə	0124		stone, home
	ɛʊ	0126		bolt, hope
	ə	1260		pillow, yellow
13 aI NL	aI	0128		I, side, china
	ɑ:	0130		I, five
	i:	0132		blind, right
	eɪ	0134		knife, mine
14 aɪə NL	aɪə	0136		fire, tyre
	ää	0138		tyre, reliable
	ɑ:	0140		fire, trial
	ɛiə	1400		tyre, reliable

15	NL aʊ	aʊ	0142		house, now, croud
		ɛʊ	0144		house, crowd
		ɪʊ	1440		now, cow
		ʉː	0146		mouse, round
		ɔʊ	1460		loud, down
		ɑʊə	0148		flower, our
		ɑː	0150		our, tower
		ɛʊə	1500		our
16	NL ɔɪ	ɔɪ	0152		bouy, toil
		ɒɪ	1520		noise, toy
		ɪe	0154		buoy,noise
		ɒɪə	0156		boil, toil
		ɔːə	0158		boil, boy
17	NL ɜ	ɜː	0160		bird, fur, curl
		ɪə	1600		year
		ø	0162		bird, fur
		ɜ(ə)	0164		girl, curl
		ɐ	0166		bird, girl
		ɔ	0168		birth, worth
		ɑɹə	0170		burner, earth
18	NL eɪ	ɪə	0172		here, really
		e	1720		serious
		iː	0176		really, serious
		jɔː	0178		here, beer
		ɪɛ	1780		here, fear, beer
19	NL ɛə	ɛə	0180		hair, pair

	ɛ	0182	ɛ̝ ɛ̞ ɛ̞ˑ ɛ̞ː	care, there
	ɜː	0184	ɜ̟ ɜ	pair, hair
20 ʊə NL	ʊə	0186	ʊˑə ʊə	your, moor
	uə	0188	u̟ə u̟ˑə u̟ˑə u̟ˑ	poor, moor
	ɔ	0190	ɔ̟ə ɔ̟ː	more, poor
	ʊwɛ	0192	ʊwɛ ʊ̟wɛ ʊ̟wɛ	brewer, sewer
21 ə₃ final open				
	reduced	0194	ə₃ ə₄	baker, china
	non-red.	0196	ɛ̞ ɪ ɐ a	china, Sandra
22 ə₄a / ___C # [fortis] ___(r)#CoV..				
	reduced	0198	ə₃ ə₄	standard, interview
	non-red.	0200	ɛ̞ a ɪ ɛ	standard, interview
23 ə₂ə₁/[non-fortis] ___#				
	red.	0202	ə₁ ə₂	hammock, pavement, accent
	non-red.	0204	ɪ ɪ̝	pavement, almond
	non-red.	0206	ɛ	accent
24 I_1 /___C				
	red. (ə)	0208	ə₁ ə₄ ɪ	houses, places
	non-red.(ɪ)	0210	ɛ̞ iː	expect, perfect

25	I_2 /___ ##			
		0212	ɪ ɪi i əɪ ɛ(ɪ)	party, city
26	ɔ 4b [+low +tense] CC [1 str.]			
	red.	2000	ə	observe
	non-red.	2002	ɒ	observe
27	p	p init.	0214	p^{hh} p^h p $\underset{\circ}{b}$ ɓ p'	pot, spy
		p med.	0216	b $\underset{\bullet}{b}$ p^h p p(ingr)	happy, capital
		p fin.	0218	$\underset{\bullet}{b}$ ʔp p p' p^h $\overset{\times}{p}$ p(ingr)	dip
28	b	b init.	0220	b $\underset{\bullet}{b^h}$ p(b)	bag, bin, bring
		b med.	0222	b $\underset{\circ}{b}$ $\overset{\times}{b}$	robber, ribbon
		b fin.	0224	b^{∂} $\underset{\bullet}{b}$ $\overset{\times}{b}$	dab, rub
29	t	t init.	0226	t $\underset{\circ}{d}$ t^h	toss. stint
		t med.	0228	ɾ t t^h d $\underset{\circ}{d}$ ɾ $^{t}\theta$ ø	letter, matter
		t fin.	0230	t t^h $\underset{\frown}{t}$ t' d ɾ ʔ	hat, sit
30	d	d init.	0232	$\underset{+}{d}$ d($\underset{\circ}{d}$)	dish
		d med.	0234	d $\underset{\bullet}{d}$ t ʒ ʃ	rudder, window
		d fin.	0236	d^{∂} $\underset{\frown}{d}$ d $\underset{\bullet}{d}$ $t^{(h)}$	red, stupid
31	k	k init.	0238	k $c^{ç}$ k^h $\underline{k}$ $\overset{\circ}{g}$	kit, cope
		k med.	0240	k k^h $\underset{\times}{k}$ $\overset{?}{\circ}k$ ø	friction, Byker
		k fin.	0242	k $\underset{\times}{k}$ x $\overset{\bullet}{g}$ q k' ø $\overset{?}{\circ}k$ ʔ	sack
32	g	g init.	0244	g $\overset{\bullet}{g}$ ɣ	ground, gape
		g med.	0246	g g $\overset{\times}{g}$ ʔ	bigger

	g fin.	0248	g̊ g	log, fig
33 tʃ	tʃ	0250	tʃ t̪ʃ tˢʃ tˢ ʃ	church, French
34 dʒ	dʒ	0252	dʒ d̪ʒ dᶻʒ	jury
35 f	f	0254	f fʼ ɯf ʋ ɸ	fetch, half
36 v	v	0256	ʋ f w	very, drove
37 θ	θ	0258	θ θ̪ ᵗθ t̪ᶿ ð̃ n̪	thing, Arthur
38 ð	ð	0260	ð θ ᵈð ʋ n̪ ð̃	them, with
39 s	s	0262	s sˢ ɕ z	soup, business
40 z	z	0264	z s	rose, hose
41 ʃ	ʃ	0266	ʃ(ʃ) ˢʃ(s)	ship, rush
42 ʒ	ʒ	0268	ʒ ʃʒ dʒ	garage, pleasure
43 h	h	0270	h ɦ hᶠ ø	happy
44 m	m	0272	m m̆ ŋ	hammer, Mary
45 n	n	0274	n nː ň m n̪ ŋ̪ ŋ	nice, rain
46 ŋ (free)	ŋ	0276	ŋ ŋ̈ ŋ̆ n ng ŋᵏ k	sing, singer, something
47 l				
/(V){V, j}.		0278	l l̪ l̠ l̪ ɭ ɦ ø	like, filling
/( )Co_V..		0280	l l̪ l̠ l̪ ɭ ɻ	cloud, acclimatise
/V_{*, C}		0282	ɭ ɦ t ø l	old, cold
/( )_#		0284	t ɭ l ɦ ø	bottle

48 [r]	r	0286	ɹ ɾ x ʁ r̥ r ɽ ɸ ɻ̊ ʃ ʋ	rich, Harry
49 [j]				
/alveolar {stop / fricative} u..		2800	ø j F j* FN	tulip, dew, issue
other contexts		0288	j j ʕ ø	pure, furious
50 [w]	w	0294	w hw ʍ w̥ w' v ʋ	wind, when, will
51 [ŋ (bound)]		2760	n ng n	walking, eating

FN Where F= voiced/voiceless alveolar fricative, and j*=F+j

APPENDIX B

Appendix (B) Social coding sheets.

T.L.S. INFORMANT PAGE 1 (SOCIAL)

(NOTE: NC refers only to missing data.)

1. Cityness of informant [multiple coding; 5 year criterion for 'immobile' informants]

/1 city	/2 big town	/3 market town	/4 other	/5 NC
Tyneside	e.g.	e.g.		
Teeside	Bristol	Grantham		
Merseyside	Nottingham	Hexham		
Clydeside	Leicester	Taunton		
London	Swansea	Shrewsbury		
Manchester	Edinburgh	Cambridge		
Birmingham	Cardiff			
Sheffield	Chelmsford			
Leeds	Peterborough			
Stoke	Reading			
Solentside	Oxford			
Belfast				
Dublin				

2. Regionality of informant [multiple coding; 2 year criterion for immobile informants]

/1 U.K. Northern /2 U.K. E & W Ridings /3 U.K. N.W.
/4 U.K. N. Midland /5 U.K. Midland /6 U.K. Wales
/7 U.K. Eastern /8 U.K. London S.E. /9 U.K. Southern
/10 U.K. S.W. /11 U.K. Lowlands /12 U.K. Highlands
/13 U.K. Ulster /14 Eire /15 New World /16 Antipodes
/17 Indian S-C. /18 Hamitic Africa /19 Germanic Europe
/20 Caribbean /21 S.E. Asia /22 Arab Africa /23 Romance Europe /24 Slavic Europe /25 S. America /26 NC

3. Regionality of both parents [multiple coding; 2 year criterion]

/1 /2 /3 /4 /5 /6 /7 /8 /9 /10
/11 /12 /13 /14 /15 /16 /17 /18 /19
/20 /21 /22 /23 /24 /25 /26NC

4. No. of moves per 5 year period before marriage.

/(1) 0 /(2) 1 /(3) 2 /(4) 3 /(5) 4 /(6) 5
/(7) 5+ /(8) NC

5. No. of moves per 5 year period after marriage.

/(1) 0 /(2) 1 /(3) 2 /(4) 3 /(5) 4 /(6) 5
/(7) 5+ /(8) NC

6. Age (1) 17-20 /(2) 21-30 /(3) 31-40 /(4) 41-50
/(5) 51-60 /(6) 61-70 /(7) 71-80 /(8) 81+

7. Sex M / F

T.L.S. INFORMANT PAGE 2 (SOCIAL)

8. School leaving age.

/(1) before legal /(2) legal minimum /(3) +1 /(4) +2
/(5) +3 /(6) +4 /(7) +5 /(8) NC

(Legal minima: Age 82+ chaotic
67-82 13
42-67 14
< 42 15)

9. Tertiary and further education.

/(1) none /(2) full-time univ. & poly. /(3) full-time tech. & nursing & secretarial /(4) full-time college of ed. /(5) block release /(6) day-release /(7) night school /(8) self-taught & corres. /(9) NC

10. Attitude to education (self) [multiple coding]

/(1) negative /(2) basic skills RRR /(3) liberal
/(4) job-oriented /(5) job-oriented & liberal /(6) NC

11. Attitude to education (offspring) [multiple coding]
(States as above)

/(1) /(2) /(3) /(4) /(5) /(6) NC

12. Distinction between education of boys & girls.

/(1) Yes /(2) No /(3) NC

13. Positive distinction between parental & school roles

/(1) Yes /(2) No (3) NC

14. Parental control of children [multiple coding]

/(1) Direct verbal /(2) Indirect verbal /(3) Direct physical
/(4) Indirect physical /(5) NC

15. Marital status

/(1) Married /(2) Single /(3) Divorced /(4) Separated
/(5) Widowed

16. Religion

/(1) active /(2) inactive /(3) anti /(4) NC

T.L.S. INFORMANT PAGE 3 (SOCIAL)

17. Nuclear family size [i.e. including breadwinner(s) & spouse]

/(1) 1 /(2) 2 /(3) 3 /(4) 4 /(5) 5 /(6) 6
/(7) 6 + /(8) NC

Note: Unmarrieds living at home with 2 parents are coded (3); unmarrieds living alone coded (1); married with no offspring coded (2).

18. Sex distribution of offspring (absolute numerical).

/(1) 0 bias /(2) F bias /(3) M bias /(4) NC
(NC includes NA)

19. Average age gap between offspring

/(1) 1 year /(2) 2 years /(3) 3-4 years /(4) 5-6years
/(5) 7-8 years /(6) 9+years /(7) NC.

(NC includes NA)

20. Distance of spouse's primary regionality.

/(1) same local authority /(2) $<$ 50 miles /(3) $\geqslant$ 50 miles
/(4) NC
(NC includes NA)

21. Micro-environmental preference in terms of sentiment.

/(1) neutral /(2) dissatisfied /(3) satisfied ambitious
/(4) satisfied stable /(5) NC

22. Micro-environmental preference in terms of housing.
(Same states as above)

/(1) /(2) /(3) /(4) /(5)

23. Interviewer's assessment of decoration, furnishing & domestic equipment.

(a) 'Taste' aspiration:

/(1) good /(2) bad /(3) indifferent /(4) NC

(b) Financial commitment to this taste:

/1 /2 /3 /4 /5 /6 /7 /8 /9 /10 /NC

[10 is high]

T.L.S. INFORMANT PAGE 4 (SOCIAL)

24. Rateable value.
(Quantitative; ∝ max. £)

25. Macro-environmental preference (finance &/or occupation no object).

(a)/(1) rural /(2) smaller town /(3) same size /

(b)/(1) south /(2) north /(3) nowhere else /(4) NC

26. Positive Tyneside consciousness.

Yes/No/NC

27. Social integration with neighbours (as claimed by informant). [multiple coding]

/(1) non-existent & unknown /(2) non-existent & known /(3) antagonistic /(4) minimal, pleasant /(5) cordial /(6) intimate /(7) NC

28. Father's occupation.

/(1) Professional & high administrative /(2) Managerial & executive /(3) Inspectional, supervisory & other non-manual, higher grade /(4) Inspectional, supervisory & other non-manual, lower grade /(5) Skilled manual & routine non-manual /(6) Semi-skilled manual /(7) Unskilled manual /(8) NC

[cf. Hall & Jones Brit. Jnl. Sociol. 1, 1950 pp.31 ff]

29. Informant's present occupation(or spouse's, if informant is not primary breadwinner).
(States as above).

/(1) /(2) /(3) /(4) /(5) /(6) /(7) /(8) NC

30. Informant's first occupation.
(States as above).

/(1) /(2) /(3) /(4) /(5) /(6) /(7) /(8) NC

31. Job preference

(a) prospects/immediate gain/NC

(b) thinking (new elements)/learned (no new elements)/NC

(c) supervised/self-deciding/NC

T.L.S. INFORMANT PAGE 5 (SOCIAL)

32. Job satisfaction (match between 31 & 29)

/(1) 3 /(2) 2 /(3) 1 /(4) 0 /(5) NC

33. Daily exposure to radio & television.

(a)/(1) predominantly radio /(2) predominantly television
/(3) radio only /(4) television only/(5) non own/(6) NC

(b)/(1) intense, non-selective /(2) intense, selective
/(3) non-intense, non-selective /(4) NC
(NC includes NA)

34. Regular drinking habit; housework as hobby.

Yes/No/NC

35. Leisure satisfaction.

satisfied/partially satisfied/disgruntled/NC

36. Hobbies.

/(1) active, expensive, rule-based, club (rackets hazard tennis)
/(2) active, expensive, rule-based, non-club
/(3) active, expensive, non-rule-based, club (hunting)
/(4) active, expensive, non-rule-based, non-club (D.I.Y.; Veteran car driving)
/(5) active, cheap, rule-based, club (amateur football)
/(6) active, cheap, rule-based, non-club (rounders)
/(7) active, cheap, non-rule-based, club (X-country)
/(8) active, cheap, non-rule-based, non-club (fell-walking, gardening)
/(9) sedentary, expensive, rule-based, club (roulette)
/(10) sedentary, expensive, rule-based, non-club (bridge, for stakes)
/(11) sedentary, expensive, non-rule-based, club (stud farm)
/(12) sedentary, expensive, non-rule-based, non-club (punter)
/(13) sedentary, cheap, rule-based, club (whist)
/(14) sedentary, cheap, rule-based, non-club (patience, scrabble)
/(15) sedentary, cheap, non-rule-based, club (potting, drama)
/(16) sedentary, cheap, non-rule-based, non-club (reading papers, painting)
/(17) active, expensive, club, collecting

/(18) active, expensive, non-club, collecting

/(19) active, cheap, club, collecting

/(20) active, cheap, non-club, collecting (sea shells)

/(21) sedentary, expensive, club, collecting (book clubs, picture clubs)

/(22) sedentary, expensive, non-club, collecting (stamps, antiques)

/(23) sedentary, cheap, club, collecting

/(24) sedentary, cheap, non-club, collecting (Shell cards, newsworthy faces, etc., green shield stamps?)

/(25) NC

37. Connection between occupation and voting behaviour

/(1) approve /(2) accept /(3) disapprove /(4) NC

38. Voting preference (usually last election).

/(1) Cons. /(2) Lab. /(3) Lib. /(4) other /(5) Comm.
/(6) Refusal /(7) Floater. /(8) NC

REFERENCES

ANDERSON, A.J.B. (1971) Ordination methods in ecology. J.Ecol., 59, pp.713-726.

BALL, G.H. (1965) Data analysis in the social sciences: What about the details? Proc. of the Fall Joint Computer Conferences, Stanford, pp.533-559.

BICKERTON, D. (1975) Dynamics of a Creole system. London: CUP.

BRENNAN, T. (1972) Numerical Taxonomy: Theory and some Applications in Educational research. Doctoral thesis, Univ. Lancaster.

BURNABY, T.P. (1966) Distribution free quadratic discriminant functions in palaeontology. Computer applications in the Earth Sciences. State Geological Survey, Lawrence, Kansas.

CATTELL, R.B. & COULTER, M.A. (1966) Principles of behavioural taxonomy and the mathematical basis of the taxonome computer program. Br. J. Math. Statist. Psychol. 19, part 2. pp.237-269.

CORMACK, R.M. (1971) A review of classification. J.Roy.Stat.Soc. (A). 134 (3).

CRYSTAL, D. (1969) Prosodic systems and intonation in English. Cambridge: CUP (Cambridge Studies in Linguistics 1).

CRYSTAL, D. & QUIRK, R. (1964) Systems of prosodic and paralinguistic features in English. (Janua Linguarum Series Minor, no. 39) The Hague: Mouton.

DOUGLAS, E. (1976) Sociolinguistic Variation in a Rural Community in Northern Ireland. In Reid, E. (ed.) pp.8-9.

ESLING, J. (1976) Articulatory setting in the community. In Reid, E. (ed.) pp.19-20.

EVERITT, B. (1974) Cluster Analysis. London: Heinemann.

GARVEY, C. & DICKSTEIN, E. (1972) Levels of analysis and social class differences in language. Lang. and Speech, 15, 1972, pp.375-384.

GILES, H. (1973a) Communicative effectiveness as a function of accented speech. Speech Monographs, 40 pp.330-331.

GILES, H. (1973b) Accent Mobility: a Model and some Data. Anthrop.Lings. Feb. 1973.

GOFFMAN, E. (1961) Encounters. NY: Bobbs-Merrill.

GOFFMAN, E. (1963) Behaviour in public places. London: Collier-Macmillan.

GOOD, I.J. (1953) The population frequency of species and the estimation of population parameters. Biometrika, 40, p.237.

HALL, J. & JONES, D.C. (1950) Social Grading of Occupations. Br.Jnl.Sociol., 1. pp.31-35.

HODSON, F.R., SNEATH, P.H.A. & DORAN, J.E. Some experiments in the numerical analysis of archaeological data. Biometrika, 53. pp.311-324.

IHM, P. (1965) Automatic classification in anthropology. In Hymes, D. (ed.), The Use of Computers in Anthropology. Mouton: The Hague.

JONES, V. (forthcoming) The TLS: an approach to data processing in sociolinguistics. SMIL.

JONES, V. & PELLOWE, J. (forthcoming) Representing dependencies in the structure of linguistic variation (Paper to III Intnl. Conf. on Methods in Dial. (Ontario 1978).).

KROEBER, A.L. (1960) Statistics, Indo-European and Taxonomy. Language, 36.

LABOV, W. (1963) The Social Motivation of a Sound Change. Word, 19.

LABOV, W. (1966) The Social Stratification of English in New York City. Washington D.C.: Center for Applied Linguistics.

LABOV, W. (1969) Contraction, deletion and inherent variability of the English copula. Language, 45. pp.715f.

LABOV, W. (1972) Some Principles of Linguistic Methodology. Lang.Soc., 1. pp.97-120.

LADEFOGED, P. (1960) The value of phonetic statements. Language, 36 pp.387f.

LAMB, S. (1965) Linguistic data processing. In Hymes, D. (ed.) The Use of Computers in Anthropology. Mouton: The Hague.

LOEVINGER, J. (1957) Objective tests as instruments of psychological theory. Psych.Rep., 3. pp.635-694.

MONOD, J. (1967) Juvenile Gangs in Paris: Toward a structural analysis. Jnl. of Research in Crime and Delinquency, 4, no.1.

MOSER, C.A. (1958) Survey Methods in Social Investigation. London: Heinemann.

NEEDHAM, R.M. (1967) Automatic classification in Linguistics. The Statistician, 17. pp.45-54.

OPPENHEIM, A.N. (1966) Questionnaire Design and Attitude Measurement. London: Heinemann.

PAHL, R.E. (1968) Readings in urban sociology. London: Pergamon. (ed.)

PELLOWE, J. (1967) Studies towards a classification of varieties of spoken English. Univ. Newcastle: unpub. MLitt thesis.

PELLOWE, J. (1970a) Establishing some prosodic criteria for a classification of speech varieties. Newc. Univ. Eng. Dept. mimeo.

PELLOWE, J. (1970b,c,d) Establishing speech varieties of conurbations: I,II,III. (Theoretical position: criteria and sampling, varieties as constructs) New.Univ.Eng.Dept. mimeo.

PELLOWE, J. (1973) A problem of diagnostic relativity in the TLS. Class. Soc.Bull. 3(1) pp.2-8.

PELLOWE, J. (1976) The Tyneside Linguistic Survey: aspects of a developing methodology. In Viereck, W. (ed.) Sprachliches Handeln - Soziales Verhalten: ein Reader zur Pragmalinguistik und Soziolinguistik. Munchen: Wilhelm Fink. pp.203-217, 365-367.

PELLOWE, J. (forthcoming) (ed.) Studies in unity and variety: collected papers of the Tyneside Linguistic Survey. in a special publication of the Philological Society.

PELLOWE, J. (forthcoming) Establishing variability in intonational systems (York Papers in Ling.)

PELLOWE, J. & JONES, V. (1978) On intonational variability in Tyneside speech. In Trudgill, P. (ed.) Sociolinguistic patterns of British English. London: Arnold.

PELLOWE, J & JONES, V. (forthcoming) Establishing intonationally variable systems in a multidimensional linguistic space. (Lang. & Sp.)

PELLOWE, J. & JONES, V. (forthcoming) Varietal variability of prosodic systems and its correlates. probably with Mouton (The Hague).

PELLOWE, J. & JONES, V. (forthcoming) Representing and interpreting the structure of linguistic variation. (Proc. Vth Intl. Symp. on Lit. & Ling. Comp. Aston 78.)

PELLOWE, J., NIXON, G. & McNEANY, V. (1972a) Defining the dimensionality of a linguistic variety space. Newc.Univ.Eng.Dept. mimeo.

PELLOWE, J., NIXON, G. & McNEANY, V. (1972b) Some sociolinguistic characteristics of phonetic analysis. In Rigault, A., & Charbonneau, R. (eds.) Proc. VII Intnl. Cong. Phon. Sci. Montreal (1971) The Hague, Mouton.

PELLOWE, J., NIXON, G., STRANG, B. & McNEANY, V. (1972) A dynamic modelling of linguistic variation: the urban (Tyneside) Linguistic Survey. Lingua 30. pp.1-30.

RADIN, G. & ROGOWAY, H.P. (1965) NPL: Highlights of a New Programming Language. CACM, 8. pp.9-17.

REID, E. (ed.) (1976) Abstracts of 1976 Research Seminar on Sociolinguistic Variation. West Midlands College: Communications Research Unit.

REID, E. (1976) Social and Stylistic Variation in the speech of some Edinburgh School Children. IN Reid, E. (ed.) (1976) p.16.

RINGAARD, K. (1965) The phonemes of a dialectal area perceived by phoneticians and by the speakers themselves. Proc. Vth Intnl. Cong. Phon. Sci. p.495. (Munster) Swirner, E. & Bethge, W. (eds.) Basel: Karger.

ROPER, J.S. (1973) PL/1 in easy stages. A programmed Learning Textbook. London: Elek Science.

ROSS, A.S.C. (1950) Philological probability problems. J.Roy.Statist.Soc. B12, pp.41-59.

RUUD, J. (1954) Vertebrates without erythrocytes and blood pigment. Nature, 173, pp.848-850.

SHUY, R.W., WOLFRAM, W.A. & RILEY, W.K. (1968) Field techniques in urban language study. Washington: Center for Applied Linguistics.

SMITH, P.M. (1976) Negotiative characteristics of interpersonal speech style shifts. In Reid, E. (ed.) (1976) pp.30-31.

SNEATH, P. H.A. (1957) The application of computers to taxonomy. J.Gen.Microbiol., 17. pp.201-226.

SOKAL, R.R. & MICHENER, C.D. (1958) A statistical method for evaluating systematic relationships. Univ. Kansas Sci.Bull, 38, pp.1409-1438.

SOKAL, R. & SNEATH, P. (1963) Principles of Numerical Taxonomy. San Francisco: Freeman.

STEVENSON, T.H.C. (1911) Census of England and Wales. London: H.M.S.O.

STRANG, B.M.H. (1969) Modern English Structure. London: Arnold.

STRANG, B.M.H. (1968) The Tyneside Linguistic Survey. (Paper read at Intnl.Cong.Dial. 1965, Marburg.) Zeitschrift fur Mundartforschung. p.788.

STRAUSS, J.S., BARTKO, J.J. & CARPENTER, W.T. (1972) The use of clustering techniques for the classification of psychiatric patients. Br.J.Psychiat.

TRUDGILL, P. (1974) The social differentiation of English in Norwich. Cambridge: CUP.

WARD, J.H. (1963) Hierarchical grouping to optimise an objective function. J.Am.Statist.Ass., 58. pp.236-244.

WEINREICH, U., LABOV, W. & HERZOG, M. (1968) Empirical foundations for a theory of language change. In Lehmann, W.P. (ed.) Directions for Historical Linguistics. pp.95-105. Austin: Univ. Texas Press.

WILLIAMS, W.T. & DALE, M.B. (1965) Fundamental problems in numerical taxonomy. In Advances in Botanical Research, 2(ed) Preston, R.D. London: Academic Press.

WILLIAMS, W.T. & LAMBERT, J.M. (1959) Multivariate Methods in plant ecology, 1. Association analysis in plant communities. J.Ecol., 47. pp.83-101.

WISHART, D. (1969) FORTRAN II Programs for 8 methods of Cluster Analysis (CLUSTAN 1). Computer Contribution 38. State Geological Survey, Univ. Kansas.

WISHART, D. (1969b) Mode Analysis. In Numerical Taxonomy, Cole, A.J. (ed.) pp.282-308. NY: Academic Press.

WOLFRAM, W.A. (1969) A Sociolinguistic Description of Detroit Negro Speech. Washington: Center for Applied Linguistics.

ZADEH, LA. (1972) Proc.Intnl.Conf. on Man and Computer. Bordeaux, France. pp.130-165. Basel: Karger.

ZADEH, L.A. (1973a) Outline of a New Approach to the Analysis of Complex Systems and Decision Processes. IEEE Trans. on Systems, Man and Cybernetics, Vol. SMC-3. pp.28-44.

ZADEH, LA. (1973b) The concept of a linguistic variable and its application to approximate reasoning. In Fu, K.S., & Tou, J.T. (eds.) Learning Systems and Intelligent Robots. pp.1-10. NY: Plenum Press.