Chapter Four

STATISTICS FOR LINGUISTS: SOME CASE STUDIES TO ILLUSTRATE TECHNIQUES AND THEIR APPLICABILITY*

INTRODUCTION

The aim of this chapter is to give detailed examples of some of the statistical techniques discussed in general terms in Chapter One. The case studies examined are taken from the linguistics literature or from work in progress. For a more complete discussion of these techniques, readers are referred to Butler (1985) and Woods et al. (1986).

MEASURES OF CENTRAL TENDENCY AND VARIABILITY

The mean, median and mode

To illustrate the calculation of the mean, median, mode, variance and standard deviation, we shall take a study of word length which formed part of an investigation into style shifts in four books of poems by Sylvia Plath (Butler, 1979). It was hypothesised that the language of the earlier poems would be formally more complex than that of the later poems, and that as part of this general expectation, word length would be higher, on the whole, in the earlier than in the later work. Here, we shall examine the data for just one book of poetry, *The Colossus*.

The word length distribution for this set of texts is shown in Table 1, and has been converted to a frequency polygon in Figure 1. A word was defined as any sequence of letters, hyphens and apostrophes bounded by spaces or punctuation marks.

* Christopher Butler, Department of Linguistics, University of Nottingham, UK.

Table 1

Word length distribution in Plath's *The Colossus*

Word length	Frequency
1	361
2	1280
3	1832
4	1500
5	1371
6	872
7	643
8	389
9	240
10	161
11	69
12	55
13	35
14	8
15	5
16	3
19	1
20	1
22	1

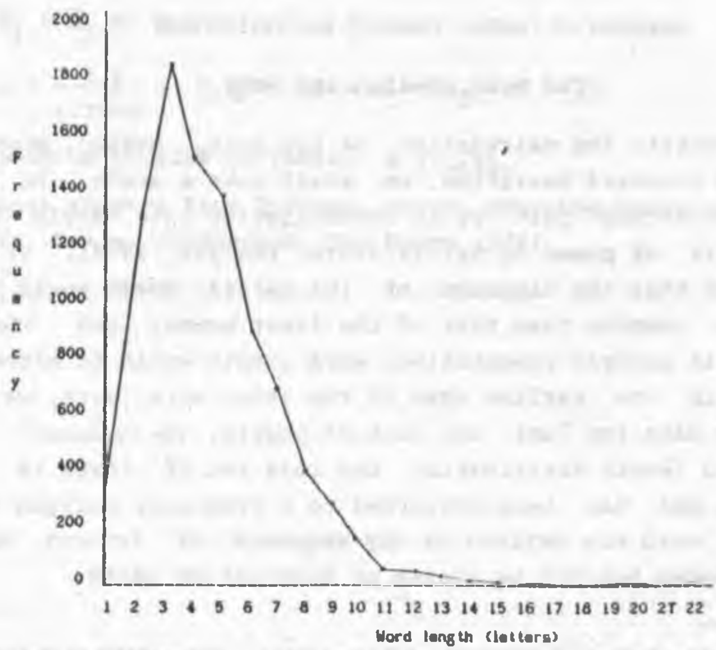

Fig. 1. Frequency polygon for length in *The Colossus*

To find the mean we use the formula:

$$\bar{x} = \Sigma fx/N$$

where $\bar{x}$ is the mean
x is a particular value of the word length
f is the frequency of that value
N is the total number of words
Σ means 'sum of'

So we have:

$$\bar{x} = (361 \times 1 + 1280 \times 2 + 1832 \times 3 \ldots + 1 \times 22) / (361 + 1280 + 1832 \ldots + 1) = \underline{4.54} \text{ letters}$$

The median is the value above which and below which equal numbers of observations fall. The total number of words is 8827 so to find a rough value for the median, we want the length of 4413rd word in ranking order. Adding up the frequencies for each length, starting with length 1, we find that the 4413rd word lies in the 4-letter category. A more exact value of the median is given by:

$$\text{Median} = L + \frac{N/2 - F}{f_m}$$

where: L = lower bound of category in which median occurs (= 3.5 if we treat each integer as representing a range from 0.5 below it to 0.5 above it)
N = total number of words (= 8827)
F = total number of words in lower categories (= 361 + + 1280 + 1832 = 3473)
f_m = frequency of the category in which the median occurs (= 1500)

thus, the median = 3.5 + (8827/2 - 3473) / 1500 = 4.13 letters.

The mode is simply that value which has the highest frequency, and is clearly 3 letters.

The distribution is strongly positively skewed (see Figure 1), with the result that the mode is lower than the median, which is in turn lower than the mean.

The variance and standard deviation

The variance is given by:

$$\text{Variance} = \frac{\Sigma f(x - \bar{x})^2}{N - 1}$$

However, a computationally more convenient expression which does not involve the subtraction of the mean is:

$$\text{Variance} = \frac{\Sigma fx^2 - (\Sigma fx)^2/N}{N - 1}$$

where: x = a word length

f = frequency of this category

N = total number of words

$\Sigma fx^2 = (361 \times 1^2 + 1280 \times 2^2 + 1832 \times 3^2 \ldots + 1 \times 22^2) = 230469$

$\Sigma fx = (361 \times 1 + 1280 \times 2 + 1832 \times 3 \ldots + 1 \times 22) = 40057$

Thus, variance = $(230469 - (40057)^2 / 8827) / (8827-1)$ = 5.52 letters and the standard deviation (s) is given by:

$$s = \sqrt{\text{Variance}} = \sqrt{5.52} = 2.35 \text{ letters.}$$

TESTING FOR SIGNIFICANT DIFFERENCES IN CENTRAL TENDENCY BETWEEN DATA SETS

The Mann-Whitney U-test

As our first illustration of hypothesis testing in relation to differences in central tendency, we shall examine part of a study by Lahey (1984) on the language of a patient suffering from cerebral atrophy. The data were taken from daily logs written by the patient over a period of 4½ years. Ten samples were taken at intervals of 6 months, each consisting of the first 30 interpretable sentences from each of the sampling periods. One variable studied was the proportion of clauses which were related in some way to other clauses in the text, and could be categorised as having a function in the larger-scale structure of the text. The proportions of such clauses were compared in the first 5 and second 5 samples, to test for changes over time. The relevant data are given in Table 2.

Lahey uses the Mann-Whitney U-test to compare the two subsamples. No justification is given in the paper for this choice, but it is sensible for the following reasons (see also the flowchart in Chapter One)

(a) It does not assume anything about the distribution of the data, or about the magnitudes of the variances for the two samples.

Table 2

Clauses with function in textual macrostructure in writing of patient with cerebral atrophy

Sample no.	No. of clauses	No. with function	% with function
1	42	42	100
2	40	39	97.5
3	36	25	69.4
4	33	28	84.8
5	33	21	63.6
6	31	21	67.7
7	31	20	64.5
8	30	18	60
9	35	19	54.3
10	33	20	60.6

(b) It assumes only an ordinal level of measurement, so does not attach importance to the actual magnitudes of the proportions, but rather to their rank ordering

(c) The data are being treated as 5 independent samples within each of two time spans, all the data coming from one subject (different, therefore, from the 'repeated measures' design where a number of subjects each perform under two separate sets of conditions).

We now rearrange the data for convenience, and rank the whole set of 10 proportions from lowest (= rank 1) to highest (= rank 10), as in Table 3, then find the sums of ranks for each sample (R_1 and R_2).

We now calculate the U statistic for each sample as follows:

$U_1 = N_1N_2 + N_1(N_1 + 1)/2 - R_1 = 5 \times 5 + 5 \times 6 / 2 - 38$

$= 25 + 15 - 38 = \underline{2}$

$U_2 = N_1N_2 - U_1 = 5 \times 5 - 2 = \underline{23}$

We now take the smaller of U_1 and U_2, ie. 2, and compare it with the critical value. The critical value of U for $N_1 = N_2 = 5$ is

Table 3

Ranks for data on patient with cerebral atrophy

Early group ($N_1 = 5$) Propn.	Rank	Later group ($N_2 = 5$) Propn.	Rank
100.0	10	67.7	6
97.5	9	64.5	5
69.4	7	60.0	2
84.8	8	54.3	1
63.6	4	60.6	3
Sum of ranks:	38 (R_1)		17 (R_2)

2 in a directional test at the $p \leq 0.025$ level. The observed value must be smaller than or equal to the critical value for significance, so the results just achieve significance at this level.

The sign test

As a second example of the testing of hypotheses about the difference in central tendency between two data sets, we shall take a project carried out by the author (Butler, 1982). Ninety-seven first year university and polytechnic undergraduate students were played a tape of a number of utterances, each consisting of a sentence concerned with opening a window, with a modal verb in a particular mood construction, spoken with the unmarked intonation pattern for that mood type. Written versions of the sentences were also provided. The informants had to imagine that the utterance on tape was being used to get an acquaintance of the same sex, age and status to open a window. They were then asked to rate the utterance for politeness in this directive function, on a scale from 1 (very impolite) to 7 (very polite).

The results considered here are those for just one pair of utterances: those of *Open the window, will you?* ('No 1' in what follows) and *Will you open the window?* ('No. 2'). One informant found one of these to be unacceptable as a directive, and so was discarded from the analysis. The ratings for the other 96 informants were as shown in Table 4.

Table 4

Politeness ratings for two modalised directives

No. 1	No. 2	No. 1	No. 2	No. 1	No. 2
3	4	1	5	5	6
5	4	3	6	5	4
7	5	5	4	5	6
5	6	4	5	3	4
4	4	5	5	5	4
4	4	5	4	6	4
4	3	5	6	3	4
5	5	5	6	2	5
5	6	2	4	4	4
5	4	5	4	3	6
1	5	5	6	4	4
5	3	5	4	1	5
5	5	4	5	3	[illegible]
5	4	6	6	2	4
3	5	6	7	4	4
3	3	2	4	4	5
5	6	3	5	2	5
2	5	5	6	3	4
1	5	4	5	6	6
4	4	4	4	4	5
4	6	5	2	4	5
4	6	3	4	5	6
6	4	4	5	6	6
3	5	5	4	6	5
4	4	4	4	2	6
4	5	3	5	5	6
2	3	4	4	4	6
2	2	6	6	4	5
3	5	4	4	3	5
5	5	4	4	4	5
3	6	4	5	5	6
5	5	4	4	3	4

Since the data are ordinal (one would not want to claim that politeness can be rated on a scale with exactly equal intervals), and the design is of the repeated measures type, the appropriate test is the sign test (see the flowchart in Chapter One). To perform this test, we record the sign of the difference between each pair of ratings, subtracting one from the other in a consistent manner. (Rating for No. 2 - rating for No. 1) is positive for 54 pairs, negative for 17 pairs, and zero for 25 pairs. The tied scores are dropped, and the number of pairs, N, reduced accordingly, to 71. The test statistic, x, is the number of pairs with the less frequent sign of the difference, ie. 17. Where we have a fairly large number of pairs of observations (say 25 or more), we convert the x statistic to a 'z-score' which can then be reffered to a table of values for the 'normal' distribution curve:

$$z = (N - 2x - 1) / \sqrt{N} = (71 - 2 \times 17 - 1) / \sqrt{71} = \underline{4.272}$$

No. 2 was predicted to be more polite than No. 1. The critical value of z in a directional test for $p \leq 0.001$ is 3.10, and since the calculated value is greater than this, the difference is significant at this level.

TESTS OF ASSOCIATION OR INDEPENDENCE

To illustrate the use of the chi-square test in testing for independence or association between variables, we shall look at part of a study by Connolly (1979) on diachronic shifts in Middle English syntax. The data are the frequencies of various positional arrangements of clause elements in 3 early and 3 late Middle English texts. We shall consider just one set of tests: those for the relative position of predicator (P) and direct object (O) in declarative affirmative clauses. The complete set of data is shown in Table 5.

Table 5

Frequencies of clauses with P + O or O + P orders in early and late ME texts

	Early ME			Late ME		
	Text 1	Text 2	Text 3	Text 1	Text 2	Text 3
P + O	69	91	76	128	103	117
O + P	10	16	15	3	4	8

Connolly first tests for homogeneity (i.e., for lack of any significant association between element order and text number) within each group of texts, using the chi-square test which, it will be remembered, compares the observed frequencies with those which are expected, here on the basis of the null hypothesis of no association between the variables. Note that the data are raw frequencies of occurrence of entities classified on a nominal, yes/no basis.

Table 6

Observed and expected frequencies of clauses with P + O or O + P order in early ME texts

	Text no.			Total
	1	2	3	
P + O	69 (67.31)	91 (91.16)	76 (77.53)	236
O + P	10 (11.69)	16 (15.84)	15 (13.47)	41
	79	107	91	277

The numbers in brackets in Table 6 represent the expected values for the set of early texts, calculated according to the following principle. Of the 277 clauses in the whole set of texts, 236 are of the P + O type, and the proportion of this type is thus 236/277. If there is no association between the variables, we should expect that this same proportion of the clauses would be P + O in each individual text. So we have:
Expected value of P + O for Text 1 = 236 x 79 / 277 = 67.31, etc.
We now calculate χ^2 as follows:

$$\chi^2 = \Sigma\,((\text{Observed} - \text{Expected})^2 / \text{Expected}) = (69 - 67.31)^2 / 67.31 + (91.16 - 91)^2 / 91.16 \ldots + (15 - 13.47)^2 / 13.47 = \underline{0.49}$$

In order to compare the calculated value with the critical value, we must also know the number of 'degrees of freedom' involved, defined here as (R - 1) x (C - 1), where R is the number of rows in the contingency table, and C the number of columns. Thus the number of degrees of freedom for a 3 x 2 table is (3 - 1) x (2 - 1) = 2.

The critical value for χ^2 at the $p \leqslant 0.05$ level and 2 d.f. is <u>5.99</u>; the value obtained is thus non-significant - ie. no association between element order and text number can be demonstrated.

An exactly parallel calculation for the late texts gives χ^2 = 2.78, again non-significant at the $p \leqslant 0.05$ level. However, there is a slight complication here. If we calculate the expected frequency for O + P in Text 2, we obtain a value of 4.42. For the chi-square test to be totally reliable, every expected value should be at least 5. So not quite so much credence can be placed in this result, and Connolly indicates this in his paper by bracketing his χ^2 value in this case.

Connolly now pools the frequencies in the homogeneous groups of texts, as shown in Table 7, and tests for association between element order and the period of the texts.

Table 7

Overall frequencies of clauses with P + O or O + P order in early and late ME texts

	Early texts	Late texts	Total
P + O	236	348	584
O + P	41	15	56
	277	363	640

For a 2 x 2 table, it is advisable to use a correction factor known as Yates' correction. Furthermore, in the special case of a 2 x 2 table, we may make use of the following formula (with Yates' correction built in):

$$\chi^2 = \frac{N(|AD - BC| - .5N)^2}{(A + B)\ (C + D)\ (A + C)\ (B + D)}$$

for the table

A	B	A + B
C	D	C + D
A + C	B + D	A + B + C + D = N

Note that the notation | | means 'take the absolute value, ignoring the sign'. For Connolly´s data:

$$\chi^2 = \frac{640\ (|236 \times 15 - 348 \times 41| - 640/2)^2}{584 \times 56 \times 277 \times 363}$$

$$= \underline{21.08}$$

The critical value of χ^2 for 1 d.f. (= (2 - 1) x (2 - 1)) is 10.83 at the $p \leqslant 0.001$ level. Since the observed value is higher than this, there is significant association between element order and text period at this level. Inspection of the data shows that O + P order is rarer in the later than in the earlier texts (15 x 100 / 363 = 4.1%, as against 41 x 100 / 277 = 14.8%).

CORRELATIONAL STUDIES

As part of a study of discourse development in profoundly deaf children, Prinz and Prinz (1985) measured the mean length of sign utterance (MLSU) and mean length of episode (MLE) for 24 such children whose ages ranged from 3 years 10 months to 11 years 5 months. A sign utterance was defined as 'a stretch of one child's communicative message bounded by another's message or by a pause of 1 second or more' (Prinz and Prinz 1985:11, fn.). An episode is 'an unbroken succession of relevant child utterances' (1985:11). The data from Table 5 of the Prinz and Prinz article are given in Table 8.

On the basis of this table, Prinz and Prinz (1985:12) comment '... individual differences in rate of psycholinguistic development occurred. However, there was a parallel increase in development in MLSU and MLE'. We can put this claim on a statistical basis by calculating correlation coefficients for the relationships between (a) MLSU and age in months, (b) MLE and age in months, and (c) MLSU and MLE. We shall discuss just the calculations for the correlation coefficient between MLSU and MLE.

Since the data are of the ratio type, the Pearson correlation coefficient (r) is appropriate. For the calculation of this coefficient, we need the values of x^2, y^2 and xy for each pair of values (x, y). These are shown in Table 9.

We now calculate the value of r as follows (N being the number of pairs of observations):

Table 8

Values of MLSU and MLE for 24 children of varying ages

Child	Chronological age	MLSU	MLE
1	3;10	2.2	2.3
2	4;3	3.8	2.5
3	4;9	4.4	4.5
4	5;2	3.7	3.8
5	5;6	5.5	5.4
6	5;8	6.9	5.1
7	5;9	7.2	6.6
8	5;11	7.3	6.8
9	6;5	6.2	7.3
10	6;10	7.1	8.2
11	6;11	7.3	7.9
12	7;1	8.2	9.3
13	7;3	6.6	10.7
14	8;2	6.8	9.9
15	8;3	7.2	10.9
16	8;10	7.4	8.8
17	9;2	8.1	11.3
18	9;5	8.2	9.9
19	9;10	7.9	12.1
20	10;1	8.1	13.2
21	10;6	8.2	10.8
22	10;8	8.4	14.1
23	11;5	8.4	15.3
24	11;5	8.2	16.0

Table 9

Values needed for calculation of Pearson correlation coefficient between MLSU and MLE

MLSU (x)	MLE (y)	x^2	y^2	xy
2.2	2.3	4.84	5.29	5.06
3.8	2.5	14.44	6.25	9.50
4.4	4.5	19.36	20.25	19.80
3.7	3.8	13.69	14.44	14.06
5.5	5.4	30.25	29.16	29.70
6.9	5.1	47.61	26.01	35.19
7.2	6.6	51.84	43.56	47.52
7.3	6.8	53.29	46.24	49.64
6.2	7.3	38.44	53.29	45.26
7.1	8.2	50.41	67.24	58.22
7.3	7.9	53.29	62.41	57.67
8.2	9.3	67.24	86.49	76.26
6.6	10.7	43.56	114.49	70.62
6.8	9.9	46.24	98.01	67.32
7.2	10.9	51.84	118.81	78.48
7.4	8.8	54.76	77.44	65.12
8.1	11.3	65.61	127.69	91.53
8.2	9.9	67.24	98.01	81.18
7.9	12.1	62.41	146.41	95.59
8.1	13.2	65.61	174.24	106.92
8.2	10.8	67.24	116.64	88.56
8.4	14.1	70.56	198.81	118.44
8.4	15.3	70.56	234.09	128.52
8.2	16.0	67.24	256.00	131.20
Σx = 163.3	Σy = 212.7	Σx^2 = 1177.57	Σy^2 = 2221.27	Σxy = 1571.36

$$r = \frac{N\Sigma xy - \Sigma x\Sigma y}{\sqrt{\{N\Sigma x^2 - (\Sigma x)^2\}\{N\Sigma y^2 - (\Sigma y)^2\}}}$$

$$= \frac{24 \times 1571.36 - 163.3 \times 212.7}{\sqrt{\{24 \times 1177.57 - (163.3)^2\}\{24 \times 2221.27 - (212.7)^2\}}}$$

$$= \frac{2978.73}{\sqrt{(1594.79 \times 8069.19)}}$$

$$= \underline{0.830}$$

The critical value in a directional test (since a positive correlation could be predicted) and at the $p \leqslant 0.005$ level, for 24 pairs, is 0.515. The correlation is thus significant at this level. The other relevant correlation coefficients are as follows:

Age in months / MLSU	0.818
Age in months / MLE	0.956

Both are significant at the $p \leqslant 0.005$ level.

MULTIVARIATE ANALYSIS

As an illustration of the use of two types of multivariate analysis, we shall discuss part of a project in which the author is currently engaged. The ultimate aim of the project is to develop a means of testing the validity of proposals made by people working in the framework of systemic linguistics, concerning the semantic choices open to language users. Such linguists construct 'networks' which aim to represent semantic difference or relatedness, and in recent years networks have appeared for meanings realized as verbs of physical change (Fawcett, 1980) and verbs concerned with accumulation and distribution (Hasan, 1987). It is with the latter set of items that we are concerned here.

Each of 11 native speakers of English was given a set of cards, on each of which was one of the following words: *accumulate, buy, collect, distribute, divide, gather, give, scatter, share, spill, strew*. They were asked to sort the cards into piles, as many or

Table 10

Similarity matrix for 11 words in a semantic field

Words	Accumulate	Buy	Collect	Distribute	Divide	Gather	Give	Scatter	Share	Spill	Strew
Accumulate											
Buy	1										
Collect	9	0									
Distribute	0	0	0								
Divide	0	0	0	6							
Gather	9	0	1	0	0						
Give	0	0	0	6	0	0					
Scatter	0	0	0	1	0	0	0				
Share	0	0	0	5	5	0	0	0			
Spill	0	0	0	0	0	0	0	7	0		
Strew	0	0	0	0	0	0	0	10	0	1	

as few as they wished, according to similarity in meaning, and then to put a rubber band round each pile. A 'pile' could consist of a single card. A table was then constructed showing, for each possible pair of words, how many informants had put that pair of words in the same pile. This <u>similarity matrix</u> for the pairs of words is shown in Table 10.

Two statistical techniques, <u>hierarchical cluster analysis</u> and <u>multidimensional scaling</u>, were applied in an attempt to discover structure in the meaning relationships between the items. As discussed in Chapter One, these are examples of <u>multivariate</u> techniques, in which a number of different variables are involved for each of a set of subjects (here, each word is rated for its similarity in meaning with respect to each of 10 other words). In this study, the MDS(X) package of programs, produced at the University of Edinburgh and University College Cardiff, was used to carry out the analyses of the similarity matrix.

```
CONNECTEDNESS METHOD

                 0 0 0 0 0 0 0 0 1 0 1
                 2 3 1 6 5 4 7 9 0 8 1

10.00000000      . . . . . . . . . XXX
 9.00000000      . . XXX . . . . . XXX
 9.00000000      . XXXXX . . . . . XXX
 7.00000000      . XXXXX . . . . XXXXX
 6.00000000      . XXXXX . XXX . XXXXX
 6.00000000      . XXXXX XXXXX . XXXXX
 5.00000000      . XXXXX XXXXXXX XXXXX
 1.00000000      . XXXXX XXXXXXXXXXXXX
 1.00000000      XXXXXXX XXXXXXXXXXXXX
 0.00000000      XXXXXXXXXXXXXXXXXXXXX

END OF METHOD
```

Fig. 2. Hierarchical clustering analysis of meaning for 11 words: connectedness method

```
DIAMETER METHOD

                 0 0 0 0 0 0 0 0 1 0 1
                 2 3 1 6 4 7 5 9 0 8 1

10.00000000      . . . . . . . . . XXX
 9.00000000      . . XXX . . . . . XXX
 6.00000000      . . XXX XXX . . . XXX
 5.00000000      . . XXX XXX XXX . XXX
 1.00000000      . . XXX XXX XXX XXXXX
 1.00000000      . XXXXX XXX XXX XXXXX
 0.00000000      . XXXXX XXX XXXXXXXXX
 0.00000000      . XXXXX XXXXXXXXXXXXX
 0.00000000      . XXXXXXXXXXXXXXXXXXX
 0.00000000      XXXXXXXXXXXXXXXXXXXXX

END OF METHOD
```

Fig. 3. Hierarchical clustering analysis of meaning for 11 words: diameter method

HIERARCHICAL CLUSTER ANALYSIS

The HICLUS option in the MDS(X) package produces a dendrogram (see Figs. 2 and 3), which displays the way in which the words cluster together. Looking towards the top of the dendrogram, we can see the tightest clusters, which then merge into looser clusters as we move down the diagram. The program offers two methods of clustering. In the 'connectedness' method, the dissimilarity between a point and a cluster is taken as the smallest of the dissimilarities between the point and the points in the cluster. This method tends to join points to existing clusters, and often gives results which are hard to interpret. The 'diameter' method takes the dissimilarity between a point and a cluster as the largest of the dissimilarities between the point and the points in the cluster. For data which the model fits perfectly, the two methods give the same results.

It can be seen from Figs. 2 and 3 that the two methods give quite similar results for our data. Both suggest that the items coded 1, 3 and 6 *(accumulate, collect, gather)* form a cluster, as do 8, 10 and 11 *(scatter, spill, strew)* and 4 and 7 *(distribute, give)*. Items 5 and 9 *(divide, share)* join the *distribute/give* cluster at a lower level, and 2 *(buy)* is weakly related to the *accumulate/collect/gather* cluster.

MULTIDIMENSIONAL SCALING

The MINISSA option in the MDS(x) package produces diagrams (see Fig. 4) which are a pictorial representation of the relationships in the data analysed.

The analysis can be carried out in 2, 3, or more dimensions (discussion of the most appropriate dimensionality for a given set of data is beyond the scope of this article); Figure 4 shows a 2-dimensional analysis. The results confirm those of cluster analysis to a large extent: 1, 2, 3 and 6 are reasonably close together, as are 4, 5, 7 and 9, as well as 8, 10 and 11.

In further work on this area, a larger group of informants will be used to group sets of lexical items, and the information given by the multivariate analyses will be compared with the groupings predicted by the semantic networks constructed by systemic linguists.

ACKNOWLEDGEMENTS

I am grateful to the following for permission to make use of published material: Professor M Lahey (Tables 2 and 3), Dr J Connolly (Tables 5, 6 and 7), Dr P M Prinz (Tables 8 and 9), Ablex Publishing Corporation (Tables 2, 3, 8 and 9).

MINISSA: RANKING: SIMILARITES
FINAL CONFIGURATION TASK NUMBER 1
DIMENSION 2 PLOTTED AGAINST DIMENSION

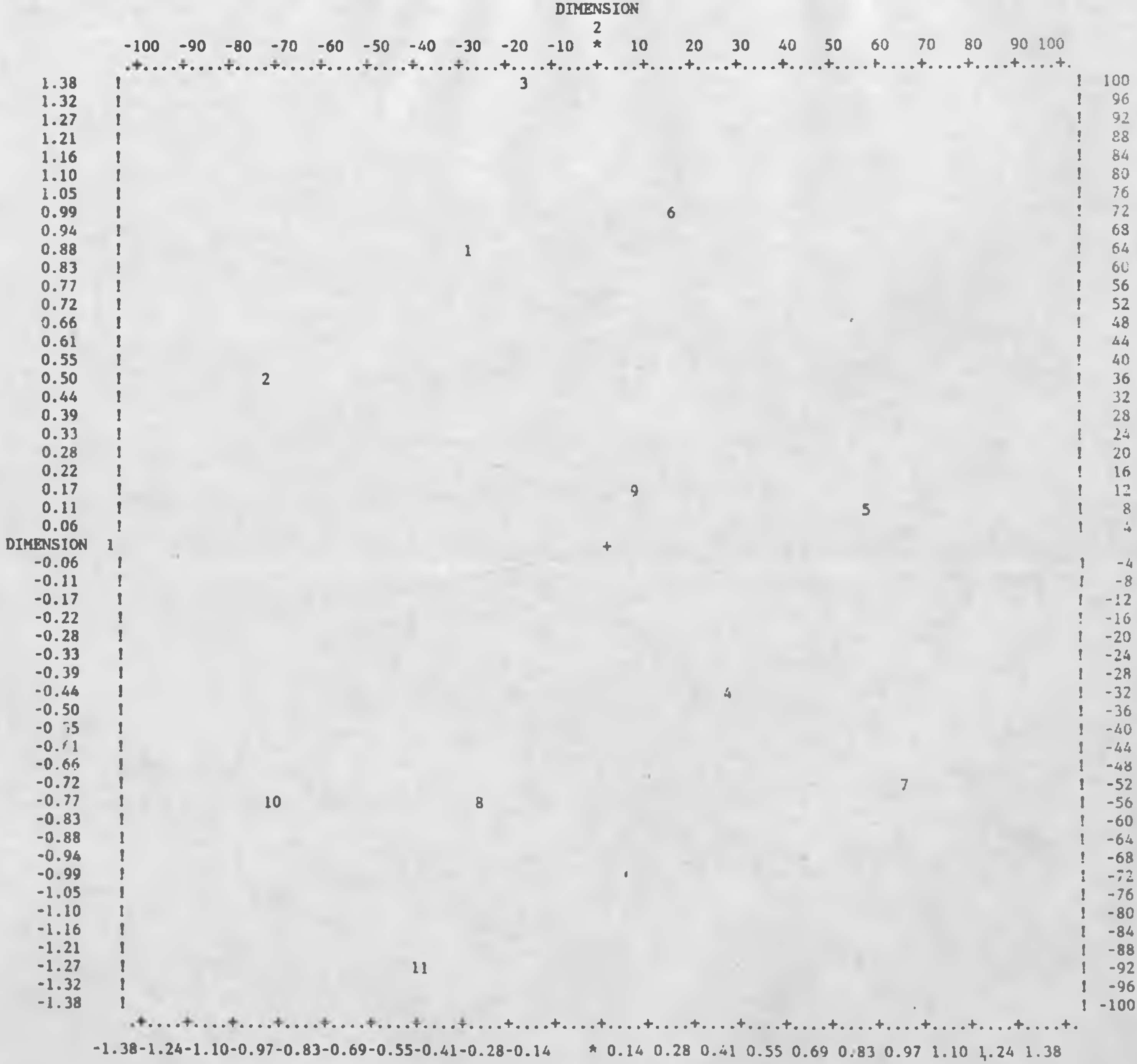

Fig. 4. Multidimensional scaling analysis of meaning for 11 words