CORRELATION METHODS OF COMPARING IDIOLECTS IN A TRANSITION AREA

David W. Reed
University of California

John L. Spicer
University of Minnesota

Introduction. The transition area as a crucial but difficult problem in American linguistic geography has been ably presented in a recent article by Alva L. Davis and Raven I. McDavid Jr.[1] The relative lack of pattern in the data from transition areas, compared to the easily mapped patterns found in relic and focal areas, is probably responsible in part for the slight attention that has been paid to transition areas before the appearance of this article. Davis and McDavid properly conclude that transition areas will assume increased importance as research in American linguistic geography moves away from the numerous focal and relic areas of the Atlantic coast toward the increasingly heterogeneous transition areas of the western United States. Certainly the collection and analysis of dialect vocabulary in California has revealed to the present writers a situation even more varied and complex than that of northwestern Ohio.

Davis and McDavid conclude that it will be necessary to collect much more complete information from transition areas if one is to correlate speech patterns with the historical and cultural complex. They suggest no new methods, however, by which the limited data now available from transition areas may be more adequately analyzed and understood. In the process of analyzing our own material from California, we have come to the conclusion that the speech patterns of transition areas grow much clearer when viewed as quantitative rather than as qualitative phenomena.

In areas of a high degree of uniformity, speech patterns can be adequately described in purely qualitative terms; but the quantitatively distributed data of transition areas require statistical methods of analysis. For example, if we wish to contrast the speech of two focal areas like the Boston area and the Virginia Piedmont, we may compile a list of features which serve as an index of difference between the two. Such a list might include the use of the terms (*hay*) *cock*, *cow yard*, and *pail* in the Boston area, contrasting with *shock*, *cow lot*, and *bucket*[2] in the Virginia Piedmont. On the other hand, if we wish to describe the differences between speech patterns in neighboring communities in a transition area like northwestern Ohio, we may find no such absolute qualitative differences, but merely differences in the degree of correspondence over a large number of items.

This paper is an attempt to apply the statistical method of correlation to the problem of ordering and establishing the degrees of relationship between the responses of the ten informants from northwestern Ohio. In a later paper we hope to describe statistical methods for clarifying the distribution of a single item in a transition area when the responses of a large number of informants are available.

[1] Northwestern Ohio: A transition area, Lg. 26.265–73 (1950).

[2] Cf. Hans Kurath, A word geography of the eastern United States 12–3, 39–40, 47–8, 54–6, figs. 5A, 42, 58, 61, 62, 66 (Ann Arbor, 1949).

Previous uses of the correlation method in linguistic analysis. Although correlation as a method of quantitative analysis had long been familiar in physical anthropology and ethnography, it was not until 1928 that the Polish anthropologist Jan Czekanowski transferred the method to the problem of differential diagnosis of the Indo-European languages.[3] In Czekanowski's study, nine branches of the Indo-European family are compared as to the presence or absence of twenty linguistic traits. Since Czekanowski's list of linguistic traits was so small that it did not guard sufficiently against the danger of statistical errors, Kroeber and Chrétien[4] applied the method to a larger sample of seventy-four linguistic traits selected at random by Kroeber from Meillet's Les dialectes indo-européens. While the results attained in this study were promising, and corresponded closely to the conventional internal classification of the Indo-European languages, a later attempt by the same two scholars to include Hittite in their scheme of classification[5] produced results which cast considerable doubt on the reliability of the method. (For example, the results indicated that Hittite was more closely related to the various Indo-European languages than several of them were to each other.)

In the course of their second article, Kroeber and Chrétien advanced a tentative explanation for these curious findings. Since the list of seventy-four linguistic traits was chosen with the Indo-European languages in view, it was 'not sufficiently pertinent to Hittite'.[6] Not fully satisfied with this explanation, Chrétien published two further studies interpreting the earlier results and testing their significance.[7] He came to the probably overcautious conclusion[8] that 'the sample method, whereby we choose a random selection of elements out of the total statistical population, is not likely to be satisfactory ... It seems to me that the answer is to employ the entire population.'

The chi-square test of reliability which Chrétien employed, however, would seem merely to indicate that the sample of seventy-four linguistic items was too small to afford reliable results and establish an order of relationships for comparisons from which the resulting correlation coefficient fell between +.35 and −.35. Outside of this zone, the ordering of relationships was still significant. In other words, in the application of correlation statistics to linguistic data, the investigator must employ some test of significance to safeguard his conclusions.[9]

[3] Jan Czekanowski, Na marginesie recenzji P. K. Moszyńskiego o książce: Wstęp do historji Słowian [On P. K. M.'s marginal criticism of the book: Introduction to the history of the Slavs], Series 2, Vol. 7 (Lud, 1928; reprinted Lwow, 1928). Title cited from Kroeber and Chrétien (op.cit. in fn. 4), as corrected by Chrétien.

[4] A. L. Kroeber and C. D. Chrétien, Quantitative classification of the Indo-European languages, Lg. 13.83–103 (1937).

[5] Kroeber and Chrétien, The statistical technique and Hittite, Lg. 15.69–71 (1939).

[6] Id. 70.

[7] The quantitative method for determining linguistic relationships, Univ. Cal. pub. in linguistics 1.11–20 (1943); Culture element distributions: 25. Reliability of statistical procedures and results, Anthropological records 8.469–90 (1945).

[8] Quant. method 19. Chrétien now agrees that this conclusion is overcautious, and suggests that the nature of the material may occasionally permit rather small samples.

[9] Chrétien's preference for the entire statistical population rather than a random sample is justified only if we wish to determine the absolute value of particular relationships in

In summary, the results of previous efforts to apply correlation statistics to the analysis of linguistic data would seem to indicate that the method is promising, provided certain cautions are observed: (1) the sample of linguistic items must be random in terms of the problem to be solved, (2) the sample must be large enough to afford reliable results, and (3) the correlations obtained must be submitted to some test of significance before conclusions are drawn. If these cautions are observed, the application of correlation statistics to data from linguistic geography ought especially to yield meaningful results, because in this field we are dealing with actual micro-units—the responses of individual informants—rather than with hypothetical macro-units like entire languages or branches of language families.

The statistical method of correlation. Correlation is the study of simultaneous variation of two or more variates.[10] The degree of similarity between the variates is expressed by a correlation coefficient which may range from +1.00 for perfect identity to −1.00 for complete dissimilarity. One of the simpler methods of calculating correlation coefficients employs the formula for Q_6 which was used by Kroeber and Chrétien.[11] To calculate the value of Q_6, one counts (a) the number of elements common to both groups of variates, (b) the number present in the first but absent in the second, (c) the number absent in the first but present in the second, and (d) the number absent in both.[12] Substituting these numbers in the following formula, one calculates the value of V:[13]

$$V = \frac{ad - bc}{\sqrt{(a + b)(c + d)(a + c)(b + d)}}$$

To give the values of the resulting coefficients a more normal distribution, V is substituted in this further formula:

$$Q_6 = \sin\left[\frac{\pi}{2} \cdot V\right]$$

order to compare them ultimately with other relationships outside the immediate field of investigation. For practical purposes, the linguist wishes simply to order whatever is being related in some fashion that will be meaningful in the immediate context. For us the significant conclusion to be drawn from Chrétien's use of the chi-square test is that in any ordering of data there will be a dead zone around the zero coefficient—a zone within which the ordering is not reliable. The larger the sample, the smaller this dead zone.

[10] For general though difficult discussions of the theory and methods or correlation, see R. A. Fisher, Statistical methods for research workers[6] (Edinburgh, 1936); G. U. Yule and M. G. Kendall, An introduction to the theory of statistics[12] (London, 1940).

[11] For a more complete description of method than can be given here, see Kroeber and Chrétien, Quant. classif. 83–5, 98–103; Chrétien, Quant. method 12–6.

[12] That is, the number of elements FROM THE LIST UNDER CONSIDERATION that are absent from both groups of variates. Another common formula (G) assumes that d is always infinite. Values of V (see below in the text) approach corresponding values of G as an upper limit. The Q_6 formula has been favored here because it gives twice as wide a spread of values as the G formula. The relative values of Q_6 and G do not differ significantly except when d is very small.

[13] V is identical with Kroeber and Chrétien's r_{hk}. V is the symbol more recently employed by statisticians for this value.

The values of Q_6 obtained by the application of these formulas may be tested for significance by a further statistical device known as the chi-square test.[14] Starting with the value of V obtained above, one applies the following equation, where N represents the total number of items: $\chi^2 = NV^2$. One then refers to an appropriate table of values of χ^2 to get a quantity known as P, or probability.[15] The values of P vary from 100 percent for certainty to a figure approaching but never reaching zero. The percentage figure indicates the proportion of times that a correlation at least as far from zero as the one actually obtained will occur on a basis of pure chance (i.e. when no observable factor influences the events other than the complex and little-understood series of causations known as chance). By convention, statisticians agree that P must be no greater than 5 percent and may preferably be held to values as low as 1 percent or even 0.1 percent if the correlation figure is to be considered significant.[16]

The northwestern Ohio material. Applying the method of correlation to the northwestern Ohio material is a relatively simple operation. Davis and McDavid presented three tables showing the distribution respectively of thirty-nine items of vocabulary, ten items of pronunciation, and seven items of grammar, among ten informants—two each in five communities in northwestern Ohio. Most of these items have several variants; only the pronunciation table was not organized directly to show simple presence or absence of all variants in each item. For statistical purposes, each variant of an item in the vocabulary and grammar tables was considered to be a separate item, the tables thus yielding ninety-one vocabulary items and fourteen grammar items. The pronunciation table was reorganized to indicate simple presence or absence of each pronunciation characteristic, with a resultant total of forty-three pronunciation items.

After this preliminary organization of the Davis–McDavid material, we subjected it to the statistical analyses described above.

General quantitative relationships among the speech patterns of the ten Ohio informants. Informants are listed by abbreviations in the row at the top and the column at the left of Table 1.[17] The Q_6 figures in the body of the table give the correlation between the speech patterns of the two informants whose column and row intersect at that point. The order of listing informants in the columns and rows corresponds roughly to the degree of relationship between their various speech patterns. The column at the right and the row at the bottom of the table show the arithmetical average of the nine correlations of the speech of each informant with the speech of every other informant. It will be observed that these mean figures describe a curve, ascending from 0.13 for P_1 to 0.48 for VW_1 and US_2, and descending to 0.28 for D_1.

[14] For a readily understandable account of this device, see Chrétien, Quant. method 17–9.

[15] E.g. Yule and Kendall 534–5.

[16] Fisher 77, 128–33, 230–8.

[17] P = Perrysburg, D = Defiance, O = Ottawa, VW = Van Wert, US = Upper Sandusky. In each community the informant designated by a subscript 1 (P_1 etc.) is the more old-fashioned, i.e. has had fewer contacts with speakers from outside his community than the other informant. For a thumbnail sketch of all the informants, see Davis and McDavid 266 fn. 9.

One of the most important features of this table is indicated by the solid and broken lines separating various sections of the Q_6 figures. The two sets of solid lines enclose correlation figures of 0.58 or more. The two sets of broken lines, with the exception of the correlation figure 0.28 for US_2–P_2, parenthesized in the table, enclose correlations of 0.22 or less. The correlation figures not enclosed by either solid or broken lines range from 0.23 to 0.52.

The results of the chi-square test of significance are summarized in the legend below the chart. These numbers should be interpreted as follows: A correlation figure of ±0.25 or more will occur in five percent or less of pure chance comparisons based on 148 items. Correlations of ±0.32 and ±0.43 or more will occur respectively in 1 percent and 0.1 percent or less of such comparisons. Correlations nearer zero than ±0.25 in Table 1 are therefore of doubtful statistical significance.

Informant	P_1	P_2	D_2	O_1	O_2	VW_1	US_2	VW_2	US_1	D_1	Mean
P_1	X	0.64	0.31	0.19	0.06	−.03	0.08	0.13	−.11	−.11	0.13
P_2	0.64	X	0.28	0.06	0.22	0.13	(0.28)	−.05	−.13	−.02	0.16
D_2	0.31	0.28	X	0.51	0.35	0.31	0.32	0.25	0.26	0.23	0.31
O_1	0.19	0.06	0.51	X	0.40	0.52	0.41	0.45	0.43	0.40	0.37
O_2	0.06	0.22	0.35	0.40	X	0.67	0.81	0.58	0.64	0.35	0.45
VW_1	−.03	0.13	0.31	0.52	0.67	X	0.77	0.80	0.72	0.45	0.48
US_2	0.08	(0.28)	0.32	0.41	0.81	0.77	X	0.60	0.63	0.41	0.48
VW_2	0.13	−.05	0.25	0.45	0.58	0.80	0.60	X	0.58	0.32	0.41
US_1	−.11	−.13	0.26	0.43	0.64	0.72	0.63	0.58	X	0.50	0.39
D_1	−.11	−.02	0.23	0.40	0.35	0.45	0.41	0.32	0.50	X	0.28
Mean	0.13	0.16	0.31	0.37	0.45	0.48	0.48	0.41	0.39	0.28	0.35

TABLE 1

Q_6 Correlation Coefficients for Ten Informants in Five Northwest Ohio Communities
Values of Q_6 at various levels of significance:
5%: 0.25; 1%: 0.32; 0.1%: 0.43

Two important points emerge from the analyses presented in Table 1. First, as Davis and McDavid have pointed out after a qualitative inspection of the data, there are two relatively homogeneous groups of informants: P_1 and P_2, whose predominant characteristics according to Davis and McDavid are Northern, and a group of five including O_2, VW_1, VW_2, US_1, and US_2, described as predominantly Midland. D_1, D_2, and O_1 do not belong clearly to either of these groups.

Our statistical analysis of the responses of these last three informants seems to provide a basis for more definite conclusions than Davis and McDavid were able to reach. O_1 corresponds somewhat to the Midland group of informants, but is quite as closely related to D_2.[18] D_2, on the other hand, stands approximately

[18] The difference between O_1 and O_2 is partially explained by the fact that the field record for O_1 was made by Frederick G. Cassidy in 1939, while all the other records were made in June 1949 by McDavid.

midway between the Northern and Midland groups, but is most closely related to O_1. In other words, D_2 and O_1 are genuine transition informants between two more or less well defined groups within the transition area.

The position of D_1 requires special comment. Whereas the correlation between two informants from the same community has generally been high—Van Wert 0.80, Perrysburg 0.64, Upper Sandusky 0.63, Ottawa 0.40 (see note 18)—that between the two informants from Defiance is only 0.23, a figure below our minimum level of significance. Further, D_1 shows a negative correlation with each of the Northern informants but a fairly high positive correlation with the Midland informants. In view of our data from D_2, we should expect D_1 also to be a transitional informant between our two somewhat clearly defined groups; but if D_1 is a transitional informant, he must be so only in a very special sense.[19]

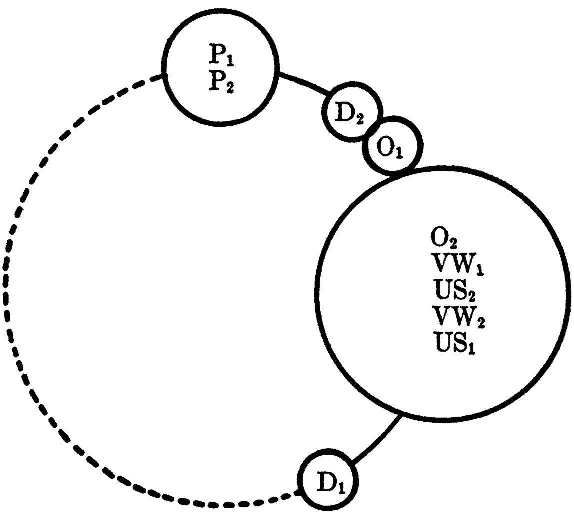

TABLE 2

Circle Diagram of Quantitative Relationships of the Ten Idiolects

If the entire series of possible relationships between all speakers in this transition area were represented diagramatically by a circle, the relative position of our ten informants on the circle would probably be that shown in Table 2.

Northern-Midland relationships among the Ohio informants. The dotted portion of the circle in Table 2 represents a hypothetical relationship, hinted at however by certain further facts. Obviously an important factor in low correlation figures is the presence of responses by one informant which are unique in the sample represented or which are shared with only a minority of other informants. We shall want to consider such responses in the light of the Northern-Midland distinction which Davis and McDavid find between the two major groups of informants. Table 3 presents a summary of those unique and minority re-

[19] McDavid informs us that D_1 was the most isolated rural informant interviewed in northwestern Ohio, and by far the oldest.

sponses which are definitely Northern or Midland in the eastern United States or which are blends of Northern and Midland forms.[20]

Table 3 shows that in minority forms O_2, VW_1, US_2, VW_2, and US_1—those informants who have intercorrelations greater than 0.58—have a majority of the Midland responses. It will be observed that in minority forms D_1 has a higher proportion of Northern than of Midland forms, and is in this respect more like P_1, P_2, D_2, and O_1 than he is like the first group of informants, with whom, strangely enough, he shows higher correlations in Table 1. Further evidence of the transitional nature of D_1 is to be seen in his relatively large number of blends, which are totally absent from the responses of the two almost purely Northern informants (P_1 and P_2) and of the two almost purely Midland informants (VW_1 and US_2). Additional confirmation of the paradoxical nature of D_1 lies in the fact that he stands in a first-place tie with VW_1 for the largest number of unique

Source	Northern				Midland				Blends				Totals		
Shared With	0	1	2	3	0	1	2	3	0	1	2	3	N	M	B
Informant															
P_1	2	7	4	7	0	0	1	0	0	0	0	0	20	1	0
P_2	3	5	4	6	0	0	1	0	0	0	0	0	18	1	0
D_2	3	0	2	5	0	1	2	1	0	0	0	0	10	4	0
O_1	1	1	1	5	0	1	2	1	0	0	1	0	8	4	1
O_2	0	0	1	1	0	1	0	4	0	0	1	0	2	5	1
VW_1	0	0	0	1	2	1	1	4	0	0	0	0	1	8	0
US_2	0	0	0	2	1	2	1	4	0	0	0	0	2	8	0
VW_2	0	0	0	3	1	0	2	4	0	0	1	0	3	7	1
US_1	0	0	1	2	0	0	1	4	1	0	1	0	3	5	2
D_1	0	1	2	8	2	0	1	2	1	0	2	0	11	5	3

TABLE 3

Northern and Midland Responses and Blends
Shared by a Minority of the Ten Informants

Midland responses but has no unique Northern responses at all. This is in sharp contrast to D_2, who is tied for first place with P_2 for the largest number of unique Northern responses, but has no unique Midland responses. Table 4 shows by means of a graded circle diagram the relative proportions of Northern and Midland responses among the ten informants, blends being counted half Northern and half Midland.

Vocabulary, pronunciation, and grammar relationships among the Ohio informants. Since the Davis–McDavid list contains items from vocabulary, pro-

[20] Certain terms in the lists are labeled Northern or Midland by Davis and McDavid. We have added labels to the following, on the basis of indications in Kurath's Word geography (pages here cited): *beller*, New England (19, 62); *sugar grove*, w. Pa., n. Va. (36, 76); *swingletree*, Delaware River, s. N. J. (13, 47, 58); *sook*, general Midlands w. of the Susquehanna (14, 30, 38, 63); *sick in his stomach*, s. Pa. (78); *ground worm*, Chesapeake Bay and e. Pa. (46, 74).

nunciation, and grammar, it was felt that it might prove instructive to break down the correlation figures into the corresponding three components. Table 5 provides such a breakdown. It can best be read in connection with Table 1.

The reliability figures in the legend of this table are of great importance to its interpretation. Thus, in the small sample of fourteen grammar items, a correlation must attain ± 0.73 to be deemed significant. Only four of the grammar correlations do in fact attain this level. In the main, therefore, we shall be justified in ignoring the grammar correlations in this table as based on insufficient evidence. The correlation figures for vocabulary and pronunciation, on the other hand, are of considerably greater significance. Thirty of the pronunciation

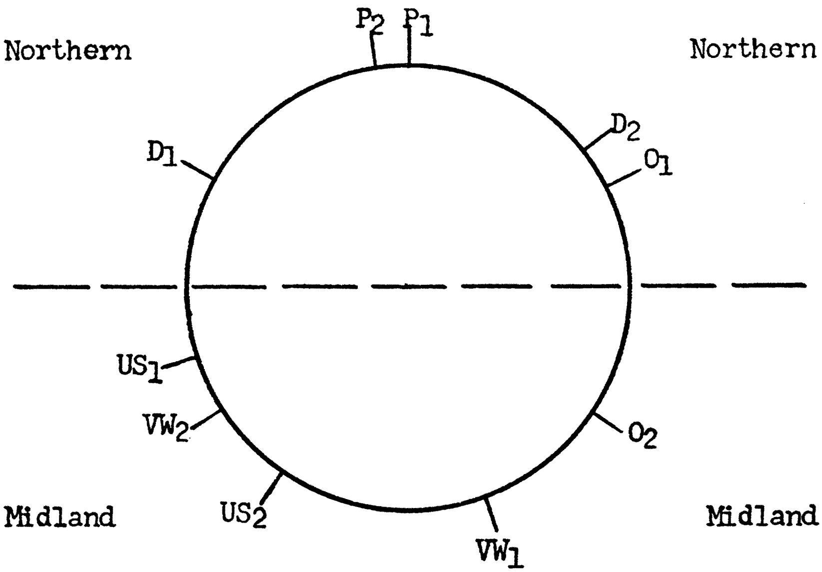

TABLE 4
Circle Diagram of North-Midland Relationships of the Ten Idiolects

correlations and 24 of the vocabulary correlations attain the 5% level of significance. It will be noted that the pronunciation correlations, and especially their means, are generally higher than the vocabulary correlations. This indicates that pronunciation in the area is more homogeneous,[21] while vocabulary shows a greater range of difference.

The position of our troublesome informant D_1 is worth investigating in Table 5. Although to a great extent the D_1 correlations here are what we should expect as

[21] As A. H. Marckwardt justly observes, however, this indication is perhaps illusory, since Davis and McDavid greatly simplified the phonetic transcriptions of the field records before constructing their pronunciation table. If they had taken into account the full range of symbols and diacritics used in the field records, pronunciation would probably seem less homogeneous than vocabulary. Phonemic analysis of the speech of each informant would provide the only satisfactory approach to a comparison of pronunciation and vocabulary on the score of homogeneity.

Informant		P_1	P_2	D_2	O_1	O_2	VW_1	US_2	VW_2	US_1	D_1	Mean
	v		0.58	0.28	0.06	−.28	−.22	−.16	−.11	−.29	−.05	−.02
P_1	p	X	0.67	0.48	0.40	0.51	0.33	0.43	0.56	0.05	−.14	0.38
	g		0.85	−.22	−.22	0.60	−.14	0.48	0.00	0.00	−.84	0.06
	v	0.58		0.45	0.17	0.11	0.05	0.20	−.19	−.17	0.14	0.15
P_2	p	0.67	X	−.13	−.05	0.41	0.17	0.51	0.51	0.05	−.23	0.21
	g	0.85		0.29	−.37	0.14	0.29	0.00	−.48	−.48	−.48	−.12
	v	0.28	0.45		0.37	0.20	0.29	0.35	0.16	0.20	0.28	0.29
D_2	p	0.48	−.13	X	0.70	0.71	0.52	0.45	0.79	0.52	−.03	0.45
	g	−.22	0.29		0.56	−.06	−.06	−.23	0.89	−.23	0.65	0.18
	v	0.06	0.17	0.37		0.38	0.51	0.16	0.32	0.44	0.45	0.32
O_1	p	0.40	−.05	0.70	X	0.40	0.70	0.52	0.84	0.68	0.32	0.50
	g	−.22	−.37	0.56		0.40	0.06	0.23	−.23	−.65	0.23	0.00
	v	−.28	0.11	0.20	0.38		0.76	0.80	0.56	0.65	0.35	0.39
O_2	p	0.51	0.41	0.71	0.40	X	0.59	0.76	0.76	0.70	0.58	0.60
	g	0.60	0.14	−.06	0.40		0.44	0.98	−.22	0.22	−.63	0.21
	v	−.22	0.05	0.29	0.51	0.76		0.84	0.78	0.87	0.53	0.49
VW_1	p	0.33	0.17	0.52	0.70	0.59	X	0.68	0.87	0.52	0.26	0.52
	g	−.14	0.29	−.06	0.06	0.26		0.44	0.63	0.00	0.44	0.21
	v	−.16	0.20	0.35	0.16	0.80	0.84		0.51	0.65	0.33	0.41
US_2	p	0.43	0.51	0.45	0.52	0.76	0.68	X	0.89	0.70	0.72	0.63
	g	0.48	0.00	−.23	0.23	0.98	0.44		−.22	0.22	−.63	0.14
	v	−.11	−.19	0.16	0.32	0.56	0.78	0.51		0.67	0.32	0.34
VW_2	p	0.56	0.51	0.79	0.84	0.76	0.87	0.89	X	0.52	0.26	0.67
	g	0.00	−.48	0.89	−.23	0.00	0.63	−.22		0.22	−.22	0.07
	v	−.29	−.17	0.20	0.44	0.65	0.87	0.65	0.67		0.47	0.39
US_1	p	0.17	0.05	0.52	0.68	0.70	0.52	0.70	0.47	X	0.64	0.49
	g	0.00	−.48	−.23	−.65	0.44	0.00	0.22	0.22		0.22	−.03
	v	−.05	0.14	0.28	0.45	0.35	0.53	0.33	0.32	0.47		0.31
D_1	p	−.14	−.23	−.03	0.32	0.58	0.26	0.72	0.48	0.64	X	0.29
	g	−.84	−.48	0.65	0.23	0.44	0.44	−.63	−.22	0.22		−.02
	v	−.02	0.15	0.29	0.32	0.39	0.49	0.41	0.34	0.39	0.31	0.31
Mean	p	0.38	0.21	0.45	0.50	0.60	0.52	0.63	0.67	0.49	0.29	0.47
	g	0.06	−.12	0.18	0.00	0.21	0.21	0.14	0.07	−.03	−.02	0.07

TABLE 5

Vocabulary, Pronunciation, and Grammar Components of the Q_6 Coefficients

Values of Q_6 at various levels of significance:

	5%	1%	0.1%
v	0.32	0.41	0.52
p	0.45	0.58	0.74
g	0.73	0.88	0.97

components of the correlations in Table 1, there is one interesting difference. The pronunciation correlations between D_1 and the Northern group—P_1, P_2, and D_2,—are all less than the vocabulary correlations between the same informants, and violate the pattern set in the rest of the table. These negative correlations constitute three of the five negative pronunciation correlations on the chart. (The other two are also for the comparison of a Northern informant and a transitional informant, P_2–D_2 and P_2–O_1.)

Cartographic representation of the quantitative relationships between northwestern Ohio communities. Since linguists have found that cartographic representation is one of the most effective ways of presenting dialect phenomena, the question naturally arises whether quantitative results of the sort obtained here can be mapped in any meaningful way. Can we adapt the qualitative method of the isogloss for single items to represent the quantitative distribution of large numbers of items? Spicer had arrived independently at a method which seemed likely to achieve these results when it was called to our attention that the German anthropologist Wilhelm Milke had successfully applied the same method to indicate the distribution of cultural similarities.[22]

The first step in preparing to map the quantitative distribution of large numbers of items is to obtain correlation coefficients for the comparisons of whole communities rather than for the comparisons of individual informants. This can be done for the northwestern Ohio material only on the assumption that the two informants in each community—since after all they were chosen largely on the basis of their family background—are adequate representatives of the community speech pattern as a whole. This assumption underlies all the work of a linguistic atlas, but it is obviously a dangerous one; a greater number of informants per community would yield much more reliable results in a quantitative study. For optimum accuracy, it would be further necessary to include both a greater number of communities and a greater number of linguistic items—perhaps all the items in the normal questionnaire of the Linguistic Atlas. Nevertheless, the assumption of representativeness is justifiable for the purpose of illustrating the cartographic method.

Once the initial assumption has been made, we add together the values of a, b, c, and d for the four individual comparisons between any two communities (e.g. P_1–D_1, P_1–D_2, P_2–D_1, P_2–D_2); a single Q_6 correlation between the two communities can be calculated from the totals. Then, taking any one community as a point of reference (1.00), we enter on a map at the appropriate points the correlation coefficients between the point of reference and each of the other communities. Finally we superimpose a scale of isogrades[23] on the communities.

[22] The quantitative distribution of cultural similarities and their cartographic representation, transl. by Chrétien and Kroeber, American anthropologist 51.237–52 (1949).

[23] Milke's term for these lines is *isopleth*, a word used with a slightly different meaning in meteorology. Because of the difference in meaning and because the phrase *graded area* is well established in dialect geography, we have preferred to coin the somewhat barbarous *isograde*. (The only pure Greek alternatives are *isotax* and *isobath*. The former carries an irrelevant suggestion of *syntax* and *taxeme*, while the latter seems to have been used as the name of a patented floating inkstand.)

Figure 1 consists of five maps, taking each community successively as the point of reference. By the use of isogrades we can express in terms of a continuous distribution the similarities which are directly expressible only as points on the map. Milke designates this distribution as the 'potential of similarity.'[24] Although

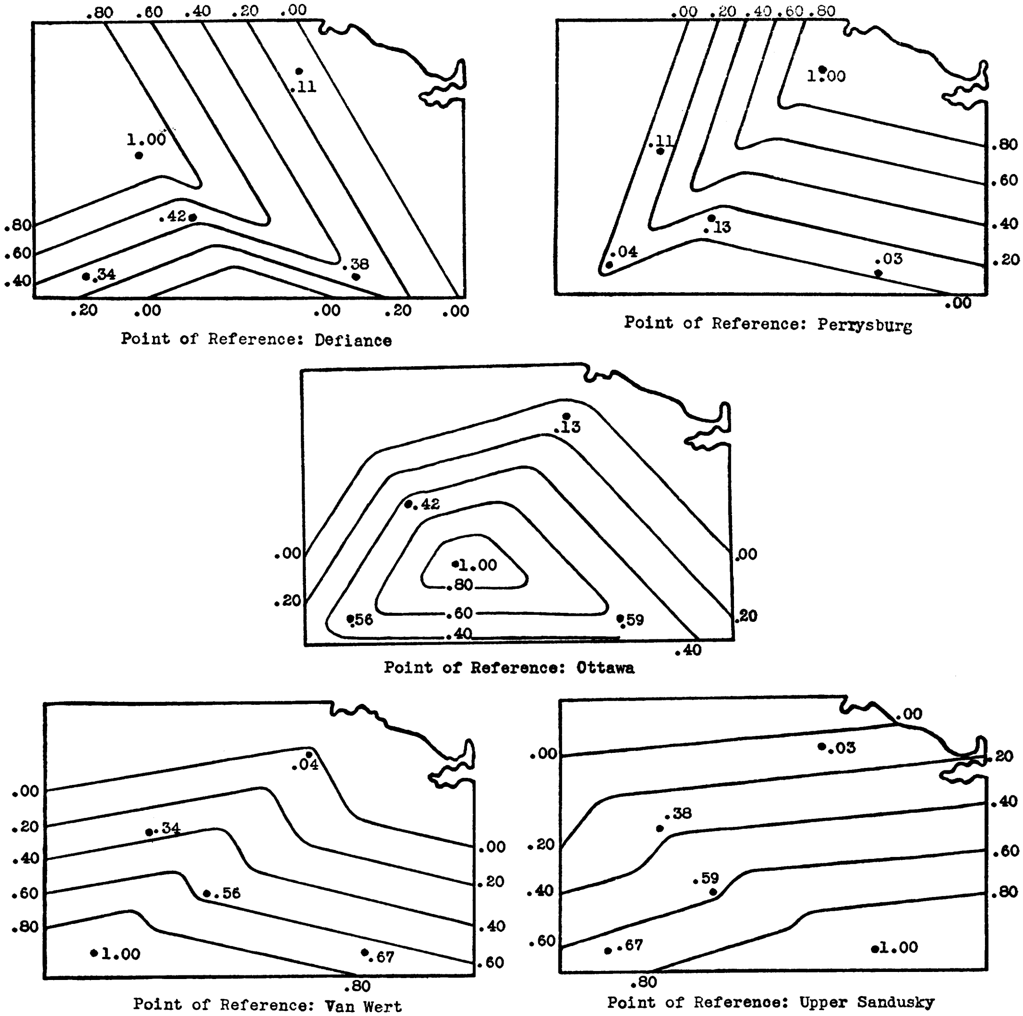

FIGURE 1. Northwest Ohio: Isogrades of Linguistic Similarity (Coefficient Q_6)

the patterns of isogrades differ considerably according to the community selected as the point of reference, all the maps in Figure 1 tend to show the north-south cleavage in dialect patterns already observed in the qualitative and statistical analyses. Wide-spaced isogrades, of course, indicate gradual transitions, while narrow-spaced lines indicate more abrupt changes.

[24] Cf. Milke 248.

Summary of results. Our correlation statistics have shown that there are two rather clearly defined groups of informants—one predominantly Midland, the other predominantly Northern—and three genuine transitional informants (D_1, D_2, and O_1). Of these three informants, D_2 and O_1 seem to represent a transitional stage normal to the body of data presented by Davis and McDavid. D_2 stands approximately halfway between the Northern and the Midland groups, O_1 halfway between D_2 and the Midland group (Tables 1 and 2). D_2 and O_1 are the only informants to give unique Northern responses, except the 'Northern' informants themselves, P_1 and P_2 (Table 3). D_1 stands out immediately in the data of Table 1 because of his negative correlation with P_1 and P_2 and his relatively low correlation with D_2, the other informant from the same community. Although D_1 behaves as a normal transition informant in showing approximately the same proportion of Northern to Midland minority forms as do D_2 and O_1, he is tied for first place with VW_1 for the largest number of unique Midland responses, and unlike D_2 and O_1 has no unique Northern responses (Table 3). D_1 is also characterized by the largest total number of blends, a fact of some interest in view of the sharply defined layers of Northern and Midland responses from this informant. Finally, D_1 differs from the Northern informants more sharply in pronunciation than in vocabulary, although this is contrary to the general tendency among all informants (Table 5). The differences between D_2 and O_1 on the one hand and D_1 on the other can be most clearly made graphic by circle diagrams like Tables 2 and 4. Cartographic representation of the quantitative differences between communities, by means of isogrades, is a valuable auxiliary device and tends to support our conclusions as to the north-south differences in these dialect patterns (Figure 1).

Conclusion. In the present article we have illustrated some of the methods we intend to use in dealing with the larger body of California material. These methods can all be applied to establish the degree of relationship between whole idiolects or the speech patterns of whole communities. The conventional Linguistic Atlas method of compiling field records of the responses of two or three informants in a community, apparently insufficient for the adequate study of transitional communities like Defiance, Ohio, is still less adequate in the study of so large and complex a transition area as California. But even the small statistical sample presented here has been shown capable of yielding some hints when subjected to proper quantitative analysis. Obviously a greater number of responses from a larger number of informants in more communities will be essential for securing answers to many of the questions posed by the transition area.