



Quantitative Classification of Indo-European Languages

A. L. Kroeber; C. D. Chrétien

Language, Vol. 13, No. 2. (Apr. - Jun., 1937), pp. 83-103.

Stable URL:

<http://links.jstor.org/sici?sici=0097-8507%28193704%2F06%2913%3A2%3C83%3AQCOIL%3E2.0.CO%3B2-4>

Language is currently published by Linguistic Society of America.

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at <http://www.jstor.org/about/terms.html>. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at <http://www.jstor.org/journals/lisa.html>.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

The JSTOR Archive is a trusted digital repository providing for long-term preservation and access to leading academic journals and scholarly literature from around the world. The Archive is supported by libraries, scholarly societies, publishers, and foundations. It is an initiative of JSTOR, a not-for-profit organization with a mission to help the scholarly community take advantage of advances in technology. For more information regarding JSTOR, please contact support@jstor.org.

QUANTITATIVE CLASSIFICATION OF INDO-EUROPEAN LANGUAGES

A. L. KROEBER

C. D. CHRÉTIEN

UNIVERSITY OF CALIFORNIA

1. Method

In 1928 the Polish anthropologist Jan Czekanowski published in the ethnographical quarterly *Lud*¹ a study of the Indo-European languages in which he employed the method of differential diagnosis by quantitative correlation determinations which he had long been using with success in physical anthropology and ethnography. This method, whatever its field, rests upon the recognition of isolable and definable features or traits, which we shall hereafter refer to as elements, whose presence or absence can be determined for a number of populational groups or territorial entities, such as races, tribes, cultures, castes, or, in the present study, languages. The distribution of these is tabulated in terms of plus for presence in a particular group, minus for absence, and the question mark for unknown. Then each group is compared with each of the other groups in terms of the four-cell segregation familiar to statisticians. That is to say, four values are determined: *a* represents the number of elements common to both groups, *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. In other words, *a* and *d* are agreements, positive and negative respectively; *b* and *c* are disagreements. These four values are then substituted in a suitable formula, and a coefficient of similarity between the two groups results. When the coefficients for each pair of the groups being considered are assembled, we get a classification of the relative degrees of

¹ Jan Czekanowski, *Na Marginesie Recenzji P. K. Moszyńskiego o Książce: Wstęp do Historji Słowian* [Marginal Criticism of P. K. Moszynski's Introduction to the History of the Slavs], *Lud*, Series II, vol. VII (1928). Reprint Lwow, 1928. For an application of the method to ethnography see Stanislaw Klimek, *The Structure of California Indian Culture, Culture Element Distributions: I* (University of California Publications in American Archaeology and Ethnology 37. 1-70 [1935]).

similarity between the populational groups or territorial entities which, being objective, has genetic and historical significance. For example, to make this clear in linguistic terms, suppose for the four languages Baltic, Slavic, Indic, and Iranian, we get a high coefficient between Baltic and Slavic, and again a high coefficient between Indic and Iranian, but low coefficients for the four other possible pairs, Baltic-Indic, Baltic-Iranian, Slavic-Indic, and Slavic-Iranian. The coefficients thus make the four languages fall into two classes, Baltic-Slavic, and Indic-Iranian, and it is evident that each class has had a certain history common to its members but not shared by members of the other class.

When all the coefficients for the groups under consideration have been computed, it is usually clarifying to put them into a tabular form in which they are arranged as nearly as possible in the sequence of their values. This arrangement tends to concentrate high values along a diagonal of the table, the lowest tending to fall away from the diagonal into the corners, unless the main relationships are multiple or polygonal instead of linear. For vivid effect the table can then be converted into a graphic diagram in which certain ranges of coefficients are expressed by gradation of symbols. For instance, all values between 1.00 and .90 can be denoted by solid black squares, those between .90 and .80 by mostly black, those between .80 and .70 by half black, and so on. If the symbol values are chosen judiciously, the diagram becomes an exceedingly effective and rapidly grasped representation of the stronger relationships, wherein the salient features of the classification force themselves upon the eye and the mind through the automatic clustering of the symbols.

There are several formulas available for computing coefficients, each possessing particular theoretical or practical advantages; but experience in anthropology and ethnography has shown that ordinarily results are not vitally affected by the choice of formula. If in a given study we compute coefficients first by one formula and then by another (as we have done in this paper) we get different absolute values, but the relative rankings of the populational groups or territorial entities tend to come out surprisingly alike, especially for the more significant highest and lowest values. We need not therefore go further into the matter of formulas here; the subject is discussed more in detail in Part 4 of this paper. What is crucial in investigations by this method is authenticity of data on a sufficient number of groups, and sharp and accurate definition of the elements involved. Ideally we should cover all the data; practically this is impossible. The principles of statistics tell us how-

ever that a genuinely random selection of a sufficient number of elements will give us the same or approximately the same results as the complete assemblage. The present study is based on a random selection of 74 elements.

The method is not exhausted with the determination of degrees of likeness of populational groups or territorial entities. Examination can be directed to degree of 'adhesion', that is to say, of co-occurrence or association, of elements themselves. Concretely, the frequency with which an optative and a dual co-exist or fail to co-exist in a given series of languages can be determined as well as the similarity of Slavic to Latin or to Sanskrit, and from the same data. The process of tabulation and counting is simply reversed: one counts in how many languages an optative and a dual co-occur instead of counting how many elements co-occur in Slavic and Latin. The findings would express what we might call linguistic types within Indo-European instead of classes of Indo-European languages. A third step is the intercorrelation of the two sets of findings. This would express the relative participation of the several classes of languages in the several linguistic types. Only the first process, the determination of classes of Indo-European languages, will be applied in this paper.

No claim can be made that this quantitative method will yield interpretations of a different order or kind from those already made in Indo-European linguistics by non-statistical methods. But these latter are, in part at least, applications of insight; and, being subjective, the best insight may sooner or later overshoot its mark. What statistical analysis can do is to validate and correct insight, or, where insight judgments are in conflict, help to decide between them. In short, it increases objectivity, sharpens findings, and sometimes forces new problems.

2. Previous Results

Czekanowski² investigated the relationships of nine Indo-European languages: Lithuanian, Old Church Slavic, Gothic, Old Irish, Latin, Greek, Vedic, Avestan, and Armenian. He employed twenty-two

² *Op. cit.* For other studies by Czekanowski applying the method to Slavic dialects see *Z Badań nad Zrózniczkowaniem Morfologicznem Dialektów Polskich* [Investigation of Morphological Differentiation of Polish Dialects], *Prac Polonistycznych* (Warsaw: 1927); *Różnicowanie się Dialektów Prastowiańskich w Świetle Kryterjum Ilościowego* [Differentiation of Ancient Slavic Dialects in the Light of Quantitative Criteria], *Sborník Pracé, I. Sjezdu Slovanských Filologu v Praze 1929* (First Congress of Slavic Philologists in Prague, 1929), Prague, 1931.

elements, which we give here, without comment, as translated from his list:

1. Surd aspirates. 2. Augment. 3. *mē*. 4. Verb-ending in *r*. 5. Conjunctive. 6. 1000 = *gheslo*. 7. *-bhis*, etc. 8. Internal *a*. 9. Reduplicated perfect. 10. Sigmatic aorist. 11. *ō* = *ā*. 12. *-mis*, etc. 13. 1000 = *teya*. 14. *ō* = *ā*. 15. Optative. 16. Dual in verbs. 17. Relative pronoun *ye/o*. 18. Dual in nouns. 19. Spirantization of *k̃*, *g̃*, (*h*). 20. *s* becomes *š* after *i*, *u*, *k*, *r*. 21. *gʷ*, *kʷ*, *gʷh* become *g*, *k*, *gh*. 22. *t*, *d* + *t* become *ss*.

TABLES I AND II

I: Czekanowski, 20 Traits

| | Li | Sl | Go | Ir | La | Gr | Ve | Av | Ar |
|--------------|------|------|------|------|------|------|------|------|------|
| Lithuanian | 1. | .96 | .57 | -.72 | -.91 | -.57 | -.50 | -.28 | -.21 |
| Old Slavonic | .96 | 1. | .65 | -.44 | -.81 | -.64 | -.46 | -.16 | -.43 |
| Gothic | .57 | .65 | 1. | -.21 | -.48 | -.51 | -.65 | -.42 | -.62 |
| Old Irish | -.72 | -.44 | -.21 | 1. | .90 | .21 | .21 | -.07 | -.52 |
| Latin | -.91 | -.81 | -.48 | .90 | 1. | .48 | .50 | .02 | -.32 |
| Greek | -.57 | -.64 | -.51 | .21 | .48 | 1. | .65 | .42 | -.02 |
| Vedic | -.50 | -.46 | -.65 | .21 | .50 | .65 | 1. | .95 | .11 |
| Avestan | -.28 | -.16 | -.42 | -.07 | .02 | .42 | .95 | 1. | .43 |
| Armenian | -.21 | -.43 | -.62 | -.52 | -.32 | -.02 | .11 | .43 | 1. |

II: Moszynski, 19 Traits

| | Li | Sl | Go | Ir | La | Gr | Ve | Av | Ar |
|--------------|------|------|------|------|------|------|------|------|------|
| Lithuanian | 1. | .96 | .22 | -.80 | -.85 | -.40 | -.28 | .19 | .06 |
| Old Slavonic | .96 | 1. | .31 | -.51 | -.78 | -.45 | .09 | .33 | -.17 |
| Gothic | .22 | .31 | 1. | .02 | -.07 | -.34 | -.61 | -.39 | -.70 |
| Old Irish | -.80 | -.51 | .02 | 1. | .95 | .12 | -.09 | -.33 | -.80 |
| Latin | -.85 | -.78 | -.07 | .95 | 1. | .19 | -.18 | -.49 | -.65 |
| Greek | -.40 | -.45 | -.34 | .12 | .19 | 1. | .59 | .33 | -.17 |
| Vedic | -.28 | .09 | -.61 | -.09 | -.18 | .59 | 1. | .95 | .38 |
| Avestan | .19 | .33 | -.39 | -.33 | -.49 | .33 | .95 | 1. | .45 |
| Armenian | .06 | -.17 | -.70 | -.80 | -.65 | -.17 | .38 | .45 | 1. |

From the presence and absence in the nine languages of twenty of these elements he computed by the formula known as Q_6 the coefficients which are given in Table I. Applying the same method to a somewhat different selection of nineteen elements chosen by Moszynski, he obtained the coefficients given in Table II. These two tables he then transformed into the graphic diagrams which we reproduce as figure 1 (for Table I) and figure 2 (for Table II).

The outstanding feature of both tables is the high coefficients between Lithuanian and Old Church Slavic, between Old Irish and Latin, and

between Vedic and Avestan. For these the coefficients range between .90 and .96 as against maximum coefficients of .65 (Table I) and .59 (Table II) and many negative coefficients between other languages. In short this statistical treatment first of all affirms the well-recognized Balto-Slavic, Italo-Celtic, and Indo-Iranian groups.

As between the two tables, Moszynski's agrees better with the general opinion of Indo-European linguists. It gives Balto-Slavic and Indo-Iranian, whose closer relationship has long been accepted, positive

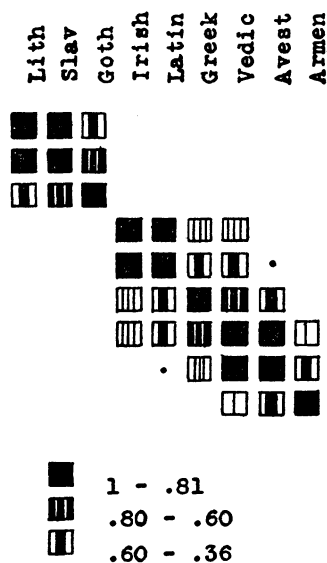


Fig. 1

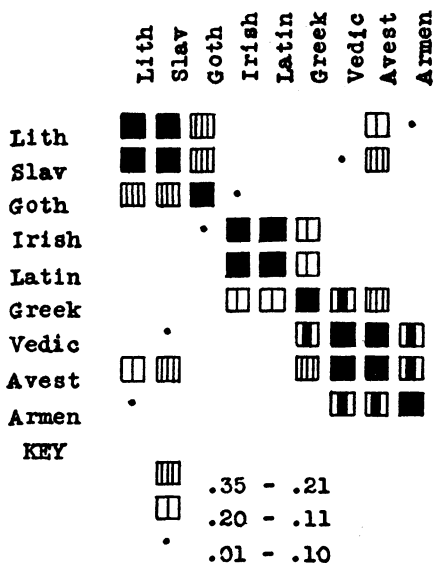


Fig. 2

Figs. 1, 2.—1, left, Czekanowski, 20 elements; 2, Moszynski, 19 elements, formula Q_6 .

coefficients with each other three times out of four (.33, .19, .09, —.28) instead of the all-negative coefficients of Table I (— .16, —.28, —.46, —.50). Armenian is closer to Balto-Slavic and Indo-Iranian and more divergent from Italo-Celtic than in Table I. This again is more in agreement with current opinion. These results suggest that Moszynski's is a more genuinely random selection of elements.

Where both tables disagree violently with accepted opinion is in linking Gothic more closely with Balto-Slavic than with Italo-Celtic: .22 and .31 vs. —.07 and .02 by Moszynski's list; and, even more fla-

grantly, .57 and .65 vs. —.48 and —.21 by Czekanowski's. This strongly controverts the prevalent classification, and will be discussed later.

A minor point is that in both tables it is the Indic member of Indo-Iranian which is the closer to Italo-Celtic and Greek, and the Iranian which is the closer to Balto-Slavic, Gothic, and Armenian. The first half of this finding is difficult to reconcile with geographical position.

3. New Data and Results

Because of the historical import of linking Germanic (Gothic) so closely with Balto-Slavic and so remotely with Italo-Celtic, it seemed desirable to compile a larger list of Indo-European elements, in order to check against the possibility that the results from both Czekanowski's and Moszynski's lists involved a statistical error due to too small a sample or to a preselected one. Such a larger list was compiled by Kroeber who extracted from Meillet's *Dialectes Indo-Européens* seventy-one elements with nearly complete distributions. This list was revised by Chrétien in the following manner: no new elements were added; duplications were eliminated; multiple elements were divided into single elements; a few elements of vocabulary were dropped; and the distributions were checked with Brugmann's *Grundriss*.

The validity of the list thus compiled depends on its being a random selection. It must be emphasized that Kroeber, who originally selected the list, is not an Indo-European linguist, but an anthropologist and American Indian linguist who was choosing elements at random in an unfamiliar field. Chrétien, who revised the list, simply put the elements into a form usual among Indo-European linguists. The authors feel therefore that the random quality necessary for valid results has been achieved.

For various reasons it was impossible to get enough data on Albanian and Tocharian. It will be noticed that our basis is broader than Czekanowski's. For example, we do not use Gothic but Germanic: thus a given element is considered as it is present or absent in primitive Germanic, not in one descendent of primitive Germanic. In the main this is true of all the languages listed: we are dealing with the principal language-groups of the Indo-European family. At the same time our nine speech-groups correspond substantially to Czekanowski's.

The list of elements follows. The order given is that of their occurrence in Meillet's *Dialectes*; the chapter references are to Meillet. We have preserved this order so that the random character of the selection will be obvious.

LIST OF ELEMENTS

References

- Br: K. Brugmann and B. Delbrück, *Grundriss der Vergleichenden Grammatik der Indogermanischen Sprachen* (2nd ed., Strassburg: 1897-1916).
 K: E. Kieckers, *Einführung in die Indogermanische Sprachwissenschaft* (Munich: 1933), vol. I.
 M¹: A. Meillet, *Les Dialectes Indo-Européens* (Paris: 1922).
 M²: A. Meillet, *Introduction à l'Étude Comparative des Langues Indo-Européennes* (7th ed., Paris: 1934).
 Sommer: F. Sommer, *Handbuch der Lateinischen Laut- und Formenlehre* (2nd and 3rd ed., Heidelberg: 1914).

Chapter II

1. Assimilation of **e* and **a*:
 M¹ 25; Br 1. §§116-20, 123-24, 128; M² 98.
2. Assimilation of **ə* and **i*, instead of **ə* and **a*.
 M¹ 25, 62; Br 1. §§194-201.
3. Voiced explosive aspirate plus voiceless consonant becomes voiced explosive plus voiced aspirate (law of Bartholomae):
 M¹ 25.
4. Genitive and ablative singular of *ā*-stems in *-āyā-* (*-(i)y-*):
 M¹ 26.
5. Genitive plural **-ōm* replaced in vowel stems by *-n-ām*:
 M¹ 26; Br 2.2. §§251-253.
6. Imperative third person in *-u*:
 M¹ 26.

Chapter III

7. Assimilation of **r* **l* and **vr* **vl*:
 M¹ 33-5; Br 1. §498; M² 117, 119.
8. Assimilation of **n* **ne* and **vn* **vm*:
 Br 1. §430; K §§34-5, 38; Sommer §§36-7, 41-2.
9. **p* ... *k* becomes *k* ... *k*:
 M¹ 33.
10. Genitive singular of *o*-stems in *ī*:
 M¹ 35; Br 2.2. §153.
11. Formative suffix **tiēn*, **tiēn*:
 M¹ 37; Br 2.1. §231.
12. Superlative in **-σημο-*, **-isημο-*:
 M¹ 37; Br 2.1. §§158-159.
13. Future in *-bō*:
 M¹ 37.
14. Passive in *-r-*:
 M¹ 35-6.
15. Subjunctive in *-ā-*:
 M¹ 36.

16. Subjunctive in -s-:

M¹ 36-7.

17. Use of verbal adjective in *-to- as past participle:

Br 2.1. §291; Br 2.3. §§834, 837-8, 840.

Chapter IV

18. Simplification of geminated consonants:

M¹ 43.

19. Present participle masculine and neuter with *-ǵo- inflection by analogy to feminine *-ǵā- inflection:

M¹ 45.

Chapter V

20. *Centum* becomes *satem*:

M¹ 51; Br 1. §596.

21. Labio-velars become velars:

M¹ 49; Br. 1. §648.

Chapter VI

22. Assimilation of *o and *a:

M¹ 54; Br 1. §§139-43, 146-8; M² 98.

23. *ā and *ō not assimilated:

M¹ 55; Br 1. §78.

24. Assimilation of *ā to *ō:

M¹ 55; Br 1. §78.

25. Assimilation of *ō to *ā:

M¹ 55; Br 1. §78.

Chapter VII

26. Shift of *-tt- to -st-:

M¹ 57-61.

Chapter VIII

27. Loss of medial *ə before a consonant:

M¹ 63; M² 101.

28. Loss of medial *ə before a consonant and following a syllable containing *o:

M¹ 68-9; M² 101.

29. Loss of medial *ə before a consonant and after a vowel plus *i:

M¹ 66-7, 70.

Chapter IX

30. Shift of *-uǵi- to *-uǵi-:

M¹ 71-4; Br 1. §320.

Chapter X

31. Voiced explosive aspirates remain unchanged:
M¹ 74; M² 87.
32. Voiced explosive aspirates become voiced explosives:
M¹ 74; M² 87.
33. Voiced explosive aspirates become spirants:
M¹ 74-6; M² 87.

Chapter XI

34. Voiceless explosive aspirates and voiceless explosives kept distinct:
M¹ 78-80; M² 85.
35. Voiceless explosive aspirates partly assimilated to voiceless explosives:
M¹ 78-80; M² 85.
36. Voiceless explosive aspirates wholly assimilated to voiceless explosives:
M¹ 78-80; M² 85.
37. Voiceless explosive aspirates become spirants:
M¹ 78-80; M² 85.

Chapter XII

38. *s becomes š after *i, u, r, k* regularly:
M¹ 84; Br 1. §819.
39. *s becomes š after *i, u, r, k* only when a vowel of the same word follows:
M¹ 84-5.
40. *s becomes *h* when (a) initial, (b) intervocalic, or (c) before or after a consonant not a stop:
M¹ 86-87.

Chapter XIII

41. Voiceless explosives become spirants:
M¹ 91; M² 85.
42. Voiced explosives become voiceless explosives:
M¹ 90-1; M² 88.

Chapter XIV

43. Augment:
M¹ 97.

Chapter XV

44. Reduplicated perfect regular:
M¹ 103.
45. Reduplication occasional:
M¹ 105-6.
46. No reduplication:
M¹ 104.
47. Perfect is preserved:
M¹ 103.

48. Preterite is derived partly from perfect and partly from aorist:
M¹ 108.
49. Preterite is derived entirely from aorist:
M¹ 104.

Chapter XVI

50. Verbal suffix **-iē-/*-īo-* used for derivatives; suffix **-i-* for states:
M¹ 109.
51. Verbal suffixes **-iē-/*-īo-* and **-i-* used for derivatives:
M¹ 113.
52. Verbal suffix **-iē-/*-īo-* used for derivatives and states; suffix **-i-* not so used:
M¹ 109.

Chapter XVII

53. Verbal abstracts of the type root plus **o/*ā* frequent:
M¹ 115; Br 2.1. §§90-2.
54. Verbal abstracts of the type root plus **o/*ā* not frequent, but more than sporadic:
M¹ 115; Br 2.1. §§90-2.
55. Verbal abstracts of the type root plus **o/*ā* sporadic:
M¹ 115; Br 2.1. §§90-2.
56. Comparative in **-iēs-, *-īos-, *-is-*:
M¹ 114; Br 2.1. §§423, 429-30, 432, 435, 440.
57. Comparative in **-isen-, *-ison-*:
M¹ 115; Br 2.1. §§425, 430, 436, 439.
58. Suffix **-tero-, *-toro-, *-tro-* used for comparative:
M¹ 114; Br 2.1. §§238, 240.
59. Suffix **-tero-, *-toro-, *-tro-* used in certain words originally comparative, but which have lost the comparative force:
M¹ 114; Br 2.1. §§238, 240.
60. Participles formed by suffix **-lo-*:
M¹ 114-5.
61. **o-* stems are feminine as well as masculine:
M¹ 116.
62. Suffix **-tūt-* forms abstract nouns commonly:
M¹ 115; Br 2.1. §343.
63. Suffix **-tūt-* forms abstract nouns rarely:
M¹ 115; Br 2.1. §343.
64. Suffix **-tūt-* not used:
M¹ 115; Br 2.1. §343.
65. Collective numbers in **-o-* (Sk *trayāḥ*):
M¹ 116; Br 2.2. §81.
66. Collective numbers in **-no-* (Lat *trīnī*):
M¹ 116; Br 2.2. §82.
67. Comparative in **-iēs-, *-īos-, *-is-* lacks feminine:
M¹ 115.

Chapter XVIII

68. Case suffix in **-bh-* replaced by case suffix in **-m-*:
M¹ 119; Br 2.2. §§275, 287.
69. Locative plural in **-su*:
M¹ 123; Br 2.2. §262.
70. Breakdown of original case system and amalgamation of the functions of dative, ablative, locative, instrumental:
M¹ 120-2.

Chapter XIX

71. Nominative plural of **o*-stems in *-oi* under influence of demonstrative:
M¹ 124-5; Br. 2.2. §219.
72. Nominative plural of **ā*-stems in *-āi* on analogy of **o*-stems:
M¹ 124-5.
73. Genitive plural of **ā*-stems uses demonstrative form (**-āsōm*):
M¹ 125; Br 2.2. §256.

Chapter XX

74. Forms of **bheya-* 'to grow' partly replace **es-* 'to be':
M¹ 126.

For the benefit of anyone who would care to examine these elements from the point of view of systematic grammar, the following classification is given:

1. Phonology: Total 30.
 - a. Vowels: Total 12.
1, 22, 2, 7, 8, 23, 24, 25, 27, 28, 29, 30.
 - b. Consonants: Total 18.
41, 34, 35, 36, 37, 42, 31, 32, 33, 3, 20, 21, 26, 38, 39, 40, 9, 18.
2. Morphology: Total 44.
 - a. Nouns: Total 17.
4, 10, 71, 72, 68, 5, 73, 69, 70, 61, 11, 53, 54, 55, 62, 63, 64.
 - b. Adjectives: Total 8.
65, 66, 56, 67, 57, 58, 59, 12.
 - c. Verbs: Total 19.
13, 43, 44, 45, 46, 47, 48, 49, 14, 15, 16, 6, 60, 19, 17, 50, 51, 52, 74.

It will be noticed that we confine our study to phonology and morphology, and that there is a fairly even distribution of elements over these fields. This evenness is, of course, purely random, since no effort was made to attain it when Kroeber selected the list.

Here follows the tabulation of occurrences, which we give to permit checking both on our presence and absence decisions and on our counting.

TABULATION OF OCCURRENCES

| | Ce | It | Gr | Ar | Ir | Sk | Sl | Ba | Ge | | Ce | It | Gr | Ar | Ir | Sk | Sl | Ba | Ge |
|----|----|----|-----|-----|-----|----|-----|-----|-----|----|----|----|----|-----|----|----|----|-----|----|
| 1 | - | - | - | - | + | + | - | - | - | 38 | - | - | - | ? | + | + | - | - | - |
| 2 | - | - | - | - | + | + | - | - | - | 39 | - | - | - | ? | - | - | + | (+) | - |
| 3 | - | - | - | - | + | + | - | - | - | 40 | - | - | + | + | + | - | - | - | + |
| 4 | + | - | - | + | + | + | - | - | - | 41 | - | - | - | + | - | - | - | - | + |
| 5 | - | - | - | - | + | + | - | - | (+) | 42 | - | - | - | + | - | - | - | - | + |
| 6 | - | - | - | - | + | + | - | - | - | 43 | - | - | + | + | + | + | - | - | - |
| 7 | - | - | + | + | - | - | - | + | + | 44 | - | - | + | ? | + | + | - | - | - |
| 8 | + | + | - | + | - | - | - | + | + | 45 | + | + | - | - | - | - | - | - | + |
| 9 | + | + | - | - | - | - | - | - | - | 46 | - | - | - | - | - | - | + | + | - |
| 10 | + | + | - | - | - | - | - | - | - | 47 | - | - | + | (+) | + | + | - | - | - |
| 11 | + | + | - | - | - | - | - | - | - | 48 | + | + | - | - | - | - | - | - | + |
| 12 | + | + | - | - | - | - | - | - | - | 49 | - | - | - | - | - | - | + | + | - |
| 13 | + | + | - | - | - | - | - | - | - | 50 | - | - | - | + | - | - | + | + | - |
| 14 | + | + | - | - | - | - | - | - | - | 51 | + | + | - | - | - | - | - | - | + |
| 15 | + | + | - | - | - | - | - | - | - | 52 | - | - | + | - | + | + | - | - | - |
| 16 | + | + | - | - | - | - | - | - | - | 53 | - | - | + | - | + | + | + | + | - |
| 17 | + | + | - | - | + | + | + | + | - | 54 | - | - | - | - | - | - | - | - | + |
| 18 | - | - | - | + | - | - | + | + | - | 55 | + | + | - | - | - | - | - | - | - |
| 19 | - | - | - | - | - | - | + | + | + | 56 | + | + | + | - | + | + | + | - | - |
| 20 | - | - | - | + | + | + | + | + | - | 57 | - | - | + | - | - | - | - | + | + |
| 21 | - | - | - | + | + | + | + | + | - | 58 | + | - | + | - | + | + | - | - | - |
| 22 | - | - | - | - | + | + | + | + | + | 59 | - | + | - | - | - | - | - | - | + |
| 23 | - | + | + | + | - | - | - | - | - | 60 | - | - | - | + | - | - | + | - | - |
| 24 | - | - | - | - | - | - | - | + | + | 61 | - | + | + | + | - | - | - | - | - |
| 25 | + | - | - | - | + | + | + | - | - | 62 | + | + | - | - | - | - | - | - | - |
| 26 | - | - | + | ? | + | + | + | + | - | 63 | - | - | - | - | + | - | - | - | + |
| 27 | - | - | - | + | + | + | + | + | + | 64 | - | - | + | + | - | + | + | + | - |
| 28 | ? | + | + | (+) | (+) | - | (+) | (+) | (+) | 65 | - | - | - | - | + | + | + | + | - |
| 29 | - | - | - | - | - | + | - | - | - | 66 | - | + | - | - | - | - | - | + | + |
| 30 | - | - | - | - | + | - | + | + | + | 67 | + | + | + | - | - | - | - | - | - |
| 31 | - | - | - | - | - | + | - | - | - | 68 | - | - | - | (-) | - | - | + | + | + |
| 32 | + | - | - | + | + | - | + | + | - | 69 | - | - | - | + | + | + | + | + | - |
| 33 | - | + | + | - | - | - | - | - | + | 70 | + | + | + | - | - | - | - | - | + |
| 34 | - | - | + | (+) | + | + | (-) | - | - | 71 | + | + | + | - | - | - | + | + | - |
| 35 | - | - | (-) | + | - | - | (+) | - | - | 72 | - | + | + | - | - | - | - | - | - |
| 36 | + | + | (-) | - | - | - | (+) | + | + | 73 | - | + | + | - | - | - | - | - | - |
| 37 | - | - | + | (-) | + | - | - | - | + | 74 | + | + | - | - | + | + | + | + | + |

We have marked a number of entries in the tabulation as (+) and (-). The ordinary plus and minus indicate that the statement of the element is universally true or not true. When enclosed in parentheses they indicate that the statement is generally but not universally true or not true. We have counted (+) as +, and (-) as -. The question

TABLE III

Kroeber-Chrétien, 74 elements, formula Q_6

| | Ce | It | Gr | Ar | Ir | Sk | Sl | Ba | Ge |
|----------|------|------|------|------|------|------|------|------|------|
| Celtic | 1. | .87 | -.25 | -.46 | -.30 | -.30 | -.18 | -.30 | -.12 |
| Italic | .87 | 1. | .09 | -.48 | -.66 | -.63 | -.40 | -.35 | .11 |
| Greek | -.25 | .09 | 1. | .25 | .22 | .28 | -.84 | -.16 | -.17 |
| Armenian | -.46 | -.48 | .25 | 1. | .22 | .05 | .31 | .28 | -.18 |
| Iranian | -.30 | -.66 | .22 | .22 | 1. | .91 | .32 | .12 | -.22 |
| Sanskrit | -.30 | -.63 | .28 | .05 | .91 | 1. | .20 | .10 | -.54 |
| Slavic | -.18 | -.40 | -.84 | .31 | .32 | .20 | 1. | .92 | -.11 |
| Baltic | -.30 | -.35 | -.16 | .28 | .12 | .10 | .92 | 1. | .32 |
| Germanic | -.12 | .11 | -.17 | -.18 | -.22 | -.54 | -.11 | .32 | 1. |

mark indicates that information is lacking, and the entry omitted from computation.

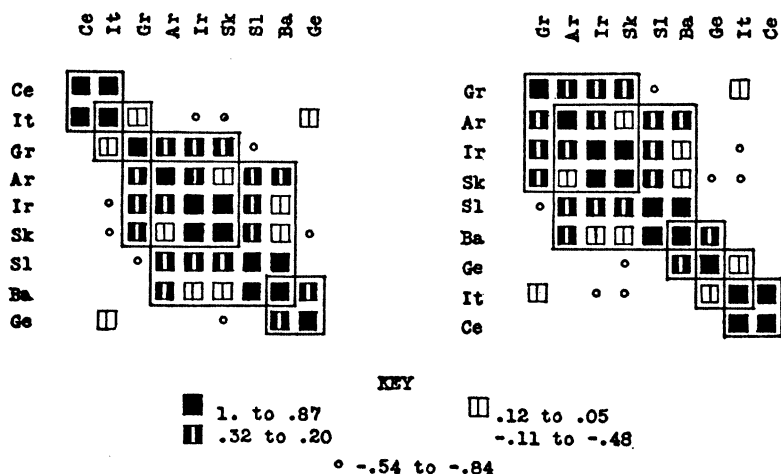


Fig. 3a

Fig. 3b

Figs. 3a, 3b.—74 elements, formula Q_6

From the tabulation just given we have determined the values of a , b , c , and d for each pair of language-groups, and then, using formula Q_6 , have computed the coefficients. These will be found in Table III. The graphic representation of this table we give as figures 3a and 3b.

In a situation like the present one, where the primary relationships more or less go round a ring, the order of arranging the languages in the table is arbitrary. In the order given, however, the related languages

are grouped together, as will be seen in figure 3a. Only Germanic-Italic falls off the diagonal. Figure 3b gives a slightly different order, which brings Germanic-Italic on the diagonal, but drops Greek-Italic into the corner.

It is now time to examine the findings of Table III. The three outstandingly high coefficients are Baltic-Slavic .92, Iranian-Sanskrit .91, and Italic-Celtic .87. From these there is a long drop to the next highest coefficients, Baltic-Germanic .32, and Iranian-Slavic .32. The three high coefficients, then, agree with all non-statistical judgments, even to ranking the Italic-Celtic similarity a little lower than the Baltic-Slavic and the Iranian-Sanskrit.

In figures 3a and 3b we have indicated by square boxes those groups of languages which have positive coefficients *with each other*. These boxes give us some very interesting results. In the first place they affirm the existence of a *satem* group of languages. And the range of the coefficients within this group is interesting because it seems to coincide with geography. Armenian, which lies geographically between Slavic and Iranian (disregarding intervening languages), with Baltic out beyond Slavic and Sanskrit out beyond Iranian, occupies this same position according to the coefficients: .31 with Slavic, but only .28 with Baltic; .22 with Iranian, but only .05 with Sanskrit. Likewise Baltic is nearer to Iranian (.12) than to Sanskrit (.10); Slavic is nearer to Iranian (.32) than to Sanskrit (.20); Iranian is nearer to Slavic (.32) than to Baltic (.12); Sanskrit is nearer to Slavic (.20) than to Baltic (.10). Confining ourselves, for the moment, to these five languages, we see that the coefficients correspond in range to the geographical positions of the languages. This is very interesting and suggests the possibility of a correlation between geographical factors and degree of similarity. We must point out here that this verdict of our coefficients corresponds with the general linguistic opinion.

Though our coefficients show the existence of a *satem* group, they do not indicate a *centum* group. Of the remaining languages, Greek has the highest number of positive coefficients. Its closest affinities are to Sanskrit (.28), Armenian (.25), and Iranian (.22). The only other positive coefficient is the very low one (.09) with Italic. In other words, Greek has more in common with the *satem* languages than with the *centum*—except the one characteristic which serves to distinguish *satem* and *centum* languages! This all seems to mean one thing: that the division into *centum* and *satem* languages was a purely arbitrary, not an organic division, but that, so far as the *satem* languages were

concerned, it happened, accidentally, to be right. Modern linguistic opinion has generally recognized the difficulties of the *centum-satem* classification, and our statistical method confirms this opinion.

So far the results of the objective method of counting have been, in the main, in accord with the subjective judgments of linguists. But when we come to Germanic the two methods diverge. Germanic has only two positive coefficients, with Italic (.11), and with Baltic (.32). Linguistic opinion grants Germanic a high degree of disparity with other Indo-European languages. Some linguists would concede it a relation with Italo-Celtic, but all they mean is that of all its distant relationships, the Italo-Celtic is least distant. Our coefficients, however, link it definitely to two languages which have with each other a negative coefficient ($-.35$). The question before us now is this: when the objective and the subjective methods diverge, which shall we follow? On the one side stands Meillet, representative certainly of linguistic opinion, who devotes half of chapter XX of his *Dialectes* to the Germanic-Italic-Celtic group, and in his Conclusion describes them as a natural group. On the other side stands the statistical treatment of his own data.

If Meillet's judgment is right, then the statistical technique is inapplicable. Yet if so, why the close agreement at all previous points between the objective and the subjective, the statistical and the linguistic findings? This would appear to be one of the times when the best insight nods and makes a partially misjudged evaluation. One observes a certain affiliation which is real enough, but perhaps secondary; thereafter he notes mentally every corroborative item, but unconsciously overlooks or weighs more lightly items which point in other directions. Kroeber has done this very thing in his specialty of Californian ethnography—until the coefficients were worked out. After all, nine languages present thirty-six interrelations, and that a scholar should estimate two or three of these somewhat too high or too low is almost inevitable.

We must conclude then that, unless the method or its application (i.e. the selection of the seventy-four elements used) is faulty, the nearest relatives of Germanic are Baltic and Italic. Moreover, the next nearest (though the coefficients are negative) are Slavic ($-.11$) and Celtic ($-.12$). The arithmetical mean of Germanic with Baltic and Slavic is .11, with Italic and Celtic is $-.01$. Germanic is thus to be linked with the first two rather than the second two; or better, it occupies a more or less medial position between the two groups, with a leaning towards the Balto-Slavic.

Our coefficients differ in absolute value from both the Czekanowski and the Moszynski coefficients, but their relative rankings are about the same as those of the Moszynski table, and agree to a considerable extent with those of the Czekanowski table. It ought not be difficult to list practically complete distribution for two hundred or more elements. But the closeness with which the major findings from nineteen or twenty elements have been corroborated by those from seventy-four elements makes it seem likely that a re-tripling of the material will not revolutionize results. In short, even a quite small sample of Indo-European material is likely to yield approximately valid results, provided it is wholly random.

4. Methodology

The basic formula for determining the coefficient of association or similarity is generally considered to be Karl Pearson's 'tetrachoric R'. This is not convenient to handle as it involves elaborate computation. The formula Q_6 , which we have used, is as follows:

$$Q_6 = \sin \left[\frac{\pi}{2} \left(\frac{ad - bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}} \right) \right]$$

Some statisticians prefer to simplify this by omitting the sine and the $\pi/2$. Thus we would have

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

It is obvious that this will give the same relative results as Q_6 , since the omitted elements do not involve the statistical data.

Another formula, a good deal simpler than Q_6 , is the following:

$$Q_2 = \frac{ad - bc}{ad + bc}$$

This computes much more quickly than Q_6 and gives usually not very different results, but it cannot be used where a , b , c , or d is zero, because the coefficient is then either 1 or -1 . For a distribution like a 50, b 25, c 25, and d 0, that is, 50 agreements against 50 disagreements, a coefficient like -1 (which means complete negative correlation) is obviously misrepresentative. For comparison we add in Table IV the Q_2 coefficients on the present set of data.

The spread of coefficients by Q_2 is wider than by Q_6 but the graphical representation (fig. 4) shows that the relative rankings are identical for the positive values.

TABLE IV
Kroeber-Chrétien, 74 elements, formula Q_2

| | Ce | It | Gr | Ar | Ir | Sk | Sl | Ba | Ge |
|----|------|------|------|------|------|------|------|------|------|
| Ce | 1. | .94 | -.37 | -.67 | -.41 | -.42 | -.26 | -.42 | -.17 |
| It | .94 | 1. | .13 | -.66 | -.81 | -.82 | -.54 | -.47 | .15 |
| Gr | -.37 | .13 | 1. | .35 | .30 | .39 | -.33 | -.23 | -.25 |
| Ar | -.67 | -.66 | .35 | 1. | .29 | .07 | .42 | .38 | -.27 |
| Ir | -.41 | -.81 | .30 | .29 | 1. | .96 | .42 | .15 | -.31 |
| Sk | -.42 | -.82 | .39 | .07 | .96 | 1. | .26 | .17 | -.75 |
| Sl | -.26 | -.54 | -.33 | .42 | .42 | .26 | 1. | .96 | -.15 |
| Ba | -.42 | -.47 | -.23 | .38 | .15 | .17 | .96 | 1. | .43 |
| Ge | -.17 | .15 | -.25 | -.27 | -.31 | -.75 | -.15 | .43 | 1. |

TABLE V
Kroeber-Chrétien, 74 elements, formula W

| | Ce | It | Gr | Ar | Ir | Sk | Sl | Ba | Ge |
|----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Ce | 1. | .85 | .49 | .42 | .44 | .45 | .49 | .45 | .52 |
| It | .85 | 1. | .57 | .39 | .30 | .32 | .41 | .42 | .57 |
| Gr | .49 | .57 | 1. | .64 | .59 | .59 | .49 | .50 | .51 |
| Ar | .42 | .39 | .64 | 1. | .60 | .57 | .64 | .63 | .50 |
| Ir | .44 | .30 | .59 | .60 | 1. | .87 | .62 | .55 | .46 |
| Sk | .45 | .32 | .59 | .57 | .87 | 1. | .60 | .53 | .38 |
| Sl | .49 | .41 | .49 | .64 | .62 | .60 | 1. | .88 | .51 |
| Ba | .45 | .42 | .50 | .63 | .55 | .53 | .88 | 1. | .64 |
| Ge | .52 | .57 | .51 | .50 | .46 | .38 | .51 | .64 | 1. |

A third formula which we may use is W .

$$W = \frac{a + d}{a + b + c + d}$$

This is simply the total of agreements, positive and negative, divided by the total number of agreements plus disagreements. Table V gives these coefficients.

It will be seen from figure 5, which represents Table V, that the rank order in the main is like that of Q_6 . A closer examination of the table will show shifts of relationship, however, among those languages which lie close together. The values by this formula will lie between +1 and zero. W does not seem to have been analyzed by statistical theorists, but its logical transparency may commend it.

We have given these various tables and diagrams with this end in view: to show that formula does not matter very materially, as we have already said in Part 1. What matters is the value of the sample of data used. As we have indicated before, the sample must be genuinely random, of sufficient size, and made up of elements sharply defined.

At least two methodological doubts which may have arisen need to be considered. The first is the problem of equivalence of elements. For example, is a locative plural in *-su* more or less important than a reduplicated perfect? and if these two are not of equal importance, is it right to treat them so in our list of elements? The answer is that

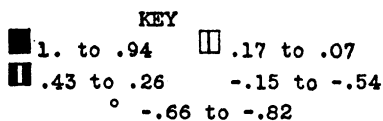
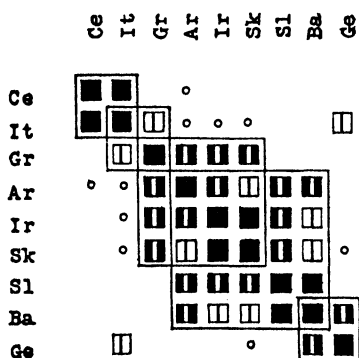


Fig. 4

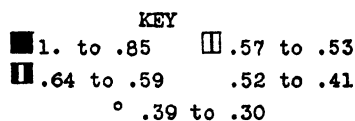
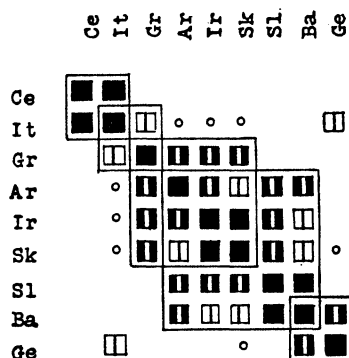


Fig. 5

Figs. 4, 5.—74 elements. 4, left, formula Q_2 ; 5, formula W

statistical analysis presupposes, ideally, the inclusion of all the elements of the field or universe of observation, and practically, at least a thoroughly random sample. If the sample is really random, and not too small, there is very little likelihood that the more fundamental elements will co-occur mainly in one grouping or pattern, and the more trivial elements in a different pattern, among the languages being compared. Why should they? And statistical experience in other fields is to the contrary.

In the second place, negative elements may occasion some doubts. The answer is that if four Indo-European languages possess augment

and five do not, the absences mean as much as the presences. Thus a common absence in Baltic and Germanic is a *point of similarity between the two*, which is certainly as important in comparing them as its co-occurrence in, let us say, Greek and Sanskrit is significant in comparing these latter. Likewise its presence in Greek and its absence in Italic is of significance in a comparison between these two. Fundamentally what are being compared are agreements (*a* and *d*) and disagreements (*b* and *c*).

Of course elements whose distribution is wholly negative in the particular universe which is being examined must be excluded. Naturally such elements are not a part of the universe under consideration. Likewise elements whose distribution is universally positive should be omitted: they prove nothing and tend to smear the results by preventing pertinent spread of coefficients.

The seventy-four elements here dealt with are prevaillingly negative in their Indo-European distribution: plus, 237 (36%); minus, 424 (64%) ?, 5. This means that an element occurs, on an average, in three of the nine languages considered. This means that in every one of the thirty-six interrelationships, common absences (*d*) will outweigh common presences (*a*). Thus for Baltic-Slavic we have $a = 23$, but $d = 42$. Where the similarity is remote the disproportion is greater: thus in Germanic-Sanskrit, $a = 3$, but $d = 25$. So long as the elements occur, in the mean, in only three languages out of nine, this heavy occurrence of common absences must be accepted. And this heavy occurrence has a meaning: it recalls to our attention the fact that the nine languages are, after all, nine different languages, already widely diversified when they are first encountered in history, in spite of the indubitable unity in their origin; and this fact does not make the investigation of the respective degrees of their inter-relationships any less sound a problem than if they were more similar to one another.

However, since experience in other fields has shown that shared absences will be felt by some as less significant than shared presences, the problem will be reapproached with the common absences, *d*, omitted. A formula that has actually been used in other fields is $a/(a + b)$, that is, the ratio of elements present in, let us say, languages I and II to the number of elements present in language I. To complete the picture, $a/(a + c)$ should be computed for language II. Except when *b* and *c* happen to be equal, the two values are unlike, and this prevents symmetrical diagramming. This difficulty can be overcome by taking the arithmetical mean of the two values (*A*) or the geometrical mean

(G). Both A and G have been employed by Driver and Kroeber in ethnography,³ and the results, in the cases tested, come out not very different from those by the other formulas. There is, however, a theoretical objection to A and G which Driver has subsequently pointed out.⁴ They omit *d* because it is wholly concerned with absences; but *b* and *c* also include absences because they refer to distributions in which an element is lacking in one entity, though present in the other. If only positive elements ought to be dealt with, the negatives implied

TABLE VI
Kroeber-Chrétien, 74 elements, elements shared (a)

| | Ce | It | Gr | Ar | Ir | Sk | Sl | Ba | Ge |
|----|----|----|----|----|----|----|----|----|----|
| Ce | — | 21 | 5 | 3 | 7 | 6 | 7 | 6 | 7 |
| It | 21 | — | 10 | 4 | 4 | 3 | 6 | 7 | 11 |
| Gr | 5 | 10 | — | 9 | 12 | 10 | 6 | 7 | 6 |
| Ar | 3 | 4 | 9 | — | 11 | 8 | 11 | 11 | 6 |
| Ir | 7 | 4 | 12 | 11 | — | 24 | 15 | 13 | 8 |
| Sk | 6 | 3 | 10 | 8 | 24 | — | 12 | 10 | 3 |
| Sl | 7 | 6 | 6 | 11 | 15 | 12 | — | 23 | 8 |
| Ba | 6 | 7 | 7 | 11 | 13 | 10 | 23 | — | 13 |
| Ge | 7 | 11 | 6 | 6 | 8 | 3 | 8 | 13 | — |

TABLE VII
Kroeber-Chrétien, 74 elements, isoglosses (b + c)

| | Ce | It | Gr | Ar | Ir | Sk | Sl | Ba | Ge |
|----|----|----|----|----|----|----|----|----|----|
| Ce | — | 11 | 37 | 40 | 41 | 40 | 37 | 40 | 35 |
| It | 11 | — | 32 | 43 | 52 | 50 | 44 | 43 | 32 |
| Gr | 37 | 32 | — | 25 | 30 | 30 | 38 | 37 | 36 |
| Ar | 40 | 43 | 25 | — | 28 | 30 | 25 | 26 | 35 |
| Ir | 41 | 52 | 30 | 28 | — | 10 | 28 | 33 | 40 |
| Sk | 40 | 50 | 30 | 30 | 10 | — | 30 | 35 | 46 |
| Sl | 37 | 44 | 38 | 25 | 33 | 30 | — | 9 | 36 |
| Ba | 40 | 43 | 37 | 26 | 28 | 30 | 9 | — | 27 |
| Ge | 35 | 32 | 36 | 35 | 40 | 46 | 36 | 27 | — |

in *b* and *c* make these as inadmissible, theoretically, as the patent double negatives of *d*. The A and G formulas, which are based on *a*, *b*, and *c*, are therefore perhaps less sound than Q₆, Q₂, or W, which are based on *a*, *b*, *c*, and *d*. It may be best to restrict them to situations in which the data have been worked out with explicitness only for positive

³ Quantitative Expression of Cultural Relationships, Univ. Calif. Publ. Am. Arch. Ethn. 31. 211-56, 1932.

⁴ In ms.

occurrences; and even in such situations the A and G results presumably have only approximate significance.

There is a method, however, which makes use of *a*, *b*, and *c*, without using *d*. This calls for two tables. In Table VI we give the elements shared (*a*) for the universe under observation. In Table VII we give the sum of disagreements (*b* + *c*). This may be used as an inverse index of relationship. Every occurrence of *b* or *c* means that an isogloss passes between the two languages being compared. The fewer these

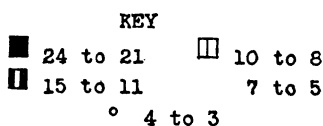
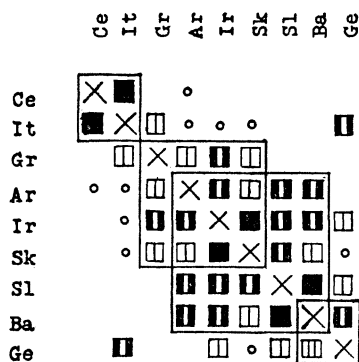


Fig. 6

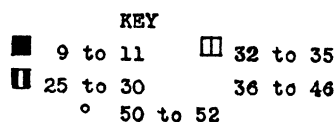
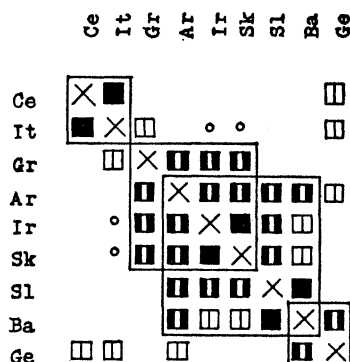


Fig. 7

Figs. 6, 7.—74 elements. 6, left, elements shared (*a*); 7, isoglosses (*b* + *c*)

separating isoglosses the closer are the two languages. These two tables are represented in figures 6 and 7 respectively.

The general resemblance of these two tables and diagrams to the others is evident. As between the two tables, where *a* is high, *b* + *c* is low; where *a* is low, *b* + *c* is high. All in all, the truest picture is probably given by the use of *a*, *b*, *c* and *d*; but if any one hesitates about the significance of *d*, he will still reach closely similar results by using *a* alone, or *b* + *c*.