63 The ABC's of Lexicostatistics (Glottochronology)

SARAH C. GUDSCHINSKY

INTRODUCTION

1. Lexicostatistics is a technique which attempts to provide dates for the earlier stages of languages much as carbon 14 dating provides dates for archaeological finds. This contrasts with previous linguistic methods which, although able to reconstruct to some extent the history of language, have been unable to provide dates apart from written historical records.

2. By simple inspection of comparable word lists, for example, the fact of the relationship of closely related languages can be discovered. But no one can say on the basis of simple inspection precisely how closely related two languages are (Swadesh, 1950, pp. 157, 164).

3. By the methods of comparative linguistics, it is possible to chart the phonemic changes by which contemporary languages have developed from a common parent language, and to reconstruct some of the vocabulary of the parent language (see Paragraph 15). This method permits the investigator to decide, to some extent, the historical order of dialect differentiation. That is, he can say that languages A and B diverged from each other before such and such a phonological change, which is peculiar to language B, took place. Or he can say that the separation of languages A and B from each other must have taken place after their separation from language C, because they share phonological features which do not occur in C. The method does not, however, permit the investigator to say at what date the separation of languages A and B took place (Hockett, 1953).

4. A method for determining the chronological relationships of cultural elements to one another by use of various kinds of linguistic evidence has been suggested by Sapir (1916, pp. 434-436). The relative antiquity, for example, of the culture items *bow*, *arrow*, and *spear* is attested by the fact that these terms cannot be analysed into constituent morphemes as can the morphologically transparent terms *railroad* or *capitalist* which represent recent additions to the culture. The assumption is that sound changes and shifts of meaning over a long period of time have obscured the original morphemic content of the older terms. Similarly, the archaic *-en* plural of *oxen* attests the ancient use of these animals, since it is assumed that words using archaic morphological processes, and the cultural elements to which the words refer, are of ancient origin. Although these and other linguistic clues discussed by Sapir have considerable value in determining something of the relative age of cultural items, and the chronological order in which they became a part of a given culture pattern, this method does not provide any exact dates. At best this method can provide the basis for such statements as: "This element was probably a part of the culture pattern before such and such sound changes took place in the language." or "This item probably entered the culture pattern of tribe A during a period of close contact with the culture of tribe B from whose language the terminology was borrowed."

5. Sapir also suggested (1921, pp. 217-220) that marked similarities in the basic morphological structure of otherwise dissimilar languages indicated remote common origin of the languages, since the effects of borrowing or other influence of one language on another seldom penetrate to the structural core or nucleus of the language affected. The use of this principle increases the number of languages that can be postulated as belonging to a given

linguistic grouping, and gives insight into linguistic relationships at deep time depths but it cannot tell us when the languages whose relationship is postulated began to diverge from one another.

6. Such historical estimation is not sufficient for the needs of anthropologists, historical linguists, and archaeologists, who want to know at just what date linguistic changes took place, and who also want to know just how the language developments correlate with cultural changes, migrations, etc., of which there is evidence from other lines of investigation (Swadesh, 1950, p. 157). Lexicostatistics is an attempt to provide the more precise dating that is needed.

BASIC ASSUMPTIONS OF LEXICOSTATISTICS

7. The first basic assumption of lexicostatistics is that some parts of the vocabulary of any language are assumed, on empirical evidence, to be much less subject to change than other parts (Swadesh [1951a], p. 12). This basic core vocabulary includes such items as terms for pronouns, numerals, body parts, geographical features, etc. This concept is similar to Sapir's idea of a basic nucleus of morphological structure discussed in Paragraph 5. Terms for new items in the material culture, on the other hand, are frequently borrowed along with the cultural items. Such terms are also easily lost with a change in the material culture, or the borrowing of a new item, or for other reasons. The contrast between the basic core vocabulary and general vocabulary may be seen in the following illustration of French loan words in English: "As against perhaps 50 percent of borrowed correspondences between English and French in the general vocabulary, we find just 6 percent in the basic vocabulary. Residual correspondences are found to be 27 percent. Thus the archaic residuum after 5000 years turns out to be five times greater than 2000 years of accumulated borrowings" (Swadesh, [1951a] p. 13).

8. The second basic assumption of lexicostatistics is that the rate of retention of vocabulary items in the basic core of relatively stable vocabulary is constant through time. That is, given a certain number of basic words in a certain language, a certain percentage of these words will remain in the language after a thousand years of vocabulary loss; that same percentage of the residue of words will remain after a second thousand years; and after a third period of a thousand years, the same percentage of the words remaining at the end of the second period will remain; and so on. Complete empirical evidence that the rate of loss is constant through time is still lacking (Lees, 1953, pp. 121-122), since the assumption has not yet been checked for a time span greater than 2,200 years and this span does not provide adequate evidence for a constant rate of loss over a long period of time.

9. The third basic assumption of lexicostatistics is that the rate of loss of basic vocabulary is approximately the same in all languages. This assumption has been tested in thirteen languages in which there are historical records. The results range from a retention of 86.4 % to 74.4 % per thousand years—an average of 80.5 % (Lees, 1953, pp. 118-119). This is not, however, conclusive evidence that all languages change at this rate, especially since all but two of the thirteen languages tested are Indo-European. (See also Kroeber, 1955, p. 91).

10. The fourth assumption of lexicostatistics is a corollary of the third, namely, that if the percentage of true cognates within the core vocabulary is known for any pair of languages, the length of time that has elapsed since the two languages began to diverge from a single parent language can be computed (Lees, 1953, pp. 116-117), provided that there are no interfering factors through migrations, conquests, or other social contacts which slowed or speeded the divergence (Swadesh, 1950, pp. 158-160; Gudschinsky, 1955, p. 149).

TECHNIQUES OF LEXICOSTATISTICS

11. In applying the lexicostatistical techniques developed from the basic assumptions, the steps are: collecting of comparable word lists from the relatively stable core vocabulary (Paragraphs 12-14); determining the probable cognates (Paragraphs 15-23, 25-28); computing the time depth (Paragraphs 31-36); computing the range of error (Paragraphs 37-45); and, optionally, computing the dips (Paragraphs 50-52).

WORD LISTS

12. The first essential in making a lexicostatistical comparison of two or more languages

is the collection of comparable word lists in the various languages. (Lexicostatistics provides a quick way of estimating linguistic relationships on the basis of a relatively small body of data. For this reason it is a useful tool in linguistic surveys. For a detailed description of gathering data in a number of dialects in minimum time, see Swadesh, 1954a.) A convenient list for this purpose is Swadesh's 200 word list. The use of this list has several advantages: it is made up of noncultural items that have been specifically chosen as a part of the core vocabulary. These items have been tentatively tested (see Paragraph 9) for percentage of retention in languages with written historical records. Later tests may well indicate that a different assortment of words would be more useful, but any revised list must be tested to ascertain whether or not the same rate of vocabulary loss applies. Meantime, this list has been used in a number of comparisons, and will yield results that can easily be compared with studies already made. It does not seem wise to start with a list shorter than 200 words, since the shorter the list of words used, the greater the probable error (see Paragraph 41). Furthermore, it is sometimes impossible to get the entire list in all of the languages investigated so that the comparisons must be made with fewer items than in the original list. For these reasons it would be good if a longer list of satisfactory items could be worked out. Swadesh is at present experimenting with the use of a list of only 100 items (see Swadesh, 1955, for a detailed analysis of the 200 word list and the suggested revision to 100 words). The reasons given for eliminating some of the items (e.g., the repetition of some roots in such pairs as woman-wife, the nonuniversality of such words as ice and snow, etc.) seem valid to this author. The gain in quality of test items, however, is balanced by some loss in terms of statistical accuracy. Kroeber (1955, p. 97) has suggested that a list of 1000 items would be preferable, and doubts that deep time depths can be explored by use of a list as small as 200 words. (Anyone choosing to use Swadesh's new list of 100 items must use .86 as the "constant" in the time depth formula of Paragraph 32.)

13. In gathering the data, each English word should be translated by the most common conversational equivalent (Swadesh 1951a, p. 13). If there is an equal choice of two or more expressions, one should be chosen purely at random (by flipping a coin if necessary) to avoid any bias in the direction of choosing known cognates, since nonrandom choice could considerably skew the final results. It is essential, for statistical reasons, that the error be random error, so that the accumulating errors tend to cancel each other out instead of compounding each other.

The same meaning of each English word should be translated in each case. For example "know" is understood as referring to facts rather than to persons. Translation from English of isolated forms in general insures that the resultant forms in each language will be comparable root stems rather than affixes or other items which are not comparable (Lees, 1953, p. 115). This is not, however, always the case, and the procedure of Paragraph 18 is used to eliminate the irrelevant material.

14. Greater time depths may be explored by the methods of lexicostatistics if the list is filled in with the reconstructed forms of the postulated common parent language of a linguistic family or stock (Swadesh, 1953a, pp. 41-42). A comparison of Proto-Romance with Proto-Germanic, for example, might be expected to give a more accurate picture of the historical facts than a comparison of modern French with modern German. Such comparisons are dependent on preliminary comparative studies (see Paragraph 15), and are limited by the fact that reconstructed forms for the entire list are seldom available.

COGNATE COUNT

15. When the word lists have been compiled, the next step is to compare the words of the two lists in order to ascertain how many of the pairs of words are probable cognates (Swadesh, 1950, pp. 157-158). True cognates are developed from the same word in a common parent language, and only true cognates are conclusive evidence of genetic relationship. The most accurate estimate of whether or not the pairs of words in a given comparison are cognate is arrived at by the careful use of the comparative method in reconstructing the proto-language. The major assumption of the comparative method is that while the phonemes of the parent language develop differently in the different daughter languages, the development

is consistent in each kind of linguistic environment within each daughter language. The investigator working on reconstruction matches the words of two (or more) languages by similarity of form and meaning. The phonemes in the same relative position in both members of a matched pair are compared—as initial consonant with initial consonant. If the two languages are related, the same pairs of phonemes will occur in many pairs of words (e.g., many words in language A beginning with *t*ᵛ may be matched in language B by words of similar meaning which begin with *č*). Each such recurring pair of phonemes is assumed to represent a different phoneme or allophone of the common parent language. The investigator on the basis of his data postulates what phoneme is represented by each pair. He also postulates the phonemic system of the parent language and on this basis reconstructs the probable form of the morphemes from which the observed forms in the daughter languages have developed. A full discussion of this method is beyond the scope of this paper, but the interested student should read Bloomfield (1933, pp. 297-320) and Pike (1950). (For a listing of additional sources, see Pike, 1950, bibliography.)

16. When detailed comparative studies are not available, probable cognates can be estimated by an "inspection method," which, although cruder and subject to a greater margin of error, can be used for time depth estimates. The careful use of the following procedures will in general discover the pairs of words which may be considered as probable cognates within a margin of error not great enough to invalidate the method or render the results useless, even though in any one particular instance the conclusion might not reflect the actual historical facts. (Fairbanks [1955] has experimented with an "inspection method" [the term is his], testing the number of dissimilar cognates and similar noncognates in eight comparisons within Indo-European. His criteria were somewhat less strict than those suggested is this paper. For example he ignored vowels, he required agreement in only two consonants of each word, and he made no provision for regularly recurring correspondences [criterion d of this paper]. In his experiment two of the eight cases showed considerable skewing because of cognates which were not similar [pp. 118-119]. This does not completely invalidate the method, but it shows the need of caution especially in deeper time depths. Both Fairbank's experiment and Taylor's work on Arawak [see Taylor and Rouse, 1955, p. 106, in which Taylor uses criteria more strict than those presented here] imply that the skewing from the use of the inspection method rather than careful reconstruction tends to be in the direction of overestimation of time depth, since after long divergence, cognates frequently lose much of their similarity.) The procedures are based in part on the improbability of the chance occurrence of the same sequence of phonemes with the same meaning in two different languages, and in part on the assumptions of comparative linguistics discussed in Paragraph 15.

17. *Procedure 1.* Register as probable noncognates the words which are similar because one language has borrowed from the other, or because both have borrowed from a common source. Borrowings from a common source are recognizable if the forms are very similar to a word of the same or similar meaning in a language which is known to be unrelated, but with which there has been cultural contact. The Mexican Indian languages of Mazatec and Ixcatec, for example, are clearly not closely related to the Indo-European Spanish, but for some centuries, Spanish has been the official language of Mexico. Therefore such words as Mazatec *ni⁴ma⁴* and Ixcatec *ʔa³ni¹ma³e³* 'heart' are registered as noncognate because of the strong probability that they are common borrowings from Spanish *anima* rather than descendents of a native word in their common parent language.

Borrowings of related languages from each other or from a closely related common source may be more difficult to detect. In comparing the Huautla and San Miguel dialects of Mazatec, for example, the only evidence that the San Miguel word *nʔaⁱ³* 'father' is a borrowing and not a true cognate with the Huautla word *nʔaiˀ³* 'father' is the fact that the vowel cluster *ai* occurs in the San Miguel dialect only in a limited number of religious terms, whereas it is normal in the Huautla dialect (Gudschinsky, 1955, p. 148). Such clues may indicate some, though probably not all, of the borrowings from related languages or dialects.

In languages whose probability of close relationship is small, all identical or very similar

words are suspect as loan words unless clearly proved otherwise (see Paragraph 20, criterion a). The apparent closeness of the dialects as ascertained by lexicostatistical methods will be greater in proportion to the number of undiscovered loans that are registered as cognates. The probability, however, is that in most cases the number of such loans will not be great enough to seriously skew the results.

18. *Procedure 2.* Isolate the equivalent morphemes in each pair of words. If equivalent morphemes are not isolated, the investigator may be misled by the complexity of the words he is comparing. The similarity of affixes marking person, number, class, aspect, etc., may obscure the fact that the basic stem morphemes are not true cognates. For example, the person marker *-le*^{4} in the forms *me*^{3}*-le*^{4} (Huautla dialect of Mazatec) and *me*^{3}*hę*^{3}*-le*^{4} (San Mateo dialect of Mazatec) 'he wants' is irrelevant to the comparison of the stems meaning 'want.' If both members of a pair of words are compounds, one pair of the constituent morphemes may be cognate even though the words as a whole are not cognate. For example, Ixcatec *ʔa*^{3}*yi*^{32}*ʔe*^{3} and Mazatec *nʔo*^{1}*yʔe*^{4} 'guts' are not cognate in spite of the very similar *yi*^{32}*e*^{3} and *yʔe*^{4} since these are the morphemes meaning 'dung'; the morphemes which distinguish between 'dung' and 'guts' are *ʔa*^{3}- 'skin' and *nʔo*^{1} 'rope' and are clearly not cognate. (For a further illustration of the need for isolating equivalent morphemes see Taylor and Rouse, 1955, p. 107.)

If the investigator finds it impossible to isolate all of the morphemes in the languages he is comparing, he should proceed with the best guess he can make from the data available to him, recognizing that the comparing of nonrelevant morphemes may cause him to register a number of false cognates which will tend to skew final results in the direction of lesser time depth and closer relationship than is the true hisorical fact. (See Paragraph 30 for an illustration of such skewing in the comparison of Ixcatec and Mazatec.) The increased margin of error from failure to identify morphemes is not so great as to invalidate the method if the results are used with caution, and not treated as absolutes.

19. *Procedure 3.* Test the pairs of equivalent morphemes isolated by procedure 2 to determine whether or not they are sufficiently similar to be considered probable cognates. This testing is done by comparing the phonemes or phoneme clusters occurring in comparable position within the equivalent morphemes. For example, in comparing Ixcatec *cu*^{3} with Mazatec *co*^{3} 'say,' *c* is compared with *c* and *u* is compared with *o*; in comparing Ixcatec *ku*^{3} with Mazatec *kao*^{4} 'and,' *k* is compared with *k* and *u* is compared with *ao* since *ao* occurs in the position comparable to the *u*; in comparing Ixcatec' *ʃu*^{3}*wa*^{3} with Mazatec *nʧoa*^{21} 'come,' *ʃ* is compared with *nʧ* and *uwa* is compared with *oa*. (Tone is ignored in this example and others in this study because the discussion of the complicated tone problems are beyond the scope of this paper.)

Any pair of equivalent morphemes may be registered as probable cognates if a minimum of three pairs of comparable phonemes or phoneme clusters are found to "agree" according to one or more of the criteria given below. In cases in which one or both members of the pair of morphemes being tested is constituted of fewer than three phonemes, the pair can be considered as probably cognate only if all the phonemes or phoneme clusters of the shorter morpheme of the pair agree with the phonemes or phoneme clusters in comparable position in the other morpheme. (For different sets of criteria for determining probable cognates, see Fairbanks, 1955, and Swadesh, 1954c, p. 308.)

20. *Criterion a.* Identical members of a pair of phonemes occurring in comparable position in a pair of equivalent morphemes may be considered as agreeing except that complete identity between languages whose relationship is suspected of being remote may suggest recent borrowing rather than genetic relationship. (Criterion d, Paragraph 23, may be used to determine whether or not the identity of any given pair of phonemes is in accord with a pattern in the language, or whether it is peculiar to this instance. In the latter case, the morpheme pair should be registered as probable noncognates.)

21. *Criterion b.* Phonetically similar members of a pair of phonemes in comparable position in a pair of equivalent morphemes may be considered as agreeing. "Phonetically similar" here means that the two phonemes of the pair must be sufficiently alike phonetically to render them suspect as possible allophones

of a single phoneme if they occurred in the same language. In general, the members of a pair of phonemes are phonetically similar if they differ in such ways as: the presence or absence of vocal vibration as *t* and *d*; the speed of articulation as *r̄* (pronounced with a quick flap of the tongue) and *t*; a slight variation of tongue position as *t* and *ṭ* (pronounced with the tongue tip curled back), or *i* (pronounced as in 'meat') and ɪ (pronounced with the tongue slightly lower and more lax as in mitt'); the presence of secondary activity modifying one of the sounds as *k* and *kʷ* (pronounced with the lips rounded); the extent of interruption of the air stream as *θ* (pronounced with partial interruption of the air stream) and *t* (pronounced with complete interruption of the air stream). For a fuller discussion of phonetic similarity, see Pike, 1947, pp. 69-71. (This criterion should be used with caution if it yields many agreements which are not substantiated by criterion d.)

22. *Criterion c.* A conditioned member of a pair of phonemes occurring in comparable position in a pair of equivalent morphemes may be considered as agreeing with a phonetically dissimilar member. That is, phonetically dissimilar phonemes agree if their environment is such that it could be considered a conditioning factor responsible for the present phonetic shape of one member of the pair of phonemes even though, arbitrarily, it has not had the same effect on the other member of the pair. For example, in comparing the forms *c̄ɪ⁴kɨ¹* (Huautla dialect of Mazatec) and *c̆a⁴kɨ¹* (San Mateo dialect of Mazatec) 'firewood,' the *ɪ* and *a* are considered as agreeing since it is possible that the *c̆* might have been responsible for the change from *a* to *ɪ* (which is pronounced with the tongue closer to the palate than *a*) in the Huautla dialect, even though the change did not occur in the San Mateo dialect. A discussion of conditioning factors may be found in Pike (1947, pp. 84-96).

23. *Criterion d.* Regularly corresponding members of a pair of phonemes occurring in comparable position in equivalent morphemes may be considered as agreeing even though they are not phonetically similar. By regularly corresponding is meant that the same pair of phonemes or phoneme clusters occur in comparable position in a number of different pairs of equivalent morphemes. For example, the Ixcatec phoneme *ɨ* agrees with the Mazatec phoneme *l* because this pair regularly corresponds in such pairs of morphemes as: Ixcatec *ɨ³wi³* and Mazatec *l³ɨ¹ ³* 'fire,' Ixcatec *ɨu³* and Mazatec *lao⁴* 'rock.'

24. In reading the work of specialists in this field, the reader should bear in mind that they differ in the degree of conservatism in their work. The reader can assess the conservatism and solidity of the work by the application of the criteria suggested in Paragraph 20-23 to the pairs of cognates which the author offers as evidence. The inclusion of a quantity of comparative data which is solid in terms of these criteria indicates that the data are conservative and reliable. If, however, only reconstructed forms (marked with an asterisk) are given, without careful documentation, the reader should realize that the proposed reconstructions and the conclusions based on them may in fact be of a highly tentative nature, and should not be accepted as conclusively proved. (See also Kroeber, 1955, p. 97.)

. . . . . .

29. *In Summary.* A total of 192 pairs of words in Ixcatec and Mazatec were compared in Paragraph 28. (Eight of the original list of words were lacking in one or the other of the languages.) Of these 192 pairs, the procedures of Paragraphs 17-23 give a total of 74 probable cognates and 118 probable noncognates. The time depth based on these figures is computed in Paragraphs 34-36; the range of error of the time depth is computed in Paragraphs 44-45; the Ixcatec-Mazatec lexical relationship in dips is computed in Paragraphs 50-51.

30. A careful comparative study would probably result in an estimated 78 cognates and 114 noncognates, since in the author's opinion it is likely that two of the 74 pairs registered as probable cognates are not true cognates, and it is also likely that six of the pairs registered as probable noncognates can be proved to be true cognates on the basis of reconstruction. On the other hand, an investigator completely unacquainted with both languages and unable to isolate the equivalent morphemes and without additional data beyond the 200 word list would be expected to arrive at a total of 72 probable cognates and 120 probable noncognates, since failure to isolate the equivalent morphemes would have resulted in registering four noncognates as probable cognates, but

lack of additional data would have resulted in registering as probable noncognates six pairs which may well be true cognates. See Paragraphs 46-48 for a discussion of the degree to which the time depth estimate is skewed by such inaccurate registering of probable cognates.

COMPUTATION OF TIME DEPTH

31. For use in the time depth formula, the number of probable cognates ascertained by the techniques of Paragraphs 17-23 must be converted to percent of cognates. This is done by dividing the number of probable cognates by the total number of pairs of words compared (Swadesh, 1950, p. 158).

32. Time depth is computed by the formula $t = \log C/(2 \log r)$ (Lees, 1953, p. 117). In this formula t stands for indicated time depth in millenia; C stands for the percent of cognates (Paragraph 31); r stands for the "constant" (also called "index" in Swadesh, 1955, p. 122), that is, the percent of cognates assumed to remain after a thousand years of diverging (Paragraph 8). (In the illustrative material in this paper the value .805 has been used for r, following Lees [1953, p. 119].) Log means "logarithm of" so that log C means the logarithm of the percent of probable cognates registered, and 2 log r means twice the logarithm of the constant.

33. The formula is solved by the following steps: (a) The logarithm of C and the logarithm of r are ascertained from Table 1. (For any who may be rusty on the use of logarithms, the following example is given. The logarithm of .38 is .968; it is found at the point where a line from .3 on the vertical scale of Table 1 meets a line from .08 on the horizontal scale. The logarithm of .39 is found at the point where a line from .3 on the vertical scale of Table 1 meets a line from .09 on the horizontal scale. The logarithm of .385 is halfway between these; half the difference between .968 and .942 subtracted from .968 gives .955 which is the logarithm of .385. Table 1 has been included in the text as more convenient to use than a full logarithmic table; it contains only those values of N that are necessary for computing the time depth.)

(b) The logarithm of r is multiplied by two.

(c) The product of the multiplication in (b) is divided into the logarithm of C.

(d) The quotient of the division in (c) is the indicated time depth in millenia. It may be changed to years by multiplying by 1,000.

COMPUTATION OF TIME DEPTH ILLUSTRATED

34. In the comparison of Ixcatec and Mazatec, 74 of the 192 pairs were registered as probable cognates (Paragraph 29). Dividing 74 by 192 gives .385 (38.5 %). This is the value to be used for C in the time depth formula.

35. The formula may now be filled in to read $t = \log .385/(2 \log .805)$. It is solved as follows: (a) The logarithm of .385 is found from Table 1 to be .955. The logarithm of .805 is

TABLE 1. NATURAL LOGARITHMS

N	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.1	—2.303	—2.207	—2.120	—2.040	—1.966	—1.897	—1.833	—1.772	—1.715	—1.661
.2	—1.609	—1.561	—1.514	—1.470	—1.427	—1.386	—1.347	—1.309	—1.273	—1.238
.3	—1.204	—1.171	—1.139	—1.109	—1.079	—1.050	—1.022	—.994	—.968	—.942
.4	—.916	—.892	—.868	—.844	—.821	—.799	—.777	—.755	—.734	—.713
.5	—.693	—.673	—.654	—.635	—.616	—.598	—.580	—.562	—.545	—.528
.6	—.511	—.494	—.478	—.462	—.446	—.431	—.416	—.400	—.386	—.371
.7	—.357	—.342	—.329	—.315	—.301	—.288	—.274	—.261	—.248	—.236
.8	—.223	—.211	—.198	—.186	—.174	—.163	—.151	—.139	—.128	—.117
.9	—.105	—.094	—.083	—.073	—.062	—.051	—.041	—.030	—.020	—.010

To obtain the natural logarithm of numbers less than .1:
multiply the number by 10 and subtract 2.303 from the ln (natural logarithm) obtained,
or multiply by 100 and subtract 4.605 from the ln obtained,
or multiply by 1,000 and subtract 6.908 from the ln obtained, etc.

NOTE: In the operations described in the text, it is possible to leave out of account the negative value of the logarithms; one negative number divided by another gives a positive quotient.

SOURCE: By permission from *Introduction to Statistical Analysis*, by Wilfrid J. Dixon and Frank J. Massey, Jr., copyright 1951, McGraw-Hill Book Company, Inc.

found to be .217. (b) The product of 2 x .217 (that is 2 log r) is .434. (c) The quotient of .434 (2 log r) divided into .955 (log C) is 2.200; that is, the indicated time depth, t, for Ixcatec-Mazatec is 2.2 millenia or (multiplied by 1,000) 2,200 years.

36. The indicated time depth for Ixcatec-Mazatec computed in Paragraph 35 may be stated in either of the following ways: Ixcatec and Mazatec are estimated to have been a single homogeneous language 2,200 years ago; Ixcatec and Mazatec are estimated to have begun to diverge from a common parent language about 245 B.C. (In the computations given here for illustrative purposes, the time depths, and the dates arrived at by subtracting the time depth from the present date, are not rounded off. It should be noted, however, that the range of error computed in Paragraphs 44-45 indicates that these dates must be taken at best as an approximation somewhere within a few years of correct. The dates have no significance in terms of single years or even decades.)

COMPUTATION OF RANGE OF ERROR

37. It is exceedingly improbable that any two successive random samplings of the basic vocabulary of a pair of languages would yield exactly the same percent of probable cognates. For this reason it is necessary to qualify the statement of time depth in such a way as to give an estimate of its accuracy. The usual way of qualifying a time depth statement is to state it as a range of years rather than as a specific number of years, and to state the degree of probability (or level of confidence) at which the range of years was computed. For example the time depth for Mazatec Ixcatec may be stated as 2,200 years $\pm$ 200 years at 7/10 confidence level (see end of Paragraph 45). (The computation of range of error is based on the assumption that all changes in the basic vocabulary are random, producing a "normal curve.")

38. Statistical methods permit computation of range of error at any level of confidence or probability. Computations are usually made, however, at one of three levels: "standard error" which is 68 % confidence level; (For convenience, standard error will be referred to as 7/10 confidence level) (see Paragraphs 39 and 44); 9/10 (90 %) confidence level; or 5/10 (50 %) confidence level which is also called "probable error." The higher the level of confidence (i.e., the more certainly the true answer lies within the range cited) the wider the range of years. Narrowing the range of years lessens the probability that it includes the true answer.

39. The first step in computing range of error at any level of confidence is the computation of "standard error" (7/10 confidence level). Standard error is computed by the formula $\sigma = \sqrt{C(1-C)/n}$ (Lees, 1953, p. 124, formula 11). In this formula σ stands for standard error in terms of percent of cognates; C means the percent of cognates (see Paragraph 34—this is the same C used in working the time depth formula); n means the number of pairs of words compared. The formula is solved by the following steps: (a) C is subtracted from 1. (b) The remainder of the subtraction in (a) is multiplied by C. (c) The product of the multiplication in (b) is divided by n. (d) The square root of the quotient of the division in (c) is found. (e) The square root found in (d) is the range of error of the percent of cognates at the 7/10 confidence level.

40. Standard error in years is computed by the following steps: (a) The range of error of the percent of probable cognates (found in step (e) of Paragraph 39) is added to C (found in Paragraph 31). (b) The sum of the addition in (a) is worked through the time depth formula exactly as the original C was (Paragraph 32). (c) The new time depth obtained from (b) is subtracted from the original time depth as computed in Paragraph 32 to give the number which is added to and subtracted from the original time depth as computed in Paragraph 32 to give the range of error in years at 7/10 confidence level. (The range of error at 9/10 confidence is obtained by multiplying the standard error of the percent of cognates [found in Paragraph 39] by 1.645 [Dixon and Massey, 1951, Table 4]. The product of this multiplication is the range of error, at 9/10 confidence level, for the percent of cognates. From it, the range of error in years at the 9/10 confidence level can be computed by the same steps used for the computation of the range of error at 7/10 confidence level (Paragraph 40]. The range of error at 5/10 confidence level is obtained by the same steps, using the figure .674 instead of 1.645.)

41. Note that standard error, and therefore any range of error, is larger if the number of

comparisons made is small, but decreases as the number of cases increases because there is division by the number of cases. This makes it important to use a list of words of sufficient length (Lees, 1953, p. 126).

42. An improved word list and more careful collection of data and ascertaining of probable cognates will reduce the actual error, but these will not show up in this method of computing the range of probable error since the accuracy of the investigator cannot be included in the formula. Lexicostatistics operates admittedly with a wide margin of error due to inaccuracy in choice of words, mistakes in determining cognates, etc. This is the price of using the method at all, and is legitimate if one does not abuse it by relying on it for a degree of accuracy that is not basically possible.

43. In very deep time depths where the percent of cognates is small the choice of a single false cognate or the rejection of a single true cognate may make considerable difference in the resulting date (Swadesh, 1953a, p. 41). If, for example, in a list of 200 comparisons there is only one cognate (.5%) the estimated time depth is 12.2 millenia, but if there are two cognates (1%) the time depth is 10. 6 millenia. This is a difference of sixteen centuries dependent on the recognition of a single cognate.

COMPUTATION OF RANGE OF ERROR ILLUSTRATED

44. The range of error at 7/10 confidence level can now be computed for the Ixcatec Mazatec time depth by the formula $\sigma = \sqrt{C\,(1\text{-}C)/n}$ as follows (see Paragraph 39 for the steps followed here): (a) The percent of cognates computed in Paragraph 34 is .385. This number subtracted from 1.000 leaves a remainder of .615 (1-C). (b) .615 multiplied by .385 gives a product of .236775 [C (1-C)]. (c) .236775 divided by 192 (the number of pairs of words compared) gives a quotient of .0012332 [C (1-C)/n]. (d) The square root of .0012332 is .03511 [$\sqrt{C\,(1\text{-}C)/n}$]. (The simplest way to find square root is by reference to a manual of mathematical tables.) This is rounded off to give a standard error at 7/10 confidence level of .035.

45. The range of error in years, at 7/10 confidence level, is computed as follows (following the steps outlined in Paragraph 40): The range of error computed in Paragraph 44 (which is the range of error of the percent of cognates) is added to the original percent of cognates computed in Paragraph 34; that is, .385 plus .035 is .42. (b) This new C is worked through the time depth formula $t = \log C/(2 \log r)$; $t = \log .42/2 \log .805$; $t = .868/.434$; $t = 2.000$ millenia or 2,000 years. (c) The new time depth is subtracted from the time depth computed in Paragraph 35 that is 2,200 years minus 2,000 years is 200 years. (d) The range of error at 7/10 confidence level may now be stated in any of three ways: Ixcatec and Mazatec were a single homogeneous language $2{,}200 \pm 200$ years ago; Ixcatec and Mazatec were a single homogeneous language between 2,000 and 2,400 years ago; Ixcatec and Mazatec began to diverge from a common parent language between 445 B.C. and 45 B.C.

From the standard error the range of error at 9/10 confidence level is computed as $2{,}200 \pm 324$ years. The range of error at 5/10 confidence level is $2{,}200 \pm 140$ years (see Paragraph 40).

46. The percent of cognates likely to be verified by comparative study, and the percent of probable cognates likely to be registered by a person with no knowledge of the two languages involved are given in Paragraph 30. At this point we are ready to work these two estimates through the time depth formula and from the results to estimate the probable degree of skewing of time depth figures due to weakness in the criteria or to the inexperience of the investigator.

47. The more conservative estimate is 78 probable cognates (rather than the 74 probable cognates on which the illustration has so far been based). 78 probable cognates out of 192 comparisons is .406 (40.6 %). Worked through the time depth formula (Paragraphs 32-33) this gives an estimated time depth of 2,078 years. The range of error computed at 7/10 confidence level is .035 (computed according to the steps in Paragraph 39) or 191 years (following the steps of Paragraph 40). This makes the most conservative estimate for the time of Mazatec Ixcatec divergence $2{,}078 \pm 191$ or 1,887-2,269 years ago. Note that the figure 2,200 years (Paragraph 36) obtained by the criteria of Paragraphs 20-23 is within this range.

48. The least accurate estimation of cognates, that arrived at by the use of the criteria suggested in this paper, by an investigator without

sufficient knowledge of the language to isolate the equivalent morphemes, without help from the comparative method, and without data beyond the 200 word list, is 72 probable cognates (Paragraph 30). This is .375 (37.5 %) and gives a time depth of 2260. Note that this figure also is within the range of error, at 7/10 confidence level, of the most conservative estimate (Paragraph 47).

49. On the basis of Paragraphs 47 and 48, it is evident that in this particular comparison, the result arrived at by the use of the criteria in this paper are only very slightly skewed from the results arrived at by the use of the more conservative methods. In other comparisons the skewing may be greater, but the investigator can, in general, estimate the direction of the skewing, and take account of it in assessing the reliability of his results.

DIPS

50. As has been demonstrated, the dating arrived at by lexicostatistical techniques is very tentative, and can be seriously misleading to anyone who assumes that the dates are absolutes in terms of years or months, and uses them without due caution. For this reason it may be convenient to consider the data in terms of dips (i.e., degrees of lexical relationship) rather than in terms of historical dates, so that the relative lexical relationships can be discussed apart from any implication of absolute time (Gudschinsky, 1955, pp. 141-142) which may be more confusing than helpful. The dip expresses a true degree of objective lexical relationship even though borrowing or other factors has destroyed the time relationship. A knowledge of this present relationship is invaluable in practical decisions regarding homogeneity of speech areas for vernacular schools, production of literature, etc.

51. The formula for computing lexical relationship in dips is $d = 14 (\log C/2 \log r)$. Having once worked the time depth formula, however, the results may be converted to dips by multiplying the time in millenia by 14, or the time in years by .014. In the Ixcatec Mazatec example used in this study, the lexical relationship expressed in Paragraph 36 as 2,200 years may be expressed as 30. 8 dips.

Similarly, the range of error in years may be converted by multiplication to range of error in dips. The range of error at 7/10 confidence level is given in Paragraph 45 as 200 years. Multiplied by .014 this gives a range of error of 2.8. dips; that is to say, at 7/10 confidence level, the Ixcatec Mazatec relationship is 30.8 $\pm$ 2.8 dips.

52. Swadesh has suggested a classification of dialects, languages, stocks, and phylums on the basis of lexicostatical results (1954c, p. 326), as follows:

Term	*Divergence Centuries*	*Cognate Percent*
language	0-5	100-81
family	5-25	81-36
stock	25-50	36-12
microphylum	50-75	12-4
mesophylum	75-100	4-1
macrophylum	over 100	less than 1

(Swadesh has used .81 as the constant in determining the value in centuries of the various percents.) These labels may be defined in terms of dips as: language, 0-7 dips; family, 7-35 dips; stock, 35-70 dips; microphylum, 70—105 dips; mesophylum, 105-140 dips; macrophylum, more than 140 dips.

This particular classification is, of course, still tentative. Its empirical usefulness with a large number of languages remains to be demonstrated. But without question, the quantified data resulting from this technique makes possible a more objective classification of lexical relationships than has hitherto been possible (Swadesh, 1950, pp. 162-163).

THE VALUE OF LEXICOSTATISTICS

53. For the anthropologist and historian, the lexicostatistical data suggest the order of the development of languages and dialects. That is, by studying a number of pairs of languages or dialects within a related group, or within a dialect area, those pairs which show greatest time depth are assumed to be representative of older splits in the dialects, and those showing lesser time depth show more recent splits so that a progressive splitting is implied (Gudschinsky, 1955). This suggested order of splitting may help in correlating the linguistic data with known or suspected migrations, cultural developments, etc.

54. The lexicostatistical data also imply the geographical location and cultural contacts

of ancient dialects, since the dialects were presumably relatively homogeneous until the time at which the evidence shows the beginning of their divergence. Then the dialects closest linguistically must have been closest geographically and longest in cultural contact. Such linguistic geographical relationships have been charted by Swadesh (1950, pp. 164-167) and Hirsch (1954). (For an extensive discussion of time depth and geographical location see Kroeber [1955]. For use of the principles of Paragraphs 53 and 54 see Taylor and Rouse, 1955.)

55. In using lexicostatistical data, it must be remembered that even when further experiment with the word list and the constant make possible a greater degree of accuracy, no individual study will be more accurate than the data available and the care used in ascertaining the probable cognates. Also, regardless of the degree of accuracy possible in determining when certain languages or dialects diverged from each other, it is not possible to determine by lexicostatistics what language was spoken by the people responsible for the artifacts found in any given place (Swadesh, 1954b; Kroeber, 1955, p. 104).

56. The archaeologist or nonlinguist who is curious to try this material is urged to do so. All that he needs beyond what is given here is the historical records or informants from which to obtain the lexical data.

REFERENCE NOTE

The problems and literature of lexicostatistics and glottochronology are discussed generally in Hymes (1960a, 1960e), in Bergsland and Vogt (1962) and in the comments to these articles (esp. Hymes, 1962b) by a variety of scholars. For recent comment, see also Hoijer (1961). For recent work of new scope, see Dyen (1962a, 1962b, 1962c) and Carroll and Dyen (1962), and cf. Elmendorf (1962b). Elmendorf (1962a), Diebold (1960), and Dyen (1962b) restate Salish relationships discussed in Swadesh (1950) and indicate the importance a well-worked body of data may acquire. For recent work on lexicostatistics, apart from glottochronology, see also Cowan (1959), Ellegard (1959), Gleason (1959), and Kroeber (1960a).

References not in the general bibliography:

CARROLL, JOHN B., and ISIDORE DYEN
1962. High Speed Computation of Lexicostatistical Indices. *Lg.*, *38:* 274-278.

COWAN, H. K. J.
1959. A Note on Statistical Methods in Comparative Linguistics. *Lingua*, *8:* 233-246.

DIXON, WILFRID J., and FRANK J. MASSEY, JR.
1951. *Introduction to Statistical Analysis.* New York: Wiley.

DYEN, ISIDORE
1962a. The Lexicostatistical Classification of the Malayopolynesian Languages. *Lg.*, *38:* 38-46.
1962b. The Lexicostatistically Determined Relationship of a Language Group. *IJAL*, *28:* 153-161.
1962c. Lexicostatistically Determined Borrowing and Taboo. *Lg.*, *38:* 60-66.

ELLEGARD, ALVAR
1959. Statistical Measurement of Linguistic Relationship. *Lg.*, *35:* 131-156.

ELMENDORF, W. W.
1962a. Lexical Innovation and Persistence in Four Salish Dialects. *IJAL*, *28:* 85-96.

1962b. Lexical Relation Models as a Possible Check on Lexicostatistic Inferences. *AA, 64:* 760-770.

FAIRBANKS, GORDON H.
1955. A Note on Glottochronology. *IJAL, 21:* 116-124.

FERNANDEZ DE MIRANDA, MARIA TERESA
1951. Reconstrucción del Protopopoloca. *Revista Mexicana de Estudios Antropológicos, 12:* 61-93.

GREENBERG, JOSEPH H., and MORRIS SWADESH
1953. Jicaque as a Hokan Language. *IJAL, 19:* 216-222.

GUDSCHINSKY, SARAH C.
1955. Lexico-statistical Skewing from Dialect Borrowing. *IJAL. 21:* 138-149.

HIRSCH, DAVID I.
1954. Glottochronology and Eskimo and Eskimo-Aleut Prehistory. *AA, 56:* 825-838.

HOCKETT, CHARLES F.
1953. Linguistic Time-Perspective and Its Anthropological Uses. *IJAL, 19:* 146-152.

LEES, ROBERT B.
1953. The Basis of Glottochronology. *Lg., 29:* 113-127.

SWADESH, MORRIS
1951b. Kleinschmidt Centennial III: Unaaliq and Proto Eskimo. *IJAL, 17:* 66-70.
1953a. Mosan I: A Problem of Remote Common Origin. *IJAL, 19:* 26-44.
1953b. Comment on Hockett's Critique. *IJAL, 19:* 152-153.
1953c. The Language of the Archaeologic Huastecs. *Notes on Middle American Archaeology and Ethnology, 4:* 223-227.
1954a. On the Penutian Vocabulary Survey. *IJAL, 20:* 123-133.
1954b. Time Depths of American Linguistic Groupings. With Comments by G. I. Quimby, H. B. Collins, E. W. Haury, G. F. Ekholm, and Fred Eggan. *AA, 56:* 361-377.

TAYLOR, DOUGLAS, and IRVING ROUSE
1955. Linguistic and Archaeological Time Depth in the West Indies. *IJAL, 21:* 105-115.

Dell Hymes
Language in Culture and Society. A Reader in Linguistics and Anthropology.
New York: Harper & Row
1964