Lexico-Statistic Dating of Prehistoric Ethnic Contacts: With Special Reference to North American Indians and Eskimos

Morris Swadesh

LEXICO-STATISTIC DATING OF PREHISTORIC ETHNIC CONTACTS

With Special Reference to North American Indians and Eskimos

MORRIS SWADESH

PREHISTORY refers to the long period of early human society before writing was available for the recording of events. In a few places it gives way to the modern epoch of recorded history as much as six or eight thousand years agó; in many areas this happened only in the last few centuries. Everywhere prehistory represents a great obscure depth which science seeks to penetrate. And indeed powerful means have been found for illuminating the unrecorded past, including the evidence of archeological finds and that of the geographic distribution of cultural facts in the earliest known periods. Much depends on the painstaking analysis and comparison of data, and on the effective reading of their implications. Very important is the combined use of all the evidence, linguistic and ethnographic as well as archeological, biological, and geological. And it is essential constantly to seek new means of expanding and rendering more accurate our deductions about prehistory.

One of the most significant recent trends in the field of prehistory has been the development of objective methods for measuring elapsed time. Where vague estimates and subjective judgments formerly had to serve, today we are often able to determine prehistoric time within a relatively narrow margin of accuracy. This development is important especially because it adds greatly to the possibility of interrelating the separate reconstructions.

Unquestionably of the highest value has been the development of radiocarbon dating.[1] This technique is based on W. F. Libby's discovery that all living substances contain a certain percentage of radioactive carbon, an unstable substance which tends to change into nitrogen. During the life of a plant or animal, new radiocarbon is continually taken in from the atmosphere and the percentage remains at a constant level. After death the percentage of radiocarbon is gradually dissipated at an essentially constant statistical rate. The rate of "decay" being constant, it is possible to determine the time since death of any piece of carbon by measuring the amount of radioactivity still going on. Consequently, it is possible to determine within certain limits of accuracy the time depth of any archeological site which contains a suitable bit of bone, wood, grass, or any other organic substance.

Lexicostatistic dating makes use of very different material from carbon dating, but the broad theoretical principle is similar. Researches by the present author and several other scholars within the last few years have revealed that the fundamental everyday vocabulary of any language—as against the specialized or "cultural" vocabulary—changes at a relatively constant rate. The percentage of retained elements in a suitable test vocabulary therefore indicates the elapsed time. Wherever a speech community comes to be divided into two or more parts so that linguistic change goes separate ways in each of the new speech communities, the percentage of common retained vocabulary gives an index of the amount of time that has elapsed since the separation. Consequently, wherever we find two languages which can be shown by comparative linguistics to be the end products of such a divergence in the prehistoric past, we are able to determine when the first separation took place. Before taking up the details of the method, let us examine a concrete illustrative instance.

The Eskimo and Aleut languages are by no means the same. An Eskimo cannot understand Aleut unless he learns the language like any other foreign tongue, except that structural similarities and occasional vocabulary agreements make the learning a little easier than it might otherwise be. The situation is roughly comparable to that of an English-speaking person learning Gaelic or Lithuanian. It has been shown that Eskimo and Aleut are modern divergent forms of an earlier single language.[2] In other words, the similarities be-

---

[1] See Radiocarbon dating, assembled by Frederick Johnson, *Mem. Soc. Amer. Archaeol.* **8**, 1951.

[2] Concrete proof of this relationship has recently been presented in two independent studies: Knut Bergslund, Kleinschmidt Centennial IV: Aleut demonstratives and the Aleut-Eskimo relationship, *Internatl. Jour. Amer. Ling.* **17**: 167-179, 1951; Gordon Marsh and Morris

tween Aleut and Eskimo are not accidental nor are they entirely due to diffusional influences of one language on the other; the nature of the similarities bears out the long-held inference that we have here distinct traces of what was in some earlier epoch a single language. To determine when the earlier single speech community separated into two, we use lexicostatistic calculation based on the percentage of like elements in test vocabularies of Eskimo and Aleut, and we find that about 2,900 years have elapsed since the common period of these now distinct languages.[3] Similar studies of Eskimo indicate that the dialects from Seward Peninsula in Alaska all the way to Greenland diverged from each other within a relatively recent period, not yet accurately calculated but probably in the last several centuries. However, the divergence between this group of dialects and those around the Yukon and at East Cape in Siberia is such as to indicate about 1,000 years of separate development.[4] By the use of these dates and others which can be calculated from the comparison of closely related dialects, one can learn much about the principal migrations and cultural influences in the prehistoric past of the Eskaleuts.

Lexicostatistic data must be coupled with other evidence, including that of archeology, comparative ethnography, and linguistic paleontology. The separate lines of study serve to verify or correct one another and to fill in details of the story. While no full correlated study of Eskimo-Aleut has yet been undertaken, the calculation of the time depth of the stock has already been in part correlated with archeologic evidence. Carbon samples obtained by Laughlin and Marsh from one of the earliest settlements on the Aleutians and subjected to laboratory tests of radioactivity were found to be about 3,000 years old,[5] practically

coinciding with the independently obtained lexico-statistic date. This indicates that the forerunners of the Eskimos and the Aleuts belonged to the same community up to the time when the islands were first settled.

The time depth of Eskaleut may be combined with other relevant considerations to formulate some possibilities as to the prehistoric migrations into the New World. This requires the dating of various linguistic groups on both sides of Bering Strait and the correlation of the dates with the facts of geographic distribution. Since very little of this work has yet been done, we cannot discuss the implications in any completeness. We may note, however, that the Nadene stock seems to show a period of divergence that may be shorter than Eskimo-Aleut,[6] a fact which suggests that the Nadenes rather than the Eskaleuts were the last group to enter America. Both of these stocks are recent enough in the New World to make it logical that their nearest linguistic relatives are to be sought in Eurasia rather than in America. A time depth of no more than 3,000 years makes it likely that such linguistic relatives can be found. This consideration lends plausibility to the case for Eskaleut relationship with Uralaltaic and Indoeuropean and to the case for Nadene relationship with Sinotibetan.[7] A proper examination of these possibilities, using vocabulary statistics in coordination with the other lines of evidence would greatly illuminate the prehistoric peopling of the western hemisphere.

DISCOVERY OF THE CONSTANT

The fact that fundamental vocabulary changes at a constant rate was discovered accidentally, but

Swadesh, Kleinschmidt Centennial V: Eskimo Aleut correspondences, *Internatl. Jour. Amer. Ling.* **17**: 209–216, 1951.

[3] Marsh and Swadesh, *Internatl. Jour. Amer. Ling.* **17**: 169, 1951. The time given in this article was 4,000 years, loosely calculated upon a retention rate of 85 per cent. The correction based on strict calculation at 81 per cent ± 2 per cent gives 2,900 ± 400 years; the way in which the corrected constant was determined is explained farther on in the present article.

[4] Calculated on the percentage agreement reported in Swadesh, Kleinschmidt Centennial III: Unaaliq and Proto-Eskimo, *Internatl. Jour. Amer. Ling.* **17**: 66–70, 1951.

[5] William Laughlin and Gordon Marsh, A new view of the history of the Aleutians, *Arctic* **4**: 75–88, 1951, especially p. 91. The exact carbon date is 3,018 ± 230 years.

[6] See Swadesh, Diffusional cumulation and archaic residue as historic explanations, *Southwestern Jour. Anthropol.* **7**: 1–21, 1951, esp. p. 14. The percentage of correspondence between Athapaskan and Tlingit indicates about 2,000 years of separation. The common period of Nadene would be somewhat older than this.

[7] See L. L. Hammerich, Can Eskimo be related to Indo-European? *Internatl. Jour. Amer. Ling.* **17**: 217–223, 1951; Robert Shafer, Athapaskan and Sino-Tibetan, *Internatl. Jour. Amer. Ling.* **18**: 12–19, 1952. Both these papers have the shortcoming of dealing with a single branch of a compared stock rather than the entirety. Although the results are nevertheless significant, it would be better to deal with Nadene as a whole, including Tlingit and Haida as well as Athapaskan, and Eskaleut rather than merely Eskimo. In the latter case, there is every reason to suppose that an even closer relationship will be found with Uralaltaic, as suggested long ago by R. Rask (see William Thalbitzer, The Aleut language compared with Greenlandic, *Internatl. Jour. Amer. Ling.* **2**: 40–57, 1921–1923, esp. p. 40).

carbon dating was the specific stimulus for the research which brought it forth. For it was the achievements of radiocarbon dating which led the author four years ago to undertake studies of rate of vocabulary change. His only expectation at that time was to find an approximate maximum rate of change. At the Viking Fund Supper Conference for Anthropologists on March 12, 1948, he presented a paper on "The Time Value of Linguistic Diversity," which showed the value of such an approximate measure of vocabulary change. The abstract of that paper, reproduced in mimeographed form by the Viking Fund, reads in part as follows:

If comparative linguistics have not been systematically exploited for time scale in studies of culture prehistory, it is probably because comparative linguists have generally been satisfied with vague relative chronologies suitable for their own purposes and unconnected with other aspects of culture history. However, the basis exists for absolute chronology— approximate, of course, but very valuable when added to other lines of evidence.

"The greater the degree of linguistic differentiation within a stock, the greater is the period of time that must be assumed for the development of such differentiation" (Sapir, *Time Perspective*, p. 76). Though the rate of change is not constant, there is definitely a maximum. Also a normal rate can probably be determined that holds good when applied to extended time periods. These questions can be studied with controlled material, available in such stocks as Indo-European, Semitic and Sino-Tibetan, where historical materials are abundant and go back as far as 6000 years. The time-diversity scales thus established can then be used in the many other situations where historical materials are lacking but where comparative linguistics has proven the common origin of two or more languages.

An index of vocabulary diversity should be based upon lists of words (or morphemes) that are relatively neutral in their cultural implications, so as to avoid the transforming effects that accompany rapid cultural change. . . .

Each application of an index gives a minimum indicated time. Where two indices give different minimum times, it is obvious that the longest minimum is the nearest correct.

The author's assumption at this time was that "the rate of change is not constant" but "there is definitely a maximum." This belief was based on the well-known fact that languages change slowly, even at their fastest. On the other hand, it was generally believed in linguistics that some languages change much more slowly than others and

in fact practically remain unchanged for thousands of years. It was not till later, after making actual measures of rate of change in supposedly slow-changing languages, that the author discovered this notion to be illusory.

Thus the first study of rate of change in fundamental vocabulary was based upon the idea of finding a useful measure of change other than a constant. On these terms, the first small experiment presented to the Viking Fund Conference consisted of comparing the percentage of agreement in Nootka-Kwakiutl with that in English-German. In the former pair, American Indian languages comprising the Wakashan stock of the northwest coast, there were about 30 per cent of cognate elements in the test vocabulary, while English and German have about 59 per cent in common.[8] This indicates that the Wakashan stock has had a much longer period of divergence than have English and German. Since we know the time in the latter case to be a few centuries more than 1,100 years, we gain some approximate notion of how great is the time depth of Wakashan. In this way historic knowledge serves to clarify prehistory.

In the fall of 1949 the author, with the support of the Phillips Fund of the American Philosophical Society and making use of the excellent collection of Salish manuscripts in the Franz Boas Collection of the Society's Library, undertook an objective classification of the languages in this ramified stock based upon vocabulary similarities. He used a test list of fundamental lexical items along the lines of his earlier work and chose as the unit for the scale of agreement and divergence the percentage of retained words in present-day English as compared with Old English of 1,000 years ago. This was considered at first merely as a convenient concrete unit, as a statistical convenience. However, as the study progressed, it was seen that the results showed remarkable inner consistency. For example, Bella Coola, the most divergent of the Salish languages, showed percentages of agreement with the twenty-five other languages all within the narrow low range of 11 to 23 per cent. In all cases where high percentages of vocabulary agreement were found, the languages are closely similar in structure also and are so situated geographically as to bear out the likelihood that they have separated from each other only in a relatively recent

---

[8] In the conference paper, the author calculated 31 and 65 per cent, with a test list and a scoring technique both slightly different from the present ones.

epoch. Nowhere in all this ramified stock were correspondences found that implied the occurrence of changeless or nearly changeless languages. On the contrary, the indications were that all the languages had changed their essential vocabularies at something like a constant rate.

This development in the Salish study demonstrated the need for further studies of languages of known history. By the time that "Salish Internal Relationships" appeared in print,[9] though not soon enough to be incorporated into the article, the author and several colleagues had made a series of such studies. Gradually the evidence accumulated showing that universal everyday vocabulary, words of the type comprising the test list, changes at a roughly constant rate. Some weaknesses in the test list were disclosed, but the essential truth of the constant shows through despite them. Certain methodological problems were encountered, but none for which the solution has not also been found. These matters are discussed in later sections of this paper.

Valuable work on the statistical analysis of lexicon dating has been done by Robert B. Lees of the University of Chicago.[10] His studies have already provided reassurance on a mathematical basis that the constant is real, that the number of historical cases is already sufficient to eliminate the possibility of chance; and he is now working out the problems of calculating the deviation error of the method. In the meantime new studies of problems of prehistory have been carried out, including those of Eskimo-Aleut already mentioned. Joseph H. Greenberg and the author, aided by a grant from the Columbia University Social Science Research Fund, have used the method with excellent effect in studies of several linguistic stocks in Africa, Australia, and America.[11] In the course of these various studies, some new applications of lexicostatistics were found. These included a technique for distinguishing between archaic residue and diffusional accumulation,[12] and a method for demonstrating remote genetic relationships.[13]

[9] *Internatl. Jour. Amer. Ling.* **16**: 157–167, 1950.

[10] See Robert B. Lees, A method of dating with Lexicon statistics, p. 3 (mimeograph edition of a paper read at the Michigan Linguistic Institute 1951).

[11] Already complete Joseph H. Greenberg, The genetic classification of Australian languages, paper presented at 1951 Annual Meeting of the Linguistic Society of America.

[12] Swadesh, Diffusional cumulation and archaic residue as historical explanations, *Southwestern Jour. Anthropol.* **7**: 1–21, 1951.

[13] Presented as, An experiment in remote comparative

RETENTION RATE

Of the three main aspects of language—sounds, structure, lexicon—the third lends itself best to the requirements of a statistical time index. Though words are readily borrowed, it has long been known that the borrowings take place primarily in the "cultural" part of the vocabulary and that the "intimate" vocabulary resists change. It was not difficult to form a list of about two hundred relatively stable lexical items, consisting of body parts, numerals, certain objects of nature, simple. universal activities. A simple clear-cut criterion of what constitutes a change could be set up, namely, the substitution of a new element from whatever source as the most usual everyday expression of the given notion. In comparing two time periods of a given language or two languages developed out of the same earlier language, the agreement could be stated as a percentage of cognate elements in the total number of items compared. The retention index could be calculated in terms of some convenient time period, say millennia or centuries, so as to enable one to compare the different cases with each other. The first two counts made by the author were present-day English as compared with Old English and modern Spanish as compared with classical Latin. The method of comparison is illustrated here with the first ten items in the list:

Old Eng.	Mod. Eng.	Cognates	Latin	Spanish	Cognates
eall	all	+	omnēs	todos	−
and	and	+	et (or −que)	y	+
dēor	animal	−	animal	animal	+
æsc	ashes	+	cinis	cenizos	+
æt	at	+	in	a	−
bæc	back	+	dorsum	espaldo	−
fūl	bad	−	malus	malo	+
rind	bark	−	cortex	corteza	+
belg	belly	+	venter	vientre	+

The semantic criterion is applied strictly. Although the word "deer" is still used in modern English, it is no longer the general word for "animal" but refers to a definite category of animals and a new word of Romance origin has taken over its old function. Similarly modern "rind" no longer refers to the bark of a tree, and "bad" rather than "foul" is now the everyday expression

linguistics, in the 1951 meeting of the Linguistic Society of America. The method will be published in connection with a paper on the Mosan stock, now in preparation.

corresponding to Old English fūl. Consequently these cases are scored as non-correspondences. On the other hand, normal phonetic changes (e.g., belg to belly) and structural modifications (e.g., the use of a new affix, as in cenizos) are disregarded. Applying this procedure to the entire test list, the agreement between Old and modern English was found to be 85 per cent. Between modern Spanish and classical Latin, the correspondence was 70 per cent. The elapsed time period in the first case was about 1,000 years, in the second case 2,000 years. Now, if after 1,000 years 85 per cent of the original vocabulary still remains in the same function, then in a second thousand years a similar retention rate would give 85 per cent of the 85 per cent still existing at the beginning of this second period. In other words, 2,000 years at the English rate would leave 72 per cent, a little more than the Spanish retention after such a period. The Spanish retention corresponds to a rate of a little less than 84 per cent per 1,000 years. The rate of retention is thus practically identical in these two cases.

In order to determine whether the rate of retention is always constant, it is necessary to examine a number of cases where the vocabularies of two periods of the same language are known and the elapsed time is also known. A number of suitable instances are available for study. Since the time intervals are generally not simple multiples of each other, it is convenient to use logarithms to reduce all the cases to a standardized time period such as 1,000 years. The mathematics can be expressed by the formula:

$$\log r = \log c \div t.$$

That is, logarithm of the retention per 1,000 years in per cent equals logarithm of common vocabulary per cent divided by number of time periods.

Tests of rate of retention have been made for several languages by different scholars with the following results.[14]

	% per 1,000 years
Middle Egyptian 2100-1700 B.C. to Coptic 300–500 A.D. (C. Baer) (calculated as 23 centuries)	76
Classic Latin 50 B.C. to present-day Romanian (E. Cross)	77
Old High German 850 A.D. to present-day German (G. J. Metcalf and R. D. Lees)	78
Classic Chinese 950 A.D. to modern colloquial North Chinese (C. Y. Fang)	79
Latin of Plautus 200 B.C. to French of Molière 1650 A.D. (D. A. Griffin)	79

	% per 1,000 years
Dominica Carib of 1650 to present-day (D. Taylor and M. Swadesh)	80
Classic Latin 50 B.C. to present-day Portuguese (E. Cross)	82
Koiné to present-day Cypriote (E. Hamp)	83
Koiné to present-day Athenian (E. Hamp)	84
Classic Latin 50 B.C. to present-day Italian (E. Cross)	85
Old English 950 A.D. to present-day English (M. Swadesh)	85
Latin of Plautus 200 B.C. to Spanish of 1600 A.D. (D. A. Griffin)	85

The reasons for variation in the rate of retention need to be considered in detail, but the amount of variation, from 76 to 85 per cent, is relatively small. For the purposes of studying reasons for variation it would be desirable to obtain the counts for a much larger number of cases, but the number of examples given here is already sufficient to eliminate the possibility of sheer coincidence in the close agreement of the retention indices.

THE TEST VOCABULARY

The lexical test list used for studying rate of change consisted of 215 items of meaning expressed for convenience by English words. In some cases, where the English word is ambiguous or where the English meaning is too broad to be easily matched in other languages, it is necessary to specify which meaning is intended, and this is done by means of parenthetic additions. If it is understood that normal everyday meanings rather than figurative or specialized usages are to be thought of, complicated notes are not necessary. The list, minus 15 items recommended for omission and with one other change, is as follows:

all (of a number), and, animal, ashes, at, back (person's), bad (deleterious or unsuitable), bark (of tree), because, belly, berry (or fruit), big, bird, to bite, black, blood, to blow (of wind), bone, breathe, to burn (intrans.).

child (young person rather than as relationship term), cloud, cold (of weather), to come, to count, to cut, day (opposite of night rather than time measure), to die, to dig, dirty, dog, to drink, dry (substance), dull (knife), dust, ear, earth (soil), to eat, egg, eye.

to fall (drop rather than topple), far, fat (organic substance), father, to fear, feather (larger feathers rather than down), few, to fight, fire, fish, five, to float, to flow, flower, to fly, fog, foot, four, to freeze, to give.

good, grass, green, guts, hair, hand, he, head, to hear, heart, heavy, here, to hit, to hold (in hand), how, to hunt (game), husband, I, ice, if.

in, to kill, to know (facts), lake, to laugh, leaf, left (hand), leg, to lie (on side), to live, liver, long, louse, man (male human), many, meat (flesh), mother, mountain, mouth, name.

narrow, near, neck, new, night, nose, not, old, one, other, person, to play, to pull, to push, to rain, red, right (correct), right (hand), river, road (or trail).

root, rope, rotten (especially log), to rub, salt, sand, to say, to scratch (as with fingernails to relieve itch), sea (ocean), to see, seed, to sew, sharp (as knife), short, to sing, to sit, skin (person's), sky, to sleep, small.

to smell (perceive odor), smoke (of fire), smooth, snake, snow, some, to spit, to split, to squeeze, to stab (or stick), to stand, star, stick (of wood), stone, straight, to suck, sun, to swell, to swim, tail.

that, there, they, thick, thin, to think, this, thou, three, to throw, to tie, tongue, tooth (front rather than molar), tree, to turn (change one's direction), two, to vomit, to walk, warm (of weather), to wash.

water, we, wet, what? when? where? white, who? wide, wife, wind, wing, to wipe, with (accompanying), woman, woods, worm, ye, year, yellow.

Sixteen items used in the studies but which are unsatisfactory for many language groups are: brother, sister, six, seven, eight, nine, ten, twenty, hundred, clothing, to cook, to dance, to shoot, speak, to work, to cry. One item, to speak, has been replaced by a near synonym of higher normal stability, to say; and one word, heavy, has been added to bring the lists to an even 200. There may be reason for questioning some other items in the list, but the more serious defects are probably contained in the seventeen items now recommended for deletion or change. For the time being it is recommended that studies continue either with the original list or with the slightly modified list so that new results will be comparable with those previously obtained. The same applies to possible additions, since any major lengthening of the test vocabulary would require recalculating the index of the constant.

Suitable items for a test list must be universal and non-cultural. That is, they must refer to things found anywhere in the world and familiar to every member of a society, not merely to specialists or learned people. Moreover, they must be easily identifiable broad concepts, which can be matched with simple terms in most languages. Of course, it would be impossible to devise a list which works perfectly for all languages, and it must be expected that difficult questions will sometimes arise. This can, however, be very simply met by omitting the troublesome item when necessary. The rules for filling in the list for each language may be stated as follows: (a) Try to find one simple equivalent for each item by disregarding specialized and bound forms and the less common of two equivalents. (b) Use a single word or element rather than a phrase, even though the meaning may be broader than that of the test item. (c) Where it is impossible to find a single equivalent, omit the form.

It is doubtless possible to devise a better test list than the present one, but from the author's own experiments along these lines it is not too easy. Many notions seem suitable but are difficult to match in some languages because of structural peculiarities; thus the locative relations, like behind, above, beneath, etc. Many promising items turn out to have finely subdivided ranges of meaning in certain cultures, e.g., to work. It is therefore suggested that an objective method may solve the problem. A long list of possible items may be tried for a number of languages chosen for their diversity, and each item scored as to how often it can be easily matched. Only those which can be matched unambiguously in most languages are suitable. The stability of the items also needs to be objectively tested by noting how often and for how long they are retained in a number of historically known situations. A stability score for individual items could be calculated, and this score taken into account in constructing the improved test list. Presumably the variation in the index of retention would be reduced by having a better test list, but we do not know to what extent the present variability can be reduced by such improvements.

An obvious way to improve the test list, if possible is to make it longer. However, once one has two hundred test items, it takes several hundred more to improve the statistical adequacy of the test in any marked degree. Considering the difficulties of finding universally suitable words for a test vocabulary, it can hardly be hoped that a list of more than about three hundred items could be devised. Even this extension would be worthwhile for the purpose of dealing with remote time depths. For instance, where there is only 5 per cent retention, it is distinctly more satisfactory to have fifteen rather than ten actual instances.

DETERMINING TIME INTERVAL

In calculating rate of retention in historic cases, it is important to have correctly dated samples and to be sure that both samples are on the same line of language tradition. In one of our instances

Middle Egyptian and Coptic are compared. The time of Middle Egyptian is given as 2100 to 1700 B.C., Coptic as A.D. 300 to 500. This makes the time interval between the two samples somewhat ambiguous, ranging anywhere from 20 to 26 centuries, and the rate of retention would accordingly vary from 73 to 78 per cent per 1,000 years. The difference is not more marked than this because we are dealing with a long time span. If the problem were one involving 400–500 years, the difference of a single century in the assumed time would make 4 per cent difference in the rate per 1,000 years. We thus see that some of the variation in the index of retention may be due not to differences in the process of linguistic change but to errors in measuring the time span between the compared samples.

To overcome this difficulty, Griffin's approach for French is excellent, since for both his early and late samples he takes the language of single authors, making it possible to date the vocabularies within a couple of generations. In using this device one must take the precaution, as Griffin does, of choosing authors who definitely use the contemporary language of their time. If the authors should be following some archaic literary tradition which prefers words of some centuries before, this would skew the results. Similarly with reference to contemporary speech forms, one must take the language of current popular usage and not a formal or literary style which imitates the usage of earlier times.

Another requirement for measuring rate of change in these cases is that the two samples be in the same line of linguistic development. If the earlier of the two is not the actual forerunner of the second but merely a close relative to it, the divergence already present in the older period will add to that which is due to the passage of the given amount of time. The cases we have used probably involve a minimum of deviation from this cause.

As already suggested, the influence of an error in calculating the time period is most noticeable for a short total period. It is therefore wise to use longer periods as much as possible. However, this would eliminate the use of data from certain areas where ancient records do not exist, for example, the New World. We have so far included one example in our material, that of Dominica Carib for a 300-year period. Other cases can and should be taken but the short-time results should be considered as less accurate, other things being equal, than those covering long periods.

So far our longest historic periods are in the neighborhood of 2,000 years. To test the method in every possible way, it would be desirable to find instances covering much longer periods. They are of course rare, but Babylonian-Assyrian offers at least one case. It would also be good to follow a single language through a series of short periods, comparing these finally with the total change for the full time span, in order to see how the rate of retention varies within a single line of development.[15]

VARIATION

It is to be expected that a certain amount of fluctuation in the calculated rate of vocabulary retention will remain, even after the obvious specific sources of error are eliminated. At least in part this variation is the usual "variance" found in all statistical phenomena. It cannot be eliminated but can be calculated. The present paper does not attempt to set forth this phase of the problem, but instead cites the work of R. D. Lees, who gives the "mean rate constant" on the basis of the present evidence as approximately 81 per cent ± 2 per cent per 1,000 years "where the limit of error is the 9/10's error calculated by a small-sample method.[16] Lees is continuing his work, and will eventually produce a full analysis of the mathematical aspect of lexicostatistics.

In addition to statistical variance, it is quite possible that the constant may be affected within limits by culture historical factors. Of course, every culture change affects the full vocabulary of a language, giving it new terms and modifying the meanings of old ones. Contact between peoples often leads to the acquisition of new concepts and of new words. One of the best illustrations of this is English, which has taken over so many words (especially from Latin and French) in the last 2,000 years that about half of the total vocabulary is identifiable as loanwords. However, within the fundamental vocabulary the borrowings are much less, there being only 6 per cent in the test list used in our studies. However, not only is the percentage small but borrowing shows no effect on the English retention rate. The 6 per cent of borrowed elements, accumulated over 2,000 years under cultural circumstances which greatly favored borrowing, exceeds the percentage in any of the

---

[15] Prof. I. J. Gelb of the University of Chicago has projected such a study covering different periods of Assyrian.

[16] *Op. cit.* in footnote 10.

other languages studied, but the retention rate is not correspondingly lower. Thus, a language which borrows some basic words from foreign languages does not necessarily change the total fundamental vocabulary faster than another language whose only changes are the substitution of one native form for another.

Are there other special factors? What we have just observed as to loanwards in English demonstrates that we cannot merely take it for granted that any given cultural influence must skew the rate of change in fundamental vocabulary. Instead, it is necessary to test carefully every hypothesis that may be suggested. For example, Elmendorf suggests that word tabus connected with mourning may have substantially changed the rate of retention of certain Salish Languages.[17] This can be checked by comparing the divergence of Twana, which is known to have had tabus, and that of a language which evidently did not have them, say Columbia, with reference to a distant relative of both of them, Bella Coola. Now, Twana has 15 per cent agreement with Bella Coola; Columbia has only 12 per cent. Quinault, which is a neighbor of Twana and which probably also had the tabuing customs, has 12 per cent agreement with Bella Coola, exactly the same as Columbia. These figures show that language with tabus have not changed any faster than those which lack them.

Until some research has been undertaken to determine causes of variation in the retention index, one tries to discover some pattern in the series of scores reported so far. But no such pattern has yet come to view. For example, what fact of history explains why Spanish should have a higher index than French and Romanian? If the explanation were that French and Romanian are on the margins of the Romance territory, then why should English have a higher retention rate than German in the West Germanic group? The search for such explanations is at present complicated by some uncertainties in the matter of dating the samples and as to the appropriateness of the test list itself. When these uncertainties have been cleared away, it will be possible to return to the problem with a real prospect of proving or disproving the presence of special factors affecting the rate of retention.

[17] William W. Elmendorf, Word taboo and lexical change in Coast Salish, *Internatl. Jour. Amer. Ling.* 17: 205–208, 1951.

CAUSE OF THE CONSTANT

Why does the fundamental vocabulary change at a constant rate? Not only curiosity but also scientific thoroughness impels us to seek an explanation. For, if we have some notion of why a given phenomenon occurs in nature, we are in a better position to direct future researches aimed at extending our knowledge still further. In the present case, we can use the constant for dating prehistory even if we do not know what produces the constant, but we shall be better guided in making our work more accurate and fruitful if we understand its mechanism. The answer is to be found in the nature of language itself.

A language is a highly complex system of symbols serving a vital communicative function in society. The symbols are subject to change by the influence of many circumstances, yet they cannot change too fast without destroying the intelligibility of language. If the factors leading to change are great enough, they will keep the rate of change up to the maximum permitted by the communicative function of language. We have, as it were, a powerful motor kept in check by a speed regulating mechanism.

The specific causes of change in language are of course various. Word tabus, existing in a variety of local customs, lead to the avoidance of words or to limitations on their use. On the other hand, the creation and use of figures of speech are widely encouraged among all human groups. Very important is the influence of new models in the form of neighboring languages and dialects, especially in view of the fact that at least some of the border population is almost certain to be bilingual. Intermarriage, trade, and warfare are some of the specific forms of contact. The influence on language is direct and indirect, since neighboring peoples not only borrow sounds, structural patterns, and words from each other, but adopt, even more readily, concepts, institutions, inventions. In other words, culture change takes place and this in turn calls for the development of new communicative symbols. Even isolated societies, in so far as such exist, experience culture change on the basis of their own reactions to their experiences, leading to the development of new vocabulary.

While it is subject to manifold impulses toward change, language still must maintain a considerable amount of uniformity. If it is to be mutually intelligible among the members of the community, there must be a large element of agreement in its details among the individuals who make

up the community. As between the oldest and the youngest generations, there are often differences of vocabulary and usage but these are never so great as to make it impossible for the two groups to understand each other. This is the circumstance which sets a maximum limit on the speed of change in language.

Acquisition of additional vocabulary may proceed at a faster rate than the replacement of old words. Replacement in culture vocabulary usually goes with the introduction of new cultural traits replacing the old ones, a process which at times may be completed in a few generations. Replacement of fundamental vocabulary must be slower because the concepts (e.g., body parts) do not change fundamentally. Change can come about by the introduction of partial synonyms which only rarely, and even then for the most part gradually, expand their area and frequency of usage to the point of replacing the earlier word.

There is one apparent exception to the slow changing norm, namely, the complete displacement of one language by another. In the special case of individuals or single families transplanted into another speech community, the offspring may grow up speaking only the new language. When a small speech community is surrounded by a larger one, under certain circumstances the smaller community becomes bilingual and eventually may drop its original language. The process of switching to a new language, of course, is not the same as that of replacing vocabulary elements, and does not constitute a real exception to the principle of a maximum rate of vocabulary change. In the past some linguists have operated with a notion of language mixture, where a new language is supposedly formed by the simple fusion of two original languages. Studies of lexical statistics, as well as other evidence, show that this phenomenon does not occur. A language may be modified under the influence of a second language, or it may be replaced by it, with the new language sometimes modified by the old one before the latter disappears, but there is never a mechanical blend of equal and equivalent parts of two tongues.

Although the evidence for the constant comes from a variety of historic and reconstructed prehistoric cases, all of it falls within the recent millennia of human history and has to do with people having a certain minimum of societal development. We must therefore be cautious in drawing inferences as to the rate of vocabulary change long ago, say 100,000 or 500,000 years back, when human society may have been substantially less complex than the most primitive historically known groups. However, this is largely a theoretical point, since linguistic reconstruction does not reach back anywhere near those remote times.

DIVERGENCE

The dating of prehistoric events by vocabulary statistics is possible because of linguistic divergence. The simplest case is that in which a migration or an invasion, taking place fairly suddenly, breaks the community into two parts and where conditions prevent them from maintaining any effective contact with each other after the separation. The vocabulary changes that take place in the two groups will then be independent of each other, the words of fundamental vocabulary which are displaced in one community may or may not be lost in the other. After say 1,000 years, there will be two distinct languages, each of which will have retained a certain portion of the fundamental vocabulary of the earlier common form. This percentage, measured by means of our test list, would be approximately 81 per cent. Since the changes in the two languages are independent of each other, by the law of chance, they may be expected to coincide with each other in 81 per cent of 81 per cent of the cases, in short 66 per cent. In 2,000 years, when each language has only 66 per cent of the earlier common vocabulary, the two will agree in only 43 per cent. The relationship of time to percentage of correspondence of contemporary languages has to be figured as twice the time since their common earlier state, because divergence is going on simultaneously in both lines. The mathematical formula for time depth of divergent development is therefore:

$$d = \log c \div 2 \log r.$$

That is, time of divergence is equal to the logarithm of common percentage of vocabulary divided by twice the logarithm of the retention rate. To save the effort of consulting the logarithm table and carrying out the division, one may use a prepared chart, like the following based on $r = 81$ per cent ($r^2 = 66$ per cent) and divergence time given in centuries to the nearest half century.

95	1	70	8	45	19	20	38	10	54
90	2.5	65	10	40	21.5	18	40	9	57
85	4	60	12	35	25	16	43	8	59.5
80	5	55	14.5	30	28.5	14	46.5	7	63
75	6.5	50	17	25	33.5	12	50	6	66.5
								5	70.5

The first column is percentage of cognate vocabulary and the second column divergence time in centuries.

Dates calculated by the author in various earlier studies have been on the basis of $r = 85$ per cent, which gives slightly longer time spans. The comparison is shown here for intervals of tens of centuries to show the difference in the calculations:

Centuries:	10	20	30	40	50	60	70	80	90
r (%) = 81	66	43	29	18	12	8	5	3	2
r (%) = 85	72	52	37	27	19	14	10	7	5

Sometimes the two languages compared belong to different epochs. Say, language A is of the present while language B is of 2,000 years ago. The calculation is not difficult. One adds half of the time depth of B to the divergence time indicated by the common percentage of A and B. The formula, with dB representing the time depth of B, is

$$d = (\log c \div 2 \log r) + (dB \div 2).$$

SLOW DIVERGENCE

Often the divergence of one language into two takes place not from a clean separation of the old community into two parts but as a result of incomplete contact over a long period of time. Say, a language is spoken over a fairly large area, perhaps with four villages A, B, C, D. Say that there is still a fair amount of social intercourse between each pair of neighboring villages. After a lapse of time in this situation, one may find that villages A and B understand each other readily, and the same holds for B-C and C-D; but A understands C and B understands D with difficulty, and A and D perhaps cannot understand each other at all. Such a condition, described as a dialect chain (or net, if it is bi-dimensional), would be reflected in the lexical statistics of the dialects. In such a situation, the time depth calculated as between A and D presumably corresponds to the actual time from the first settling of the area or from the time when an earlier close social relationship among the villages broke down.

The type situation could be illustrated by many actual cases. The circumstance giving rise to the situation is that of incomplete inner cohesion of extended speech communities. If this relationship continues long enough, the divergence between even the close dialects may some day reach the point of difficult intelligibility and eventually of mutual non-intelligibility. But the earlier relationship will be reflected in the vocabulary correspondences. Such language chains seem to be evi-

denced at various points in the Salish linguistic map, for example, in the following languages of the Interior Division, shown here with the percentages of common vocabulary:

	Li	Sh	Ok	Cm
Lillooet	—	48	33	25
Shuswap	48	—	50	34
Okanagon	33	50	—	54
Columbia	25	34	54	—

Any neighboring pair of languages in this sequence share about half of their vocabulary; A with C and C with D have a third of their vocabulary in common; A and D have only a fourth. We infer from these figures that this series of languages has developed out of a single undifferentiated language in about 34 centuries—the time indicated by the minimum agreement of 25 per cent. The dialects which gave rise to the present languages must have constituted a chain in essentially the same order of geographic location from the beginning.

Contrariwise, geographic neighbors which do not show such lexical relationships could not have been in close contact continuously. Let us take as a contrasting example percentages in the following four geographically consecutive Salish languages:

	Sh	Li	Fr	Nt
Shuswap	—	48	19	19
Lillooet	48	—	28	26
Lower Fraser	19	28	—	58
Nootsack	19	26	58	—

In this sequence, A is rather closely related to B, and C to D, but there is a distinct break between B and C. Moreover, B is not closer to C than to D. This implies that the divergence of C and D must have taken place when the forerunner of this group was no longer in the relationship of a dialectal chain with B. The fact that C and D are closer to B than to A suggests that there was an old dialect chain long ago, while C and D were undifferentiated. In other words something like 39 centuries ago (based on 19 per cent minimum agreement in the chain) there was a dialect chain ABX. Not more than 32 centuries ago (based on 26–28 per cent agreement between B and C or D) the chain was broken into two parts, $A, B : X$ either through some sudden event, as migration or invasion, or because the slow linguistic change had passed the point of easy inter-dialectal influence. The unified language X split up into a dialect chain CD only after it ceased to be influenced by B. Further elucidation of prehistoric

relations among these languages can be obtained by considering them with reference to their neighbors in still other geographic directions.

The foregoing examples are illustrative of the kind of inference that can be drawn from lexical statistics about the prehistoric relationships among human communities. The amount of detailed prehistory that can be read in this way is of course limited, but it represents an important advance over what was previously possible.

DIFFUSIONAL EVIDENCE

The evidence of vocabulary statistics can often be coordinated with other linguistic evidence, particularly that of diffusion. An interesting case in the Salish stock is the close agreement of a series of distantly related languages in certain points of phonetics. Tillamook, Cœur d'Alene, Comox, and the Puget Sound group of languages show evidence of having once made the sound shift of Salish w to g^w; they also agree in the change of k to ch and partly also on the treatment of Salish y. The common features of this set of languages in part stand out against the typical traits that unite them with their closer linguistic relatives; thus Comox and Cœur d'Alene are more like each other in their treatment of original y and w, even though they are a thousand miles apart, than like their immediate geographic neighbors and their immediate sister languages linguistically speaking. These facts could be easily reconciled if the languages mentioned were in geographic contact with each other, since it is known that phonologic changes can move by diffusion across language boundaries. Since they are not now in contact, the inference is that they once were and that they continued so long enough to admit the diffusional imitation of a series of sound changes. The best assumption geographically is that the area in which these groups lived as neighbors must have been somewhere in the area of Puget Sound or Vancouver Island. It follows then that Tillamook and Cœur d'Alene have migrated to their present areas in relatively recent times.

Is there any way of dating these events? Within limits, *yes*. The diffusional situation must have existed after the time when Comox branched off from its nearest linguistic relatives, Seshelt and Pentlatch, for neither of these languages shares special features with Tillamook and Cœur d'Alene. Comox-Seshelt-Pentlatch show evidence of having once been a dialect chain, with 45 per cent

vocabulary agreement between the extremes, and this figure corresponds to 19 centuries of divergence. The contact of Comox with Puget, Tillamook, and Cœur d'Alene was therefore within the last 19 centuries, and the Tillamook and Cœur d'Alene migrations must have taken place since that time. For the present, we cannot be more exact than this, but it is probable that subsequent studies may narrow the period within which the events could have taken place.

Borrowed words as diffusional items are of particular interest because they give evidence not only of ancient contacts but also of some of the cultural items that passed from one group to the other. An illustration can be given here which fits in with the phonological diffusion already cited. In Quileute, a non-Salish language of the Olympic Peninsula, and in Nootka, an equally unrelated tongue of Vancouver Island, we find the word qaawats "potato." Related forms are found in Tillamook, qakts "potato," and in Cœur d'Alene, sqigwts "wild potato." Now, in Tillamook every present-day k represents an older labialized k^w which in turn was the product either of original k^w or of g^w from w. Thus Tillamook "potato" must have once had the form qawts, which corresponds exactly with Nootka and Quileute except for differences in the vowels reflecting nothing more than the diverse quantitative scheme of vowels in Salish as contrasted with Nootka and Quileute. A murmur vowel between w and ts and the long pronunciation of the first vowel make Salish qawts sound practically the same as Nootka qaawats. The Cœur d'Alene word has the noun-prefix s- and has a changed vowel in the first syllable of the word, but is otherwise derived from the same earlier form as the Tillamook; the vowel can be explained as a secondary change somehow related to Salish sound symbolism.[18] The Salish words thus match perfectly with the Quileute and Nootka, and borrowing must have taken place. There is at present little to indicate the direction of diffusion, whether from a Salish language to Nootka and Quileute, whether from one of these to Salish, or from some entirely different language to all of these. What is clear is that the word goes back to the same region where the w to g^w sound shift took place, and that it was in use about 2,000 years ago. It is also evident that it applied to some root native to that region and important in the diet of the local people.

---

[18] See Gladys A. Reichard, Sound symbolism in Cœur d'Alene, *Internatl. Jour. Amer. Ling.* 11: 78–91, 1945.

DATED LINGUISTIC PALEONTOLOGY

The example of prehistoric word borrowing just cited illustrates how reconstruction of ancient words and their meanings throws light on prehistoric culture even to small details. This method, known as linguistic paleontology, may be applied either with loanwords or with the native words of the linguistic group and constitutes an important means for penetrating prehistory.[19] Lexical statistics greatly improves the possibilities of linguistic paleontology by adding approximate dates to the detailed cultural information. A systematic exploitation of this advantage should bring valuable results, especially in correlating linguistic reconstructions with archeological evidence.

[19] A classical example of linguistic paleontology is O. Schrader and A. Nehring, Reallexikon der indogermanischen Altertumskunde, Berlin-Leipzig 1917–1928.