

ANAIS
DO
XXXI CONGRESSO INTERNACIONAL
DE
AMERICANISTAS
SÃO PAULO

23 a 28 de agosto de 1954

organizados e publicados por
HERBERT BALDUS

VOLUME II

EDITORA ANHEMBI

São Paulo

1955

TOWARDS A SATISFACTORY GENETIC CLASSIFICATION OF AMERINDIAN LANGUAGES

by

MORRIS SWADESH

1. In order properly to meet the needs of American prehistory, a language classification should :

- (a) treat as a single language all local speech varieties which are identical with each other or which are nearly identical within the criterion of mutual intelligibility, even if the peoples are known by separate tribal or local names,
- (b) keep apart as separate languages speech varieties which are not mutually intelligible, even if the peoples are conventionally grouped under a single name,
- (c) clearly distinguish relations of common origin from those of element diffusion,
- (d) distinguish recent from ancient common origin and indicate the approximate time depth of each relationship,
- (e) exclude from any grouping every member language which cannot be proved to belong in it,
- (f) include in each grouping every member which properly belongs,
- (g) be uniform in scale and accuracy for all parts of America,
- (h) ascertain all distant relationships within the limits of modern comparative linguistic science (nor-

mally revealing connections which go back at least about 6000 years and perhaps as high as 10,000).

No classification thus far made meets these requirements. Among those which come close to doing so, the limited to America north of Mexico, the classic study made under the direction of J. W. Powell violates all but two (c, e) of the criteria to a greater or lesser degree. More recent treatments in the main show improvement, sometimes in one respect, sometimes in several. Not infrequently, they show improvement in one respect, accompanied by serious weakening in another. Must we infer that the desired classification is completely out of reach at the present time? Is it something that must still wait another century until sufficient data can be gathered and until a large corps of comparativists has carried on many decades of detailed study?

The author believes that this need not be so, that the goal is today within reach and can be attained if we make effective use of present possibilities. He is already engaged in a project which envisages a uniformly scaled total classification of American languages within a few years. The purpose of the present paper is to explain the project, to invite criticism and above all to ask the broad cooperation of Americanists necessary for the success of the study.

2. *Diagnostic Lists.* The key to our problem lies in the use of diagnostic rather than complete data. A test list, consisting of 200 items of meaning in the non-cultural sphere has been used with very satisfactory results. The resulting classifications have been found to coordinate well with those derived from far more extensive bodies of material, embracing structural features and detailed phonology as well as much larger amounts of vocabulary.

It is of course easier by far to obtain a 200-word vocabulary of a language than an exhaustive body of well analyzed material. By appealing to anthropologists, missionaries and native speakers, it has been possible in

the first six months of systematic efforts to collect about 150 vocabularies to date and perhaps twice that many more have been promised to be submitted within the next several months. This is a good beginning toward the hundreds of languages still spoken in America.

One of the encouraging developments in the last few years is the undertaking of linguistic surveys of different areas. We have been apprised of such projects for the State of California, U.S.A., undertaken by the Linguistics Department of the University of California, for Venezuela, by Martha Hildebrandt of the Comisión Indigenista, for Colombia, by Jean Caudmont of the Instituto Colombiano de Antropología. Perhaps still others are under way which have not come to our attention, perhaps even more will be planned as this excellent idea spreads thru the hemisphere.

Besides long term surveys, like those mentioned, another approach to the problem is thru rapid surveys. C. F. Voegelin of Indiana University is even now (summer 1954) engaged in a field study of the Utoaztecan languages of the Great Basin, which involves collecting written and sound-tape recordings of a half dozen languages. Last summer the author carried out a similar study of Penutian languages under the auspices of Columbia University.

The full results of these linguistic surveys may not be published for some years, but it is very easy to extract small bodies of diagnostic material, such as the 200-word lists, and to bring them out without delay.

Data on extinct languages and older stages of contemporary languages can be extracted from published and manuscript sources for inclusion in the vocabulary collection. Several such have already been or are in the process of preparation.

Scholars are cordially invited to submit vocabularies for publication. Blank vocabularies and explanatory material will be furnished on request, by writing Morris Swadesh, 2625 Milwaukee Street, Denver 5, Colorado.

3. *Problems of Analysis.* Naturally, a vocabulary that is expertly prepared by someone acquainted both with the language and with linguistic methodology, will be more valuable than one recorded in a less skilful manner. Nonetheless, linguistic amateurs of all degrees can be of great assistance in a project of this kind. In the first place, it must be emphasized that even fragmentary and approximate concrete information in hand is a thousand times better than no information at all. But, even beyond that, it is generally possible, thru critical study and the interchange of a few letters, to improve the quality of a vocabulary to the point of high dependability.

In this project, my practice has been to first obtain a vocabulary of whatever quality from the best qualified person known to me as being in a position to obtain it and willing to do so. The request for a vocabulary is accompanied by a minimum of instructions along with the blank list. The best version of the blank form, which is bilingual in Spanish and English, has a place for pronominal affixes as well as the vocabulary proper. The accompanying materials include an explanation of the purpose of the vocabularies, and a few general instructions. They tell that the simplest everyday equivalent of each item of meaning is desired, and that the writing may be either conventional or phonetic, preferably but not necessarily along phonemic lines.

When the vocabulary comes in, the editor studies it critically for internal evidence indicating the adequacy of the writing system and for structural facts which may affect the use of the material for comparative purposes. Use is made of any explanatory notes submitted by the author. Moreover, reference is made to any related languages which may be already in the files, and to any conveniently available published material. On the basis of this critical study, the editor may write to the author for clarification of given points or may suggest possible revisions in the orthography or in the form in which the elements are given. The results of such interchanges between the editor and the author have been very satis-

factory. Sometimes even linguistically sophisticated scholars have obtained new insights into difficult phonemic or structural problems from clues suggested by the findings of other scholars and brought to their attention thru the medium of these discussions.

To give an idea of how the vocabularies are studied, one of the first points checked is the use of characteristic affixes with body part stems. If the words for *back*, *belly*, *bone*, *ear* etc. all begin with the same sounds and if these are absent in other nouns, one may infer that a possessive or body-part prefix of some kind is involved. Similarly, constant elements are sought in the verbs, in the adjectives and so forth. Again there are certain sets of words which are frequently expressed by identical or related forms. Thus, *water*, *rain*, *river*, *lake* and *sea* may include some identical expressions or may have the same basic element with different qualifiers. *Drink*, *flow* and *wash* may turn out identical with *water* or derived from that stem. *Bad* may be *not-good*. *Fog* may be *cloud* or *earth-cloud*. Other sets of similar meanings are *man-husband*, *woman-wife*, *husband-wife* (that is, *spouse*), *ear-hear*, *eye-see*, *fire-burn*, *ashes-dust-earth*, *head-hair* or *hair-feather-wing-hand*, *foot-leg*, *near-short-narrow*, *far-long-wide*, *at-in*, *bone-stone-tooth-egg*, *tree-stick*, *day-sun*, *seed-fruit-flower* or *seed-bone-stone*, *eye-flower-star*, etc.; opposites which may be defined by negatives include *dull-sharp*, *near-far*, *thick-thin*, *wide-narrow*. Since any language is likely to have a few of these identities or derivations, there is almost always some opportunity to observe phonetic variability and some outlines of structure. Still another approach to structure is to note the length of the words; any entry that is considerably longer than others may well be a compound. As for phonetics one can use the standard procedures of phonemic analysis even on a 200-word list, observing which consonants and vowel sequences occur in initial, medial and final positions, and what sounds occur in contact with what others.

4. *Comparability*. The purpose of analyzing the entries in the vocabularies and of determining phonetic

and structural relations within the language, is to make possible more effective comparison among the languages. For instance, if *near* is expressed in one language as *not-far* and in another language by a simple element, they cannot be cognate and any phonetic resemblance must be accidental. Similarly if the expression of *ear* in a language is *someone's-ear*, with a prefix for the indefinite possessor, it would be wrong to compare the initial part of this word with the root element in another language.

To facilitate comparison among many languages it is further desirable to have the orthographies as nearly alike as possible. Users of the diagnostic list should not have to consult the literature on each language in order to know, at least in general terms, what the symbols mean.

Matters of structure should be clearly indicated in as handy a fashion as possible. Hyphens can be used effectively in initial or final position, to indicate that the element is used only with affixes; within the form to show that it is made up of two or more parts. The analysis of the word in the latter case can be briefly given by a parenthetical translation of the parts with hyphens to parallel those in the native form.

Another very helpful device is to space the words uniformly in columns on the printed page, so that the corresponding columns of the two lists can be placed alongside of each other and the presence of cognates (or likely cognates) noted at a glance. In the published form of our vocabularies, the 200 items are arranged in four columns of 50 each on one side of a page. The meanings are not given there, but are given in a separate Semantic Key, also 50 in each column. The key is trilingual, in English, Spanish and Brazilian.

In addition to the various devices showing analysis briefly and simply to the extent possible in the word list itself, it is helpful to have additional information on each language easily available. This is accomplished by printing a one-page sketch of the language on the back of the vocabulary. A roughly standardized outline is

followed in these sketches, so that it is possible for a user to make a structural comparison of languages, at least as to their principal features.

5. *Time Depth.* Undoubtedly one of the most important features of any genetic classification of languages is a correct indication of time depth. It is not enough to say that languages A and B and C are related, but absolutely necessary to indicate which are closer together and which farther apart, and to what extent. Thus, Yucatec, Lacandon, Chontal and Huastec are all unquestionably Mayan languages, but the first two are very closely related, the third a bit farther removed and the last considerably more so. The degrees of relationship can be roughly indicated as follows:

AB C D

In the past the determination of degrees of relationship was either a very subjective or a very complicated procedure, often a combination of both. Scholars undoubtedly vary considerably in their ability to judge degrees of relationship after making an analysis or reading a description of each language in the group. Even after they have made lists of shared features, there is a question of weighting the relative diagnostic importance of each. Hence, it is not surprising that scholars are frequently found in disagreement as to the order of relationship among a series of languages belonging to the same genetic group.

Linguists in the past have mainly restricted themselves to making estimates of relative chronology, even tho the correlation of their results with archeology often requires some notion of absolute time. It was relatively safe to suggest that "A is more closely related to B than to C," but few ventured to suggest that "languages A and B must have separated in the neighborhood of 1200 years ago." Nevertheless, whenever archeologists correlate linguistic classifications with actual prehistoric sites of known stratigraphy or perhaps even with a definite date established by tree rings or radioactive carbon, the

linguist might well find himself forced to be more specific. He might then say, "A is too divergent from B to have been associated with site X." In general, the cautious linguist preferred not to say anything more on absolute chronology than he was forced to, and when the question had to do with a matter of 500 or 1000 years one way or another he might well be expected to declare, "It is impossible to know."

This situation has considerably changed in the last few years by the discovery of the "lexicostatistic constant," the fact that there is something like a uniform rate of displacement in unspecialized noncultural vocabulary. It is now possible to derive estimates of time depth by counting cognates in a diagnostic vocabulary, such as our 200-item list, once the constant has been determined in historical control cases.

Classifications, which once required long study and were ever beset with subjectivism, can now be achieved in little time by an objective procedure. And where chronologies were once largely confined to judgments of relative time-depth within linguistic groupings, today we can add estimates of actual divergence time.

The lexicostatistic treatment of a series of related languages consists in making cognate counts of each pair of languages, calculating the percentage of cognates over total of word pairs (usually a little less than 200, because of missing items or because of uncertainties in judging cognates), and converting the percentage into centuries of indicated time-depth by a mathematical formula or by reference to a table of percentages and time-depths. If the comparatist is already acquainted with the main facts of phonologic correspondence between two related languages, he can derive an estimate of the time of their separation in 5 or 10 minutes. The internal classification of a family or stock requires cognate counts for each pair of languages; thus, among 25 languages, there are 300 pairings. But, even in such a complicated case, the labor of determining the indicated interrelationship of the language is considerably less than by the old procedures.

The time-depth estimated by these means is a simple measure of the point at which the two languages began to diverge only if the first separation was an abrupt one, as by migration or suffering invasion. If the originally developing dialects remained in contact with each other, divergence would be more gradual according to the degree of contact which continued. This fact allows one to infer the prehistoric geographic arrangement of the dialects.

A crucial question in lexicostatistics is its accuracy. This has not yet been finally determined. We only know that in various cases where the results could be checked by archeology, the correlation has been excellent. We also know that the results within a linguistic stock seem to be internally consistent, and that the results are in accord with the relative chronology derived from other forms of linguistic evidence, insofar as these are unambiguous.

One of the known weaknesses of the method is that the persistence rates of elements in the test list are not entirely uniform. For example, body part names, as a class, are more persistent than certain of the verbs and adjectives. When sufficient studies have been made analyzing these matters from every point of view, it may be possible to revise the list into a much more dependable one. Until this is done, it is necessary to observe certain cautions. If it is impossible to obtain all or nearly all the items in the list, one needs to note what type of elements predominate. If the more persistent type (objects of nature, body parts, and such verbs as drink, eat, etc.), then the time depth will be minimized. It is desirable therefore to always try to fill the lists completely. It also follows that for remote relationships, as indicated by a low number of cognates with the less persistent elements largely gone, the indicated time depth will be lower than it should be. The constant of cognate retention in two languages for the full list is 66% after 1000 years. The constant for the more persistent elements has been calculated at 74%. We may assume that the 66% constant is approximately correct for the first 1000 years

or so, after which it begins to veer toward 74%, reaching this rate in perhaps about 5000 years. The tentative practice is to use the 66% constant thruout, but eventually a more accurate procedure will be necessary. To suggest the size of the error in the present practice, we may note that by figuring the constant as 66% for 3000 years and 74% for another 3000, we end up with 12% cognates after 6000 instead of after 5000 years.

6. *Grouping Scale.* In Amerindian Noncultural Vocabularies, an effort is made to give the affiliation of each language. In doing this, the terms family, stock and phylum are used according to a set scale, so that the reader may know from the term itself something of the degree of relationship. For example, Yucatec can be listed as of the Yucandonan subfamily, Choloyucan family, Mayan stock, Totomayan microphylum. The scale of relationships is as follows:

<i>Term</i>	<i>Time-depth Centuries</i>	<i>Cognate Percent</i>
Family	up to 25	over 36
Stock	25 to 50	36 to 12
Microphylum	50 to 75	12 to 4
Mesophylum	75 to 100	4 to 1
Macrophylum	over 100	under 1

Where two speech varieties show an indicated time-depth of under 5 centuries, that is a cognate percentage of over 81, they are considered to be dialects of the same language.

As the cognate percent falls, problems of identifying cognates may also arise. This is due partly to the accumulation of structural and phonetic changes with the passage of time. Furthermore, the low number of cognates in the test vocabulary is reflected even more in the rest of the vocabulary, which is generally much less persistent than our selected items. This means that it is hard to find enough examples of each phonologic relationship to

be entirely sure of the correctness of each assumed equivalence. In consequence, by the time one reaches the mesophylum level and sometimes even before, it is necessary to resort to reconstruction for calculating degrees of relationship. This means, if the conditions permit it, reconstructing test vocabularies for the prototypes of families and making the comparison between the reconstructed lists.

7. *Remote Relationships.* Generally speaking, when related languages are separated by less than about 4000 years, linguists have little trouble in recognizing the connection and there is not likely to be any difference of opinion as to their relatedness among scholars. For greater time depths, the quarrels begin between those who do and those who do not recognize relationship. There is frequently the claim and sometimes the real danger that relationships may be asserted on the grounds of vague similarities which could be due to chance or on the basis of solid similarities which may be due to borrowing rather than common origin. There are means of resolving these quandaries, based upon the use of diagnostic vocabularies and the counting of apparent cognates.

For this purpose, the author has used a reduction of the 200-word list, consisting of 97 of the more stable items. One counts only cognates of a given minimum phonetic complexity, for example, CVC elements, that is those consisting of at least a consonant followed by a vowel and another consonant. Calculating the probability of accidental agreements among CVC elements, it turns out that for most pairs of languages four CVC agreements represents a safe measure of caused relationship. Conservatively calculated, there is less than 1 chance in 70 that this minimum number of agreements might result from pure chance among 97 items.

Using this measure of proof, the following interesting new relationships are found to exist:

1) Sapir's Hokan-Coahuiltecan-Supanec phylum extends eastward to include Tunican and Natchez Muskogian,

and also embraces Yukian of California and Jicaque of Honduras.

2) The Kiotanoan and the Utoaztecan stocks are related to the Chibchan stock (and presumably to other groups included in Macro-Chibchan), and to the Panoan family.

3) The Totonacan family, the Mizoquean stock and the Mayan form a microphylum.

4) The Quechumaran stock is related to the Tupian and to Puelche.

5) Macro-Guaycuruan, Mapuche, Tehuelchean (Tehuelche-Ona), Alacalufan, and Yamana are all related to each other.

6) Alacalufan and Yamana compose a closer grouping, evidently a stock, within the Guacuyaman phylum.

* * *

BIBLIOGRAPHY OF AMERINDIAN GLOTTOCHRONOLOGY

MORRIS SWADESH, Lexicostatistic Dating of Prehistoric Ethnic Contacts with Special Reference to American Indians and Eskimos, *Proceedings of the American Philosophical Society* 96.453-63 (1952).

MORRIS SWADESH, Salish Internal Relationships, *International Journal of American Linguistics* 16.157-167 (1950).

JOSEPH H. GREENBERG and MORRIS SWADESH, Jicaque as a Hokan Language, *IJAL* 19.216-222 (1953).

MORRIS SWADESH, The Language of the Archeologic Huastecs, *Notes on Middle American Archaeology and Ethnology*, number 114, 1953.

MORRIS SWADESH, GEORGE I. QUIMBY, HENRY B. COLLINS, EMIL W. HAURY, GORDON F. EKHOLM, and FRED EGGAN: Symposium on Time Depths of American Linguistic Groupings, *American Anthropologist* 56.361-377 (1954).

MORRIS SWADESH, Perspectives and Problems of Amerindian Comparative Linguistics, *WORD*, in press 1954.