ON THE RATE OF REPLACEMENT OF WORD-MEANING RELATIONSHIPS

DAVID SANKOFF

Université de Montréal

A method is proposed for quantitatively comparing word-meaning lists in related languages, taking into account synonymous lexical representations. This is applied to the Indo-European languages to estimate rates of change of lexical representation for more than a thousand meanings documented by Buck. Comparisons between the distribution of rates for the lexicostatistic test list and other meanings, and between the distributions calculated from Indo-European data and those from Malayo-Polynesian, suggest generalizations of lexicostatistic theory.

The turnover rate for words taking on a given meaning was first studied by Swadesh 1955 in the context of his controversial theory of lexicostatistics. He had proposed a basic 'test list' of meanings, and he expected the words for all of these to be more persistent than the rest of the vocabulary. Furthermore, he thought that, within the test list, the meanings would all have approximately the same persistence, and that this would be independent of language and time period. Swadesh noted, however, that in Lees' 1953 data, the persistences associated with the different meanings seemed to be grouped according to a rather dispersed frequency distribution. Dyen 1964 demonstrated empirically that those basic meanings with the most persistent lexical representations tend to have this property independently of the language family concerned, as do those with the least persistent representations. Dyen, James, and Cole 1967 published a list of replacement rates for the Swadesh list,[1] calculated from comparisons in 46 pairs of Malayo-Polynesian languages.

The present paper has two objectives: (1) to generalize lexicostatistic theory so that it considers all synonyms for a meaning instead of only the most common or most easily elicited word; and (2) to apply this theory in a study of the replacement rates associated with the thousand-odd meanings listed by Buck 1949. This will involve a comparison of the behavior of the test list with that of the general vocabulary.[2]

1. THEORETICAL CONSIDERATIONS. In studies of word replacement, it is usually assumed that in every language one can elicit a unique lexical representation (the most common word, the most frequently elicited, etc.) of each meaning m in the test list. This word runs a constant risk λ_m of being replaced by a non-cognate word; i.e., the probability of being replaced within a short time, between $\text{time}_{\text{START}}$ and time_{END}, is approximately $\lambda_m \times [\text{time}_{\text{END}} - \text{time}_{\text{START}}]$. Depending on how one formalizes these assumptions (cf. Brainerd 1970, Sankoff 1969), one can obtain the probability that a word-meaning relationship will remain undisturbed by such a replacement over an arbitrary time interval, as a

[1] From this point on, I use the terminology 'replacement rate' instead of referring to the inverse entity, persistence (or retention rate).

[2] I would like to thank Dell Hymes for his comments on an earlier version of this paper.

function of λ_m and the length of the interval. As a consequence of this, one can estimate λ_m after seeing whether the word for a meaning has survived over various time periods in a number of different languages. The accuracy depends on the length of the time periods and the number of languages. In language families with little or no historical record, one can instead compare for cognation the words for the meaning m in the related languages and obtain an estimate of λ_m not in absolute terms, but relative to the rate for the other meanings.

In reality, although for some meanings it may be easy to find unique lexical representations, for others it may be difficult to choose between two, three, or more words. Moreover, any such choice inevitably wastes a certain amount of information about changes in lexical representation. If we want to take into account all the words for a meaning, on the other hand, we can no longer simply compare the words at two points in time (or in two related languages) and make a single cognation judgment. How, for example, should we compare the set of words a, b, and c at $\text{time}_{\text{START}}$ with the words a, b, x, y, and z at time_{END}?[3] Whatever the method adopted, it should score 'zero' if there are no cognate pairs between the sets of words from different times, and should score 'one' if the sets contain a single cognate pair. Consider first a measure derived by counting the words in the larger of the two sets and calculating the proportion of these having cognates in the other set. This measure scores zero or one when required, and somewhere in between when there are some cognate pairs but some unrelated words as well. Further, when the proportion of cognate pairs is high, and only then, the measure is high. For the example above, it scores 2/5. The only unfortunate property of this indicator is that it is just as sensitive to the presence or absence of very rare synonyms or usages as it is to more common ones.

To adjust for this, we assign a weight to each of the words which can take on the meaning m. The weight on word a may be interpreted either as the probability that a is used when m is expressed, or as the degree of appropriateness of a relative to the other words for m. At a given point in time, the sum of the weights of the words for any one meaning should equal exactly one. Then a natural comparison between two sets of words is given by $1 - \frac{1}{2} \times$ (the sum, over all words, of the absolute difference in weight on a word between $\text{time}_{\text{START}}$ and time_{END}). If all the words representing m at a given time are equi-probable, it is easy to prove that this second measure gives the same score as the first one, and is thus a consistent generalization of it.

Several models of the evolution of word-meaning relationship have been proposed (e.g. Sankoff 1969, 1970) to generalize the standard model of simple one-for-one lexical replacement. In these studies, our second measure is a natural generalization of the single cognation judgment of standard theory, though it does not necessarily follow that the two entities should behave alike as functions of the length of the time interval involved. For the models referred to, however, there is strong evidence from computer simulation experiments that they all behave like the simple lexicostatistic model. In particular, the exponential decline, in the simple model, of cognation probabilities (cf. Dyen et al. 1967) is paralleled in more sophisticated models by the exponential decline of the expected

[3] Two cognate words will be represented by the same letter.

value of our measure. Moreover, the exponential law parameter λ_m has the general significance, in these models, of the rate of change of lexical representation.

2. THE DATA. Buck's dictionary lists, for more than a thousand meanings, the (one or more) lexical items having the same meaning for each of 31 Indo-European languages. Also included is a summary of etymological knowledge about all these words. Although some of this information may be faulty or incomplete, the book is still a remarkably systematic and uniform presentation of a huge volume of data. It is probably the single most appropriate source of data for the type of quantitative historical survey described here. The basic data for this study consist of all Buck's cognation judgments (explicit and implicit) for each meaning, along with all information useful in estimating relative word usage probabilities, such as whether a form is usual or rare, archaic, literary, popular, etc. These data have been extracted according to a formal protocol by a number of assistants, and transferred to punch cards for the calculations.

It is also useful to have a matrix of divergence times between each pair of the 31 languages. This has been constructed as far as possible using historical sources, including archaeological estimates and the opinions of comparative linguists.

More details on these data are presented in Sankoff 1969.

3. ESTIMATING λ_m. The obvious estimator of the parameter in an exponential decline law is the natural logarithm of the observation $\div$ ($\text{time}_{\text{START}} - \text{time}_{\text{END}}$). This estimator, however, has a number of unpleasant properties which render it inapplicable. In particular, it takes on the value ∞ when there are no cognates; and it is difficult to combine, or pool, estimates from a number of interrelated observations.

Dyen et al. made elegant use of the analysis of variance to estimate the λ_m in a case where the values of $\text{time}_{\text{START}}$ and time_{END} are unknown. In the present study, information about divergence times between pairs of languages is incorporated as follows.[4] Among 31 languages there are 465 pairwise comparisons, but these comparisons are not all given the same weight in estimating λ_m. There are five modern Germanic languages and four Romance languages among the 31, and hence 20 comparisons between the two groups; but this collection of comparisons should weigh no more heavily in the estimate than, for example, the single comparison between Latin and Greek. To obtain such a balance, for each meaning m, values of the measure from all comparisons having the SAME divergence time have been averaged, and these averages summed over all DISTINCT divergence times (only 65 of these), resulting in a total S_m. Under an exponential decline law, there is a unique value of the parameter λ for which such a sum (i.e. over 65 observations) has value S_m, and this value has been calculated (numerically) to estimate λ_m.[5]

[4] For a language attested at two points of time in its history, the divergence time is $\text{time}_{\text{END}} - \text{time}_{\text{START}}$. For two related languages sharing a common history up to $\text{time}_{\text{SPLIT}}$ and attested at $\text{time}_{\text{END 1}}$ and $\text{time}_{\text{END 2}}$, respectively, divergence time is ($\text{time}_{\text{END 1}} - \text{time}_{\text{SPLIT}}$) + ($\text{time}_{\text{END 2}} - \text{time}_{\text{SPLIT}}$).

[5] The stability of this estimator is illustrated by an experiment in which all comparisons involving the 33 largest divergence times (out of 65) were ignored (and the calculation adjusted accordingly). This had very little effect on the individual estimates, and almost no systematic effect.

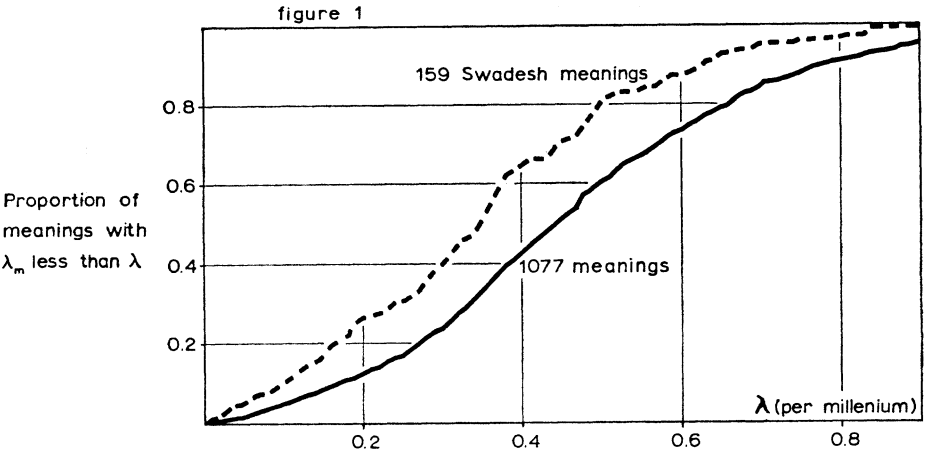

FIGURE 1.

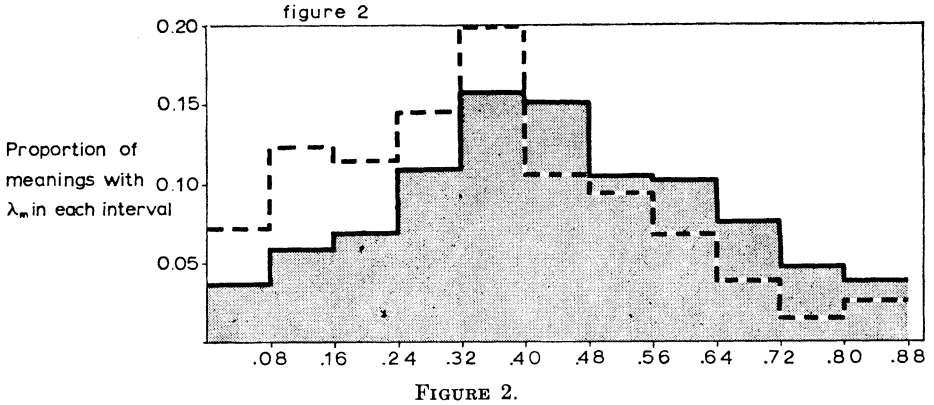

FIGURE 2.

4. RESULTS. Figures 1–3 summarize the results of these calculations.

Figure 1 represents the CUMULATIVE frequency distribution of λ_m for 1077 meanings from Buck's dictionary, compared with the corresponding curve for 159 meanings which are both in the dictionary and on the Swadesh 200-meaning list. One may read from this graph, for example, that 40% of all the meanings m have λ_m less than 0.4, but 65% of the Swadesh meanings have such low replacement rates.

Figure 2 shows the same data in terms of ordinary frequency distributions. Note that in every interval of values of λ less than 0.4, the Swadesh meanings have proportionately greater representation than the general vocabulary, and the opposite is true for each interval of values of λ greater than 0.4.

Figure 3 compares the cumulative distribution for the Swadesh meanings with a curve constructed from the data in Dyen et al. on an almost identical set of 156 meanings in the Malayo-Polynesian languages. These data have been scaled so that the mean agrees with that for the Indo-European languages.[6]

[6] Dyen et al. present their results in relative terms. They make no strong claim for absolute time values, although they make one calculation in terms of a fixed time scale (pp. 169–70).

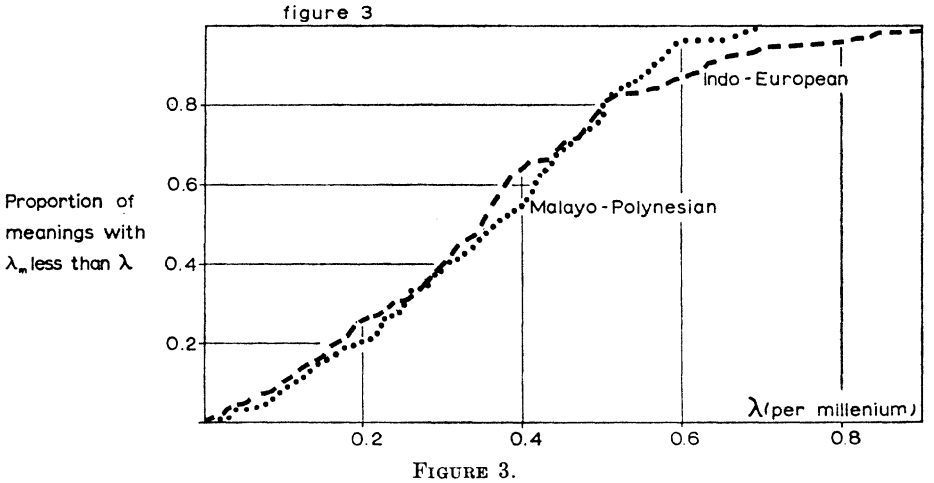

FIGURE 3.

5. DISCUSSION. As we have redefined λ_m, it no longer stands for the probabilistic rate of simple replacement of the words for a meaning, but for a generalized rate of change in word-meaning relationships. The standard model of simple replacement with the old interpretation for λ_m should be considered a special case of some more general model in which both sudden replacements and gradual usage changes occur and where a wider connotation must be given to λ_m.

Is the assumption that λ_m is a constant, depending only on m, warranted? Dyen showed that, at least within the Swadesh list, replacement rates depend strongly on the meaning. This paper presents evidence for extending this dependence, in verifying the hypothesis, in Indo-European, that the test list meanings are more persistent than the general vocabulary. This is still a long way from a proof of the assumption. Indeed, still more realistic models might allow λ_m to be a slowly varying random function of time. As it is 'λ_m is a constant' represents a second-order approximation to a descriptively adequate theory, Swadesh's original hypothesis being a first approximation.

Another way of looking at this problem may result from comparing the Indo-European results with the Malayo-Polynesian ones of Dyen et al. The similarity in the shape of the frequency distributions suggests that probability distributions of replacement rates, rather than constant replacement rates, should be the universals of lexicostatistic theory.

REFERENCES

BRAINERD, BARRON. 1970. A stochastic process related to language change. Journal of Applied Probability 7.69–78.

BUCK, CARL D. 1949. A dictionary of selected synonyms in the principal Indo-European languages. Chicago: University of Chicago Press.

DYEN, ISIDORE. 1964. On the validity of comparative lexicostatistics. Proceedings of the Ninth International Congress of Linguists, ed. by H. G. Lunt, 238–52. The Hague: Mouton.

——; A. T. JAMES; and J. W. L. COLE. 1967. Language divergence and estimated word retention rate. Lg. 43.150–71.

LEES, ROBERT B. 1953. The basis of glottochronology. Lg. 29.113–27.

SANKOFF, DAVID. 1969. Historical linguistics as stochastic process. Doctoral thesis, McGill University.

——. 1970. Lexical replacement processes. Unpublished.

SWADESH, MORRIS. 1955. Towards greater accuracy in lexicostatistic dating. IJAL 21. 121–37.

[Received 27 January 1970]