PART ONE - LEXICOSTATISTICS

The Swadesh theory of lexicostatistics (1950, 1952, 1955) provided the first quantitative comparison of related languages based on a well-defined model of language change. The stochastic nature of this model was poorly understood by linguists, in the main, and many have rejected the theory in the course of a protracted and confused controversy. Meanwhile field linguists, especially those working with language groups of unknown history, have accepted lexicostatistics and have found it to be an efficient, valid and reliable technique.

The Swadesh theory

There are serious oversimplifications of reality implicit in lexicostatistics, and it is these, rather than the stochastic aspects, which are limitations of the theory. Swadesh hypothesized, in effect, that

(i) it is possible to discover a set of basic, universal and non-cultural meanings, and he constructed a list of about 200 such meanings;

(ii) in every natural language, at a given time, there is a unique lexical representation (word) corresponding to each of these meanings; but

(iii) over short time intervals, the word representing any meaning runs a small but constant risk of being replaced by a different (non-cognate) word; and

(iv) the replacement, or non-replacement, of the lexical

representation of a meaning occurs independently of that of any other meaning, and independently over different periods of time.

To formalize (i) and (ii), we must postulate the existence, for each natural language, at all points t in time, of a lexicon, represented by a finite abstract set L_t. A well-defined equivalence relation corresponding to cognation partitions the elements of $\bigcup_{t\in T} L_t$ (T a real interval) into equivalence classes. If $k\in L_s$, $l\in L_t$ $(t,s\in T)$ are cognate, we write $\delta(k,l) = 1$. Otherwise $\delta(k,l) = 0$.

Further, we must postulate the existence of a finite abstract set M (corresponding to the universal set of meanings), and a procedure for defining, for any t, and any L_t, a unique map from M into L_t. This map, written $M \xrightarrow{t} L_t$, specifies that for each $m\in M$ there is a $l\in L_t$ such that $m \xrightarrow{t} l$ (l means m).

Hypotheses (iii) and (iv) imply that the changes over time in the image of the map $M \xrightarrow{t} L_t$ have a certain stochastic aspect. This can be modelled by the probability statement

$$P\left[\delta(k,l) = 1\right] = 1 - \lambda(t - s) + h \quad ,$$

λ a universal constant, $t > s$, and $h/t-s \to 0$ as $t \to s$; and two independence conditions; let

$$\left.\begin{array}{l} m_i \xrightarrow{s} k_i \\ m_i \xrightarrow{t} l_i \end{array}\right\} \quad \text{for } i = 1,2,\ldots,|M| \quad (|M| \text{ the number of elements in M}),$$

then $\delta(k_1,l_1)$, $\delta(k_2,l_2)$, . . . , $\delta(k_{|M|},l_{|M|})$ are independent random variables; let

$m \xrightarrow{s_i} h_i$ and $m \xrightarrow{t_i} j_i$ for $[s_i, t_i)$, $i = 1,2,\ldots,N$, a finite number of disjoint intervals,

then $\delta(h_1,j_1)$, $\delta(h_2,j_2)$, . . . , $\delta(h_N,j_N)$ are independent random variables.

This model has a number of immediate properties which form the central thesis of the Swadesh theory. These are presented here as Theorems 1, 2 and 3. For simplicity we will assume that at any time t, at most one word in L_t can belong to a cognation equivalence class. This simplifies notation and proofs, although the assumption may be relaxed without substantively affecting this development.

One further type of assumption is required to ensure a degree of randomness in the choice of replacing word during lexical replacement. To prove Theorem 1 as stated below, we require $\exists C > 1$ such that

$$P\left[m \xrightarrow{t} l \,\middle|\, m \xrightarrow{s} k\right] \leq \frac{C}{|L_t|} = O(|L_t|^{-1}),$$

for all m, $t > s$, $\delta(k,l) = 0$.

Theorem 1

$$P\left[\delta(k,l) = 1\right] = e^{-\lambda(t-s)} + O(|L_t|^{-1})$$

$$|M|^{-1} E \sum_{m \in M} \delta(k_m, l_m) = e^{-\lambda(t-s)} + O(|L_t|^{-1})$$

Proof

Let $N(t-s)$ be the number of changes (with respect to the cognation relation) of the mapping $m \xrightarrow{t} l$ in the interval $(s,t]$. Then $N(t-s) = 0$ is just the event that a Poisson process remains at zero on the interval $(s,t]$ (see, e.g. Parzen, 1960, p.252), and

$$P\left[\, N(t-s) = 0\right] = e^{-\lambda(t-s)} .$$

However,

$$\begin{aligned} P\left[\delta(k,l) = 1\right] &= P\left[N(t-s) = 0\right] \\ &\quad + P\left[N(t-s) > 0\right] \times P\left[\text{last change is to } k \text{ (or cognate)} \mid N(t-s) > 0\right] \\ &= e^{-\lambda(t-s)} + (1 - e^{-\lambda(t-s)})\, O(|L_t|) \\ &= e^{-\lambda(t-s)} + O(|L_t|) \end{aligned}$$

Then

$$E\ \delta(k,l) = e^{-\lambda(t-s)} + O(|L_t|)$$

$$\begin{aligned} |M|^{-1}\ E\sum_{m\in M}\delta(k_m,l_m) &= |M|^{-1}\ \sum_{m\in M} E\,\delta(k_m,l_m) \\ &= |M|^{-1}\,|M|\left[e^{-\lambda(t-s)} + O(|L_t|)\right] \\ &= e^{-\lambda(t-s)} + O(|L_t|) \end{aligned}$$

Definition

L'_t, L''_t, $t>s$ represent the lexicons of two languages which are independent daughter languages of the same parent language (which are said to split at time s) if

$L'_s = L''_s$, and if

$m \xrightarrow{s} k$ (in both languages)

$m \xrightarrow{t} l'$ in the first language

$m \xrightarrow{t} l''$ in the second language,

then $\delta(k,l')$ and $\delta(k,l'')$ are independent random variables.

Theorem 2

Let L'_t, L''_t, $t>s$ be as above.

Then

$$P[\delta(l',l'') = 1] = e^{-2\lambda(t-s)} + O(\min|L'_t|, |L''_t|)$$

$$|M|^{-1} E \sum_{m\in M} \delta(l'_m, l''_m) = e^{-2\lambda(t-s)} + O(\min|L'_t|, |L''_t|)$$

Proof

Assuming $m \xrightarrow{s} k$ in both languages,

$$P[\delta(l', l'') = 1] = P[\delta(l',l'') = 1 \cap \delta(k,l')=1] + P[\delta(l',l'')=1 \cap \delta(k,l') =$$

By transitivity of the equivalence relation represented by δ, the first term on the right is

$$P[\delta(k,l'') = 1 \cap \delta(k,l') = 1], \text{ which, by independence}$$
$$= P[\delta(k,l'') = 1]\ P[\delta(k,l') = 1]$$
$$= [e^{-\lambda(t-s)} + O(|L'_t|)]\,[e^{-\lambda(t-s)} + O(|L''_t|)]$$

$$= e^{-2\lambda(t-s)} + e^{-\lambda(t-s)} \left(O(|L_t'|) + O(|L_t''|) \right) + O(|L_t'||L_t''|)$$

$$= e^{-2\lambda(t-s)} + O(\min|L_t'|, |L_t''|)$$

Now the second probability on the right hand side above is, similarly,

$$P[\delta(l',l'') = 1 \cap \delta(k,l') = 0 \cap \delta(k,l'') = 0]$$

$$= \sum_{\substack{l' \in L_t'' \\ l'' \in L_t''}} P[m \xrightarrow{t} l' \text{(1st language)} \cap m \xrightarrow{t} l'' \text{(2nd language)}] \, \delta(l',l'')$$

$$\delta(l',k)=0$$
$$\delta(l'',k)=0$$

$$= \sum P[m \xrightarrow{t} l'] \; P[m \xrightarrow{t} l''] \; \delta(l',l''), \text{ by independence.}$$

Since we have fixed $m \xrightarrow{s} k$

$$P[m \xrightarrow{t} l'] \leq \frac{C}{|L_t'|}$$

The summation contains at most

$$\max(|L_t'|, |L_t''|)$$

terms which are not annihilated by $\delta(l',l'')$

and so the total is

$$\leq \frac{C}{|L_t'|} \; \frac{C}{|L_t''|} \max(|L_t'|, |L_t''|)$$

$$= O(\min|L_t'|, |L_t''|)$$

This completes the proof of the first statement of the theorem. The proof of the second parallels the analogous result in the previous theorem.

In natural languages, $|L_t|$ is several thousands and $\frac{1}{|L_t|}$ is negligible compared to the exponential term, except for very high values of t (where the theory has little applicability). In the next theorem, the results of Theorems 1 and 2 are utilized, neglecting the error terms of the form $O(|L_t|)$.

Under certain, more specific restrictions on $P\left[m \xrightarrow{t} l \mid m \xrightarrow{s} k\right]$, Brainerd (n.d.) solved for the exact form of the error term attached to the exponential laws (here formulated as Theorems 1 and 2).

Theorem 3

Insofar as we may approximate the results of Theorems 1 and 2 by

$$|M|^{-1} E \sum \delta(k,l) = e^{-\lambda(t-s)}$$

and

$$|M|^{-1} E \sum \delta(l',l'') = e^{-2\lambda(t-s)}$$

respectively; if it is known that $t-s = T$, then

$$\hat{\lambda} = \frac{-\log |M|^{-1} \sum \delta(k,l)}{T}$$

is the maximum likelihood estimator (MLE) of λ in the first formula above, and if λ is known,

$$\widehat{t\text{-}s} = \frac{-\log |M|^{-1} \sum \delta(k,l)}{\lambda}$$

is the MLE of t-s.

In the case of two independent daughter languages (Thm. 2),

$$\widehat{t\text{-}s}' = \frac{-\log |M|^{-1} \sum \delta(l',l'')}{2\lambda}$$

is the MLE of t-s.

Proof

It suffices to find the MLE of λ, the other cases being analogous.

Consider binomial trials with parameter $p = e^{-\lambda T}$.

$\delta(k,l) = 1$ is the equivalent of a success in one such trial.

$\sum_{m \in M} \delta(k_m,l_m) = r$ is the equivalent of r successes in $|M|$ trials.

The likelihood function of λ in such a case is

$$L(\lambda) = \binom{|M|}{r} e^{-\lambda T r} \left(1-e^{-\lambda T}\right)^{|M|-r}$$

$$\log L(\lambda) = \text{constant} - \lambda T r + (|M|-r) \log (1-e^{-\lambda T}).$$

$$\frac{d \log L(\lambda)}{d\lambda} = -Tr - Te^{-\lambda T} \frac{(|M|-r)}{1-e^{-\lambda T}}$$

At the MLE, $\hat{\lambda}$, this derivative should be zero,

$$Tr - Tre^{-\hat{\lambda} T} = |M| Te^{-\hat{\lambda} T} - Tre^{-\hat{\lambda} T}$$

$$\hat{\lambda} = \frac{-\log \frac{r}{|M|}}{T}$$

and the same process yields

$$\widehat{t-s} = \frac{-\log \frac{r}{|M|}}{\lambda}$$

Let $r = \sum_{m \in M} \delta$ as in Theorem 3. Swadesh (1950) derived a methodology to utilize the three results

$$\hat{\lambda} = \frac{-\log (r/|M|)}{t-s}$$

$$\widehat{t-s} = \frac{-\log (r/|M|)}{\lambda}$$

$$\widehat{t-s}' = \frac{-\log (r/|M|)}{2\lambda}$$

as follows. He first selected his list of meanings which he considered basic to all languages. He then compared Old English with Modern English ($t-s \approx 1000$ years), i.e. he compared the words in each language corresponding to the basic meanings. The etymology of words in these languages being fairly well known, he was able to decide when a pair of words corresponding to the same meaning were cognate (i.e. one was historically derived from the other, or both were derived from a common root, by a series of phonological alterations, each of which affected only a part of the word in question). This immediately led to $\lambda \approx 2 \times 10^{-4}$. Using the estimate which he obtained as a constant, he dated the relative times of separation or "split" of

various Salish (western North American Indian) languages from a common parent with the estimator $\hat{t}_s'$. After the work of Lees (1953), λ was considered to be a universal constant, $\hat{t}_s'$ could estimate absolute dates of split, and $\hat{t}_s$ could date a collection of texts from a dead language.

Criticisms of the theory

Criticisms of lexicostatistics fall into two classes. In the first class are protests based on or resulting from the stochastic nature of the model and/or the stochastic nature of the phenomena of lexical loss and replacement. The second class of criticisms refer to particular assumptions in the model , and I will discuss these in the next section.

Bergsland and Vogt (1962) presented four cases where $\hat{t}_s$ (or $\hat{t}_s'$) are not accurate (three too low and one too high), and rejected the Swadesh theory on this basis. In statistical terms, the authors constructed a sample consisting entirely of outliers and rejected an hypothesis without even considering the distribution of the test statistic. Fodor (1962) took the same approach to "disprove" lexicostatistics. Chretien (1962) calculated and published pages of ordinary binomial functions to prove, in essence, that $\hat{t}_s$ is a random variable and hence not "an acceptable mathematical formulation" of the Swadesh theory. This basic misunderstanding of the nature of statistical estimation is characteristic not only of critics of lexicostatistics, but also of many of its practitioners.

A more important criticism has been expounded, at great length, by Fodor (1965) and, more clearly, by Teeter (1963). Quoting from the latter:

> "Lexical similarities and dissimilarities do not come about in any one simple way, and any mechanical method of counting lexical similarities cannot separate those due to chance, universals, diffusion, and common origin. Lexical change is the result of many factors, and all are scrambled together in the final result."
>
> (p.641)

This diversity of causes of lexical and semantic change has received detailed study by linguists and semanticists; see, for example, Bloomfield (1933) p.392 ff., Ullman (1957) p.183 ff. Quoting from Lees (1953):

> " The reasons for morpheme decay, i.e. for changes in vocabulary, have been classified by many authors; they include such processes as word tabu, phonemic confusion of etymologically distinct items close in meaning, change in material culture with loss of obsolete terms, rise of witty terms or slang, adoption of prestige forms from a superstratum language, and various gradual semantic shifts, such as specialization, generalization, and pejoration."
>
> (p.114)

And it is just this diversity and the difficulty of "unscrambling" which, contrary to Teeter and to Fodor, justifies a stochastic model incorporating retention parameters. Consider, for comparison, the problem of constructing a model for the behaviour of gases. We have an enclosed volume containing a large number of particles of

finite dimension, undergoing rapid motion. We can assume everything is perfectly deterministic, all the particles obeying Newton's three laws of motion, and all collisions perfectly elastic. The position of any particle at any time can, theoretically, be calculated precisely if we know the initial state of the system and the time elapsed. Practically speaking, of course, this would be impossibly tedious, boring and pointless, there being so many particles, any two of which may collide, plus the walls, plus gravitational or electrical charge attractions and repulsions to consider. What is possible, interesting, and of great value (witness the fields of kinetic theory and statistical mechanics, dating from the work of men such as Maxwell, Boltzmann, and Einstein) is to consider the nature of each particle as a random process involving appropriate parameters and to consider the statistical behaviour of the model thus constructed. It is complexity and great difficulty of prediction which make a statistical model workable. In the same way, Fodor and others have inadvertently justified the proposition that some sort of stochastic process might be an appropriate model for lexical change phenomena. The question remains, what process? The Swadesh theory provides at least a first approximation to the correct answer.

Problems with Swadesh's model

Before discussing details of the model, it is appropriate to present the results of an early (1953) lexicostatistic investigation of R. Lees. He chose thirteen language pairs, each pair consisting of an historic language and a modern descendant. The

particular choice of pairs presumably stemmed from availability and not from any sampling technique. He translated each word in Swadesh's 215-word list (1950) into the 26 languages. After counting the number, r, of cognates between each language pair, he used (in effect),

$$\hat{\lambda} = \frac{-\log (r/|M|)}{t-s}$$

where $|M| \leq 215$ according to the number of indeterminate cognations and uncertainties of translation. To get an estimate of a "universal" λ, he combined the individual estimates in

$$\lambda = -\log\left(\frac{1}{13} \sum_{i=1}^{13} e^{-\hat{\lambda}_i}\right)$$

($\lambda = \frac{1}{13}\Sigma\lambda_i$ gives approximately the same result.)

Using $p = e^{-\lambda t}$ as the parameter in the binomial experiment he calculated, for each language pair,

$$\frac{(|M|p - r)^2}{|M|p(1-p)}$$

which should be approximately the square of a standard normal random variable, if the assumptions of the theory are true. Since an estimate of λ is used in calculating p, the sum of the squared variables should be χ^2_{12}-distributed. But $\chi^2_{12} = 29.5$, significant at the 1% level, suggesting rejection of the theory.

Lees, however, suggested four reasons for <u>not</u> rejecting on the basis of the χ^2 test; the large values for $|M|$ and r, uncertainty in t, possible inappropriateness of the χ^2 test, and the error in estimating λ. The first and third of these are not valid

statistically, and the fourth is a source of very little of the excess χ^2. The variability in the time parameter can be incorporated into the χ^2 calculation. This only reduces χ^2 to 25.9 - 27.5 depending on the variation assumed in t. Lees' results, then, indicate strongly that the theory is an inadequate model for the phenomena.

We turn now to the second class of criticisms of the Swadesh model, those that involve objections, evaluations or improvements related to the generalizations and simplification of reality inherent in lexicostatistic theory. The listing of assumptions earlier in this chapter will serve as a framework for classifying this latter class of criticisms.

(i) There are no universal sets of meanings, it being difficult to specify most meanings without recourse to particular natural languages. No list of meanings yet devised is completely satisfactory for sufficiently diverse languages; Hoijer (1956), O'Grady (1960), Cohen (1964), Levin (1964), Trager (1966).

(ii) The existence of synonymy proves the non-uniqueness of the meaning map $M \rightarrow L$; and no known methods of eliciting words for given meanings are completely and reliably reproducible, from speaker to speaker or even from occasion to occasion for a single speaker; Gudschinsky (1960). The existence of general and specific terms for a single entity provides a further complication.

(iii) If the parameter λ can be said to exist at all, it is constant neither from language to language; Bergsland and Vogt (1962),

Fodor (1962), from meaning to meaning; Swadesh (1955), Andreyev (1962), Ellegard (1962), and especially Dyen (1964), van der Merwe (1966), Dyen, James and Cole (1967), nor even from time interval to time interval for the same meaning; Swadesh (1962).

Judgements about cognation are unreliable, especially with respect to languages which are separated by large t-s and whose history is mostly unknown; Fairbanks (1955), Teeter (1963), Lunt (1964). An analysis of this latter problem is beyond the scope of this study.

(iv) Lexical loss and replacement do not occur independently for different meanings, neither are current and future trends entirely independent of what has happened in the past, especially in languages which have possessed an orthography for some time. This has been noted especially in connection with the independence assumption of Theorem 2, as in the interval immediately after a split we might expect parallel (to some extent, at least) evolution of the two daughter languages; Lees (1953), Hymes (1960), Teeter (1963). Also in this connection, independence of evolution does not strictly hold where borrowings, loan-translations and imitations of other types are frequent occurrences.

Towards a new theory

A number of authors have attempted to deal with one or more of these problems. Swadesh (1952) discarded more than half of the meanings in his original list. For choosing among synonyms, Gudschinsky (1956) proposed a random selection, Hymes (1960) suggested a procedure which would select cognate forms whenever they were

available, Satterthwaite (1960) and Dyen (1960) pointed out that it would be more reasonable to choose the word which is most frequently used for the meaning in question.

Little could be done about the central postulate or result of the theory; that λ is a constant, until the work of Dyen became well known. Dyen, on the basis of comparisons of a large number of Malayopolynesian languages was able to segregate meanings into groups on the basis of their individual λ's. A discussion of the mathematical implications of this ($p=e^{-\lambda_i(t-s)}$ for meaning m_i leads to $E(r/|M|) = \sum_{i=1}^{M} e^{-\lambda_i(t-s)}$) was published by van der Merwe (1966). Meanwhile, Dyen (1964) had statistically demonstrated that meanings with high λ in the Malayopolynesian languages tend to have high λ in the Indoeuropean languages and <u>vice versa</u>. This was the first new type of lexicostatistic result since the work of Lees. Later (1967) this work was refined so that Dyen <u>et al</u> were able to estimate a separate λ for each meaning on a 196-word list of the Swadesh type.

On the problem of independence, Swadesh pointed out that interaction between languages because of contact would bias estimates of t-s downward. Hattori (1953) suggested and Hymes (1960) discussed the formula

$$E(r/|M|) = e^{-1.4\lambda(t-s)}$$

as a way of taking into account parallel evolution and the effect of those meanings with lower λ than the rest of the list. The latter effect is, however, properly described by using a sum of exponentials and, for the former, it is unreasonable to expect a constant

multiplier (1.4) to express the dependence of two languages over all time. It is clear that the multiplier of $-\lambda(t-s)$ should be near zero when t is close to s and to approach 2 as t gets very large. This was noted by Gleason (1960) who rightly suggested that for all sufficiently large t, estimates of t-s could be corrected by adding a small positive constant.

One further suggestion that has been made by many authors and implemented by some, e.g. Hirsch (1954), Hattori (1957), is to attempt to construct a larger set M to provide a better (i.e. lower variance) estimate of time intervals.

The primary purpose of this paper will be to develop a formal theory of word-meaning relationship, applicable to lexical and semantic change, which incorporates most of the criticisms levelled against the Swadesh theory.

Relationship to linguistic theories

This theory is unique in that it provides a link between two previously unrelated linguistic theories, that of generative grammar, and the conventional descriptive semantics. Elsewhere (1969) we show how stochastic models, like our theory of word meaning behaviour, and Labov's (1967,1968) frequency approach to optional grammatical rules, can be derived by imposing probabilistic structure on formal grammars. On the other hand, the major phenomena and problems of descriptive and historical semantics can be elegantly formalized in terms of this same model.

PART TWO - WORD-MEANING PROCESSES

The problems of the Swadesh theory stem from its assumptions about the nature of meaning, and its oversimplified mechanism of lexical replacement. I propose a model of word-meaning relationship in which lexical replacement is a consequence of a more basic stochastic phenomenon - fluctuations in probabilities of word usage. The only aspect of a "meaning" which is relevant to this model is its representability by one or more words. I make no assumption as to the psychological or cultural nature of meaning. In fact, Thm. 4 below shows that the set of meanings as defined here can be considered a purely analytical construct. This set is completely determined by comparing word usage probabilities in certain contexts. For a natural language there is the possibility of constructing the set of meanings by empirical means (from word usage frequency data).

Whether the entities I refer to as meanings correspond well to aspects of the intuitive (or the semanticists') concept of meaning depends on whether they have important properties in common and whether they behave similarly over time. It is my thesis that these entities model the processes of historical semantics at least as closely as, say, the "meanings" of Osgood et al (1957) model psychological aspects of meaning or the "meanings" of Katz and Postal (1964) model the grammatical function of meaning.

The word-meaning relationship

The mapping type of relationship in the Swadesh theory can be represented by a bipartite graph as in Fig. 1 .

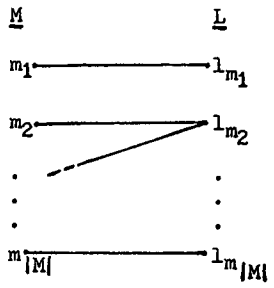


Fig. 1. Map relationship (many-to-one possible but not one-to-many).

The first generalization to be made is to allow a many-to-one (in both directions) relation, as in Fig. 2.

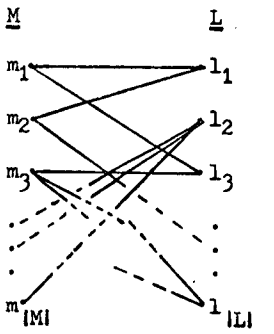


Fig. 2. Unrestricted word-meaning relationship.

The next important refinement of the model is the introduction of probability distributions on words and meanings. The frequency with which a word takes on a meaning in M has, as cited in PART 1, been recognized as important to lexicostatistics. Dyen's (1960) essay contains a clear description of how fluctuations

in these frequencies underlie the phenomena of lexical replacement. In what follows, L can be understood as in PART 1, but M is completely reinterpreted.

Definition

Let L and M be finite sets.

Let $p(\cdot,\cdot)$ be a bivariate probability distribution on $M \times L$.

Let $S_m = \{l \in L \mid p(m,l) > 0\}$.

If $S_m \neq \phi$ for all $m \in M$, and if for distinct $m,n \in M$, $S_m \neq S_n$, then M is a set of meanings on L, with respect to the distribution p, and each non-zero p(m,l) represents a word-meaning relationship between l and m.

p(m,l) should be understood as the probability that the word l will be used, and that meaning m will be intended (when no information is given about the context). The definition incorporates two restrictions on abstract meanings, neither of which is overly restrictive when considered as properties of meanings in the intuitive sense. First, if a meaning is expressible by some word or other in the lexicon, that word must have a non-zero probability of expressing it (in some context which has a non-zero probability of occurring). Second, if two meanings are to be distinct, on our level of analysis, at least one of them must be expressible by at least one word which the other is not. Fig. 3 illustrates these conditions. The latter principle, lexical distinguishability of meanings, might seem to place too much emphasis on marginal or threshold word-meaning relationships (those with very low $p(\cdot,\cdot)$).

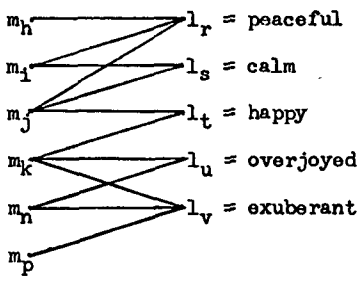


Fig. 3 . Part of word-meaning system. A line joins m and l iff $p(m,l) > 0$.

Such objections will be seen to have little importance, however, after Theorem 9 below, where M is embedded in a metric space. Here all meanings which do not differ greatly in their usage probabilities will cluster together in the metric space, and any comparisons between meanings will be in terms of the metric. Assuming lexical distinguishability facilitates the particular line of development followed here, but relaxing it (e.g. in favour of a more quantitative distinction between meanings, or in favour of a definition of meaning grouping closely related lexically distinguished entities) is not likely to radically affect the behaviour of meanings in the metric space. An important consequence of the definition of a set of meanings is

Theorem 4

Let $\mathcal{P}(L)$ be the set of subsets of L, and let M be a set of meanings on L with respect to p. If $S_m = \{l \in L \mid p(m,l) > 0\}$, then

$$m \longrightarrow S_m$$

is a one-one map from M onto a subset of $\mathcal{P}(L)$.

Proof

It need only be shown that if

$$m \longrightarrow S_m \quad , \quad n \longrightarrow S_n$$

then

$$S_n = S_m \implies m = n$$

or equivalently,

$$m \neq n \implies S_m \neq S_n \quad ,$$

but this is just the condition of lexical distinguishability in the definition.

Theorem 4 tells us that, for analytical or computational purposes, we can treat meanings as sets of words. Two meanings are distinguished by the words they do not share and are related by those they have in common. Note that the case $p(m,l) = 0$ can arise in two ways. Either $p(m,l) = 0$, for all l, in which case $S_m = \phi$ and m is not a meaning, or m is a meaning but $l \notin S_m$. From now on, no distinction will be drawn between the meaning m and the set S_m, and the latter notation will be discarded. Sometimes, an entity whose status as a meaning or not is under study, will be labelled m. If m is not a meaning, $p(m,l) = 0$, for all l; $m \notin M$; $S_m = \phi$, etc., and every attempt will be made to keep this usage unambiguous.

Interpretation of the marginal distributions

With the usage probability interpretation of $p(\cdot,\cdot)$,

$$g(l) = \sum_m p(m,l)$$

is the overall probability that l is used. The probability function g(l) underlies word-frequency distributions, e.g. those of Zipf (1945), Josselson (1953), and Juilland (1965a, 1965b).

$$f(m) = \sum_{l} p(m,l)$$

is the overall probability that m is used. This is related (at least conceptually) to the "semantic frequency lists" of Eaton (1940). Since these are probability distribution functions,

$$\sum_{m} f(m) = \sum_{l} g(l) = \sum_{m,l} p(m,l) = 1 \ ,$$

and, of course,

$$p(m,l) \geqslant 0 \ .$$

Recapitulating, a word-meaning relationship exists between m and l, or a line is drawn between m and l on a word-meaning graph like Fig. 2 or 3 , iff l can take on meaning m, which occurs iff $p(m,l) > 0$. (I.e., we require that if a word <u>can</u> take on a meaning, there is a non-zero probability that it <u>will</u> do so.) The statement $f(m) = 0$ is equivalent to saying that m is not lexically representable by elements of L, and $m \notin M$.

<u>Precision of speech</u>

In constructing a model involving the grouping of words and the distinction between meanings, provision should be made for some degree of variation to correspond to the variation which occurs in reality, from person to person and, more especially, from situation to situation. This variation is a complex effect, but

a good deal of it may be interpreted as alternation between precise and loose speech. In certain situations, and for certain topics, effective communication requires unambiguous usages, specific rather than generic terms, and other manifestations of precision which are, on the other hand, inefficient, uneconomical or just too difficult to sustain in everyday speech. This alternation may occur independently in different parts of the lexicon in a natural language, but for our model we will use a single precision parameter α. Each value of α will specify a set M_α of meanings on L. In the next few sections, the probability distributions and other entities dependent on α will be so subscripted (e.g. $p_\alpha(m,1)$, M_α).

In what manner should the system depend on α? In natural languages, as a speaker becomes more precise he draws more distinctions between words and he groups two words of similar meaning less frequently (i.e. with smaller probabilities). One measurement which is sensitive to this process in the model is the average size of the meanings

$$E_\alpha\left[|m|\right] = \sum_{m \in M_\alpha} |m| f_\alpha(m) ,$$

where $|m| = |S_m|$, the number of words connected to, representing, or simply <u>in</u>, a meaning. This measurement would be too crude, by itself, to serve as a precision parameter, since it does not distinguish between overall precision in the system and extreme precision in one part of the system but little precision in the rest. Instead, a condition should be placed on the system so that if α increases, then <u>in any part of the system</u>, this increase

would coincide with an increase in the probability weight on small meanings (i.e. $|m|$ is small) and a decrease in α would coincide with an increase on large meanings. Such a restriction may be formalized as follows.

Let $\alpha \in [0,1]$. Let $D \subset \mathcal{P}(L)$ be any set of subsets of L, meanings or not, such that

$$m \in D, \quad n \subset m \Rightarrow \quad n \in D .$$

Then it is required that

$$\sum_{m \in D} f_\alpha(m) \qquad \left(\text{or } \sum_{m \in D} \sum_{l \in m} p_\alpha(m,l)\right)$$

is monotonic and non-decreasing with α. Another way of looking at this is in terms of the lattice of subsets of L. If we choose any points in the lattice or even draw a line right across it, the probability assigned to all sets below these points, or below the line, must increase (or at least not decrease) as α, the precision, increases. A simple example will illustrate this. Let $L = \{l_1, l_2, l_3\}$. Fig. 4 depicts the lattice of subsets of L.

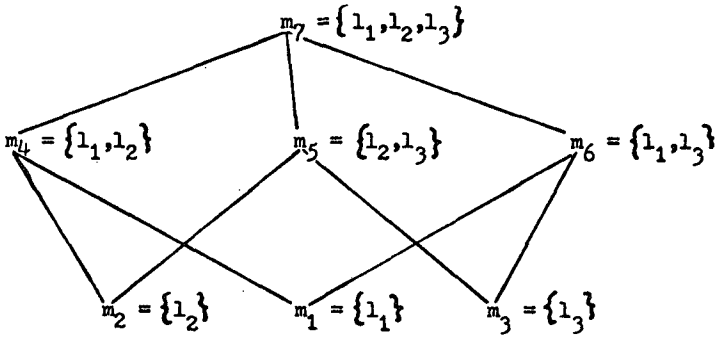


Fig. 4 . Possible meanings when $L = \{l_1, l_2, l_3\}$.

For three values of α , values of $p_{\alpha}(m,l)$ might be as in Table 1, and it is easy to verify that the precision condition holds (D can be one of $\{m_1\}$, $\{m_2\}$, $\{m_3\}$, $\{m_1,m_2\}$, $\{m_2,m_3\}$, $\{m_3,m_1\}$, $\{m_1,m_2,m_3\}$, $\{m_4,m_1,m_2\}$, $\{m_5,m_2,m_3\}$, $\{m_6,m_3,m_1\}$, $\{m_4,m_1,m_2,m_3\}$, $\{m_5,m_1,m_2,m_3\}$, $\{m_6,m_1,m_2,m_3\}$, $\{m_4,m_5,m_1,m_2,m_3\}$, $\{m_5,m_6,m_1,m_2,m_3\}$, $\{m_4,m_6,m_1,m_2,m_3\}$, $\{m_4,m_5,m_6,m_1,m_2,m_3\}$, or $\{m_1,m_2,m_3,m_4,m_5,m_6,m_7\}$).

		m_1	m_2	m_3	m_4	m_5	m_6	m_7	
$\alpha = 1$	l_1	1/4	-	-	0	-	0	0	$M_1 = \{m_1,m_2,m_5\}$
	l_2	-	1/8	-	0	1/4	-	0	$E_1[\lvert m\rvert] = 1.5$
	l_3	-	-	1/8	-	1/4	0	0	high precision
$\alpha = 0.5$	l_1	1/10	-	-	0	-	1/10	1/10	$M_{0.5} = \{m_1,m_2,m_5,m_6,m_7\}$
	l_2	-	1/10	-	0	1/5	-	1/10	$E_{0.5}[\lvert m\rvert] = 2.1$
	l_3	-	-	0	-	1/10	1/10	1/10	medium precision
$\alpha = 0$	l_1	0	-	-	0	-	1/8	1/4	$M_0 = \{m_6,m_7\}$
	l_2	-	0	-	0	0	-	1/4	$E_0[\lvert m\rvert] = 2.75$
	l_3	-	-	0	-	0	1/8	1/4	low precision

Table 1. A word-meaning system at 3 levels of precision.

The example suggests the next theprem, which confirms that the precision requirement is strong enough to imply monotonicity of the average meaning size.

Theorem 5

$E_\alpha[|m|]$ is a decreasing function of α .

Proof

$$\left.\begin{aligned} \text{Let } a(i) &= \sum_{|m|=i} f_\alpha(m) \\ b(i) &= \sum_{|m|=i} f_\beta(m) \end{aligned}\right\} \qquad \begin{aligned} &i = 1,2,\ldots,|L| \ , \\ &\beta < \alpha \end{aligned}$$

Then $a(\cdot)$ and $b(\cdot)$ are probability distributions on the integers, where $a(i)$ is the probability that an unspecified meaning will contain i words. Consider

$$D_j = \left\{ m \subseteq L \;\middle|\; |m| \leq j \right\} .$$

Clearly $m \in D$, $n \subset m \Rightarrow n \in D$. Then the precision condition requires

$$\sum_{m \in D_j} f_\alpha(m) \geqslant \sum_{m \in D_j} f_\beta(m) \ .$$

Therefore

$$\sum_{|m| \leq j} f_\alpha(m) \geqslant \sum_{|m| \leq j} f_\beta(m) \ ;$$

$$\sum_{i=1}^{j} a(i) \geqslant \sum_{i=1}^{j} b(i) \ .$$

Since $a(\cdot)$ and $b(\cdot)$ are probability distributions,

$$\sum_{i=1}^{|L|} a(i) = \sum_{i=i}^{|L|} b(i) = 1 \ ,$$

$$\sum_{i=j}^{|L|} a(i) = 1 - \sum_{i=1}^{j-1} a(i) \leqslant 1 - \sum_{i=1}^{j-1} b(i) = \sum_{i=j}^{|L|} b(i) \ .$$

Then

$$\sum_{j=1}^{|L|} \sum_{i=j}^{|L|} a(i) \leq \sum_{j=i}^{|L|} \sum_{i=j}^{|L|} b(i) \quad ;$$

$$\sum_{j=1}^{|L|} ja(j) \leq \sum_{j=1}^{|L|} jb(j) \quad ;$$

$$E_{\alpha}[|m|] \leq E_{\beta}[|m|] \quad ,$$

since $a(\cdot)$ and $b(\cdot)$ are the probability distributions of the values of $|m|$.

Regularity conditions

We have imposed a condition on the $p_{\alpha}(m,l)$ so that the probability weight must flow down the lattice of subsets of L as α increases. It would be desirable, from the viewpoints of model realism and analytical convenience, to have this "flow" behave in as continuous a manner as possible. It would be most convenient if the $p_{\alpha}(m,l)$ were required to be continuous functions of α, but there are good reasons to relax this somewhat.

Again trying to model natural languages, it would be realistic to require that the following process may occur in the system. Suppose a meaning $m' \in M_o$ is connected to $k,l_1,l_2,\ldots,l_r \in L$. (In our earlier notation, $S_{m'} = \{k,l_1,l_2,\ldots,l_r\}$, in our current notation $m' = \{k,l_1,l_2,\ldots,l_r\}$, $p_o(m',k) > 0$, $p_o(m',l_i) > 0$, $\forall\, l_i \in m'$). As α increases, the values of all the $p_{\alpha}(m',l_i)$ fluctuate but remain greater than some positive value, except for $p_{\alpha}(m',k)$ which gradually drops to zero at α_o. In terms of speech behaviour, the words $k, l_1,l_2,\ldots,l_r$ are used interchangeably (in certain

contexts) to mean m', when precision is low. As precision increases, $l_1, l_2, \ldots, l_r$ continue to be interchangeable but k is seldom usable in this sense and, at α_o, never. It is most important in what ensues to understand that the set $m' = \{k, l_1, l_2, \ldots, l_r\}$ ceases to be a meaning when the precision is α_o,

i.e.

$$m' \in M_\alpha, \quad \alpha < \alpha_o$$

$$m' \notin M_{\alpha_o}.$$

It is, however, most natural that $m = m' - \{k\} = \{l_1, l_2, \ldots, l_r\}$ be a meaning at α_o, since the interchangeability of these words is not necessarily dependent on the behaviour of k. Hence, if any psychological interpretation is to be attached to the set of abstract meanings in our model, it must be realized that as precision changes, the abstract label attached to a psychological or cognitive entity may suddenly change as lexical representability of that entity changes. If this seems strange behaviour for a symbolic system, it should seem less so later, when the M_α are embedded in a metric space and the relative position of meanings in this space becomes more important than the letters that identify them.

Returning to quantitative considerations, since m' ceases to be a meaning at α_o and m suddenly takes over its role, it is necessary that $p_\alpha(m', l_1), \ldots, p_\alpha(m', l_r)$ drop discontinuously to zero at α_o and $p_\alpha(m, l_1), \ldots, p_\alpha(m, l_r)$ jump to compensate.

We must, therefore, accept certain discontinuities of this sort in the model. For simplicity's sake, we restrict

occurrences such as this so that only <u>one</u> $p_\alpha(m,l)$ may drop continuously to zero at any particular value of α_o ($p_\alpha(m',k)$ in the example above). This is in fact a weak restriction, in that we can approximate situations where N of the $p_\alpha(m,l)$ go to zero at α_o by having them do this one at a time, at α_o, $\alpha_o + \epsilon$, $\alpha_o + 2\epsilon$, . . . , $\alpha_o + (N-1)\epsilon$ for arbitrarily small ϵ .

An appropriate continuity-discontinuity condition may be most economically phrased as in condition (iii) in the next section.

<u>Summary of development thus far</u>

We assume that there exists a finite set L (the set of words) and for each $\alpha \in [0,1]$ a finite set M_α (a set of meanings on L) and a bivariate probability distribution p_α on $M_\alpha \times L$ such that

$$\sum_{m \in M_\alpha} \sum_{l \in L} p_\alpha(m,l) = 1 .$$

The elements of M_α are in one-one correspondence with certain non-empty subsets of L.

$$m \longleftrightarrow S_m \Leftrightarrow p_\alpha(m,l) > 0, \ \forall l \in S_m .$$

This correspondence enables us to unambiguously identify S_m with m, and we may rewrite the above condition

(i) $\quad p_\alpha(m,l) > 0 \Leftrightarrow l \in m$ and $m \in M_\alpha$

As α varies between zero and 1, the following conditions must hold:

(ii) If $D \subset \mathcal{P}(L)$ such that $m \in D$, $n \subset m \Rightarrow n \in D$, then

$\sum_{m \in D} \sum_{l \in m} p_\alpha(m,l)$ is monotone non-decreasing with α.

(iii) The $p_{\alpha}(m,l)$ are continuous functions of α only where M_{α} is fixed. M_{α} changes at α_0 only as a result of discontinuities occurring, for unique m, and unique $k \notin m$, to all of

$$p_{\alpha}(m,l) \ , \ p_{\alpha}(m + \{k\},l) \ ; \qquad l \in m \quad (\text{for } m + \{k\}, \text{ read } m \cup \{k\})$$

but

$$p_{\alpha}(m,l) + p_{\alpha}(m + \{k\},l) \text{ is continuous, for all } l \in L.$$

Before enunciating the continuity and discontinuity condition (iii), we described the desired behaviour of some of the functions $p(.,.)$ at a point where the condition is relevant. We can prove that this condition implies this behaviour.

Theorem 6

In the system as described above, if α_0 is a point where M_{α} changes, then $p_{\alpha}(m + \{k\},k)$ (as in condition (iii) above) is continuous at α_0, and if it goes to zero at α_0 it is the only such function.

Proof

By condition (iii),

$$p_{\alpha}(m,l) + p_{\alpha'}(m + \{k\},l) \text{ is continuous at } \alpha_0 \text{ for all } l \in L.$$

Therefore

$$p_{\alpha'}(m,k) + p_{\alpha}(m + \{k\},k) \text{ is continuous at } \alpha_0.$$

But

$$p_{\alpha'}(m,k) \equiv 0 \quad \text{since } k \notin m \ ;$$

hence the continuity of $p_{\alpha}(m + \{k\},k)$.

Now if any other $p_\alpha(n,l')$ goes to zero at α_o, n ceases to be a meaning and M_α changes as a result. This contradicts condition (iii) unless $n = m$ or $m + \{k\}$, in which case discontinuities are prescribed by the same condition.

Existence and local behaviour

The next theorem gives assurance that the conditions on the components of a word-meaning system, as developed so far, are not contradictory. The proof consists of a construction of a particular system (which is otherwise uninteresting) and is presented as Appendix 3 in Sankoff (1969).

Theorem 7

Word-meanings systems exist.

Specifically, it is possible to construct a word-meaning system using any finite set

$$L = \{l_1, l_2, \ldots, l_{|L|}\}.$$

The regularity conditions are strong enough, however, so that aside from continuous variation in the $p_\alpha(\cdot,\cdot)$, only certain types of change in M_α are possible.

Theorem 8

Suppose M_α changes at α_o. Let M^-, M^+ be the state of M_α in small enough intervals to the left and right of α_o, respectively. Then one of A, B or C must hold.

A. For a unique m, and unique $k \notin m$, as in condition (iii),

$$m + \{k\} \in M^-, \quad m + \{k\} \notin M^+; \quad m \in M^-, \quad m \in M^+,$$

represented by

$$(\in, \notin; \in, \in), \quad p_{\alpha_o}(m + \{k\}, k) = 0,$$

B. $$(\in, \in; \notin, \in), \quad p_{\alpha_o}(m + \{k\}, k) > 0,$$

C. $$(\in, \notin; \notin, \in), \quad p_{\alpha_o}(m + \{k\}, k) = 0.$$

Proof

There are 16 ways of filling four places with $\in$ or $\notin$.

$(\in, \in; \in, \in)$, $(\notin, \notin; \notin, \notin)$, $(\in, \in; \notin, \notin)$ and $(\notin, \notin; \in, \in)$ involve no change in M_α.

$(\notin, \notin; \in, \notin)$, $(\notin, \notin; \notin, \in)$, $(\notin, \in; \notin, \notin)$ and $(\in, \notin; \notin, \notin)$ imply either $p_\alpha(m,1) \equiv 0$ or $p(m + \{k\}, 1) \equiv 0$ near α_o, and hence have no discontinuity.

In $(\notin, \in; \notin, \in)$ and $(\in, \notin; \in, \notin)$, $p_\alpha(m,1)$ and $p_\alpha(m + \{k\}, 1)$ "jump" in the same direction, hence their sum could not be continuous.

$(\notin, \in; \in, \notin)$, $(\notin, \in; \in, \in)$ and $(\in, \in; \in, \notin)$ violate condition (ii).

There remain only the three possibilities,

A. $m + \{k\}$ disappears, $\lim_{\alpha \to \alpha_o} p_\alpha(m + \{k\}, k) = p_{\alpha_o}(m + \{k\}, k) = 0$.

B. m appears, $m + \{k\}$ in M^- and M^+, $p_{\alpha_o}(m + \{k\}, k) > 0$.

C. m appears, $m + \{k\}$ disappears, $p_{\alpha_o}(m + \{k\}, k) = 0$.

These three situations are illustrated in Fig. 5A, 5 B and 5C.

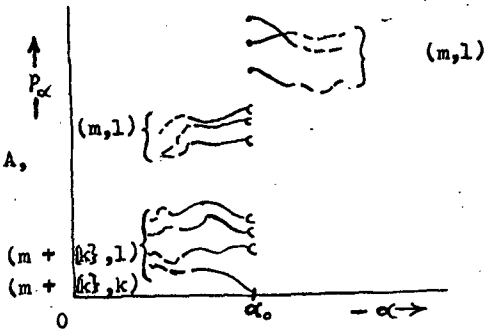


Fig. 5A;
Possibility A,
Thm. 8

N.B. right continuity instead of left continuity would be equally possible here.

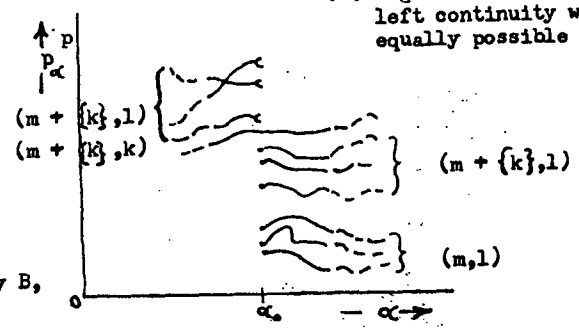


Fig. 5B;
Possibility B,
Thm. 8

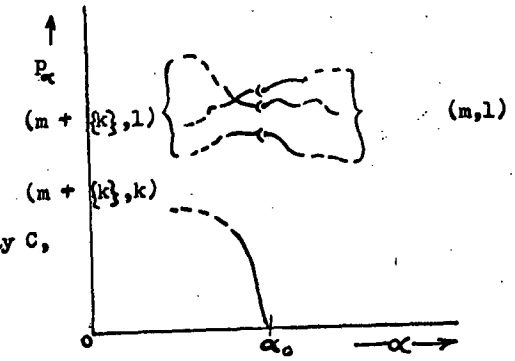


Fig. 5C;
Possibility C,
Thm. 8

Meanings as points in a metric space

The idea of distances between meanings is not new, and there have been a number of attempts to operationalize this concept. We shall examine a very natural way of defining such a distance for the meanings in a word-meaning system in terms of the functions $p_\alpha(m,l)$.

Definition

Let $m \in M_\alpha$, $n \in M_\beta$

$$d_{\alpha,\beta}(m,n) = \tfrac{1}{2}\sum_{l \in L} \left| \frac{p_\alpha(m,l)}{f_\alpha(m)} - \frac{p_\beta(n,l)}{f_\beta(n)} \right|$$

Theorem 9

$d_{\alpha,\alpha}$ defines a metric on M_α.

Proof

The norm $\sum|\cdot|$ defines a metric on probability distributions

$\frac{p_\alpha(m,l)}{f_\alpha(m)}$ defines a probability distribution on L.

It remains to prove that two such $m \in M_\alpha$ do not define the same distribution. But this follows from the fact that each $m \in M_\alpha$ defines a unique subset of L such that $p_\alpha(m,l) > 0$.

Remark

If as β increases beyond α, $p_\beta(m,l)$ changes, $d_{\alpha,\beta}(m,m)$ will have a minimum value at $\beta = \alpha$ and will increase for β on either side of α. In a neighbourhood of α, $d_{\alpha,\beta}(m,m)$ for fixed

m, then, measures distance from α . This relevance of d to the parameter as well as to the meanings will become important in later sections.

Theorem 10

If $M_\alpha = M_I$, $M_\beta = M_J$ for $\alpha \in I$, $\beta \in J$, two intervals, and if

$m \in M_I$, $n \in M_J$

$d_{\alpha,\beta}(m,n)$ is continuous on $I \times J$.

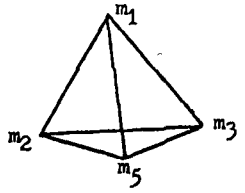
Proof

This follows from the continuity of the p_α on such intervals and from the fact that d is a continuous function of such p_α .

As α changes, the points in M_α move continuously. When M_α changes, two (at most) points experience a sudden shift in position with respect to the rest of the points. This may involve the creation or annihilation of these points. When α is close to 1, there will be few words in common between two meanings, on the average, and hence the distance between them will be close to 1. When α is close to zero, on the other hand, the reverse is true, and distances will tend toward zero. This rather succinct comparison of precise versus loose usage accords well with more intuitive notions of precision of speech. Fig. 6 and Table 2 present, as illustrations, the distances in the metric spaces defined by the 3-word system described earlier in this chapter.

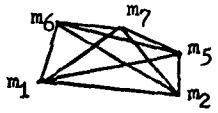
$\alpha = 1$

	m_1	m_2	m_3
m_2	1		
m_3	1	1	
m_5	1	$\frac{1}{2}$	$\frac{1}{2}$



$\alpha = 0.5$

	m_1	m_2	m_5	m_6
m_2	1			
m_5	1	1/3		
m_6	$\frac{1}{2}$	1	2/3	
m_7	2/3	2/3	1/3	1/3



$\alpha = 0$

	m_6
m_7	1/3

Table 2. $d_{\alpha,\alpha}(\cdot,\cdot)$ for system of Table 1.

Fig. 6. 2-dimensional visualization of distances in Table 2.

Diachronic word-meaning systems

We have developed, in some detail, a synchronic (i.e. at a fixed point in time) theory of words and meanings. It remains to show what relevance this has to historical linguistics and lexicostatistics.

As Ullman (1957) remarks:

> "The two [semantic relationship, simple or multiple, and semantic change] are interdependent, one being the projection of the other on a different plane. The functional analysis of meaning will entail therefore a definition of semantic change along similar lines. If a meaning is conceived as a reciprocal relation obtaining between name and sense [word and meaning], then a semantic change will occur whenever a new name becomes attached to a sense and/or a new sense to a name." (p.171)

and, as he points out, word-meaning phenomena at a fixed time have parallels in processes of change over time.

In our particular model, changes in the system as the precision parameter changes will provide the prototype for change with time.

Definition

A word-meaning system history is a word-meaning system with $\alpha \in [0,1]$ replaced by $t \in [0,T]$ (time parameter) and with condition (ii) relaxed entirely. Condition (iii) is changed so that if k and m are given as before, and $m + \{k\}$ is a new meaning or if m disappears starting at t_o, there are discontinuities in $p_t(m + \{k\},k)$ and $p_t(m + \{k\},l) + p_t(m,l)$ for one $l \in m$, but

$$p_t(m + \{k\},k) + p_t(m + \{k\},l) + p_t(m,l)$$

is continuous.

Although an adjustment to the construction necessary for Thm. 8 could adapt the existence proof of word-meaning systems to that of word-meaning system histories, it will be simpler to leave existence to be implicit in the constructions carried out later.

Theorem 11

Suppose M_t changes at t_o. Let M^-, M^+ be as in Thm. 8. Then one of A, B, C, A', B', C' holds.

A. $(\in, \notin; \in, \in)$, $p_{t_o}(m + \{k\},k) = 0$,

A'. $(\in, \in; \in, \notin)$, $p^+_{t_o}(m + \{k\},k) > 0$, $p^-_{t_o}(m + \{k\},k) = 0$,

B. $(\in, \in; \notin, \in)$, $p_{t_o}(m + \{k\}, k) > 0$,

B'. $(\notin, \in; \in, \in)$, $p_{t_o}(m + \{k\}, k) > 0$,

C. $(\in, \notin; \notin, \in)$, $p_{t_o}(m + \{k\}, k) = 0$,

C'. $(\notin, \in; \in, \notin)$, $p_{t_o^+}(m + \{k\}, k) > 0$, $p_{t_o^-}(m + \{k\}, k) = 0$.

Proof

A, B and C were the three possibilities admitted in Thm. 8. The only new restriction applies when $m + \{k\}$ appears or m disappears at t_o, and therefore applies to none of the three. A', B' and C' were discarded in Thm. 8 because they violated condition (ii). The new condition (iii) applies to all of these cases. In A' and C', $m + \{k\}$ appears so $p_t(m + \{k\}, k)$ must jump from zero at t_o.

The cases A, B and C are still represented by Fig. 5A, 5B and 5C, with α replaced by t. Cases A', B' and C' would be represented by mirror images of these three figures, except that $p_t(m + \{k\}, k)$ must exhibit a discontinuity at t_o, and one of the $p_t(m + \{k\}, l)$ must compensate for this.

Remark

The asymmetry with respect to time of the conditions for changes in M_t may be interpreted as follows. The probability that a word may be used for a meaning may drop to zero continuously, but it may not increase from zero continuously. Instead, it must at some time jump to some finite value. This distinction is not too important to the overall characteristics of word-meaning system

histories, but we note it because the particular type of histories we shall study have this property.

The development of the metric d in the previous section carries over completely when the time parameter replaces the precision parameter, except, of course, that there is no longer any necessary trend in the average distance between meanings as t increases.

Anticipating some of our later discussion, consider the case where all meanings consist of exactly one word, as in the Swadesh model. In this case, letting s and t be time as in Thm. 1,

$$d_{s,t}(m,n) = 1 - \delta(k,l)$$

where $m = \{k\} \in M_s$, $n = \{l\} \in M_t$. d then, is in a certain sense a generalization of the cognation indicator δ.

Word-meaning processes

So far, changes in M_α or M_t have been deterministic as the value of the parameter changes. (Even though the p_α or p_t are probability functions, we have not studied further properties of the random variables which are distributed according to these functions, and we will not do so. In linguistic terms, we are still dealing with langue and not parole.) To generalize the Swadesh theory, and to provide a realistic model, we must take into account unpredictability of lexical and semantic change. In probability theoretical terms, we must impose a probability measure on the set of all possible histories. We shall not do this explicitly. Rather we shall assume it is possible, and assume that the examples we construct by specifying local behaviour are well-behaved in terms of an underlying probability measure space.

Definition

A word-meaning process is a set of word-meaning system histories indexed by $\omega \in \Omega$ where $(\Omega, \mathfrak{F}, P)$ is a probability measure space. This means that any event or combination of events in which we may be interested is represented by a set, A, of histories ($\omega \in A$) where A is a member of the σ-algebra $\mathfrak{F}$, and where P(A) is well-defined for all $A \in \mathfrak{F}$.

A word-meaning process based on Brownian motion

To construct the word-meaning process which is the best model for natural languages would require the operationalizing of definitions, collection of much data and its statistical analysis. At present, we shall attempt only an heuristic investigation.

In PART 1, we emphasized the basic unpredictability of change in the word-meaning relationship. In terms of our model, (and considering only small intervals of time) this means that for $t > s$,

$$E\left[p_t(m,l) - p_s(m,l)\right] = 0$$

Furthermore, it should not be possible to predict the future behaviour of individual $p_t(m,l)$ from trends established in the past: for any $t > s_1 > s_2 > \ldots > s_r$

$$P\left[p_t(m,l) \mid p_{s_1}(m,l), p_{s_2}(m,l), \ldots, p_{s_r}(m,l)\right]$$

$$= P\left[p_t(m,l) \mid p_{s_i}(m,l)\right], \text{ the Markov condition.}$$

But these two conditions and the continuity conditions on p_t indicate that the local behaviour of $p_t(m,l)$ should resemble a

diffusion process, with zero drift. The simplest such process is the well-known Brownian motion, whose behaviour characteristics change neither with time, t, nor with position, x.

We proceed to construct a word-meaning process satisfying these properties. Let $(L, M_\alpha, p_\alpha(\cdot,\cdot))$ be a word-meaning system for a fixed α. For $t = 0$, let $p_t(m,l) = p_\alpha(m,l)$, $M_t = M_\alpha$. Let

$$n_o = E_\alpha[|m|] \cdot |M| \quad .$$

n_o is the number of word-meaning relationships in the system. Let $x_1(t), x_2(t), \ldots, x_{n_o}(t) : t \geq 0$ be n_o sample paths of a Brownian motion process, chosen independently, and $\bar{x}(t) = \frac{1}{n_o} \sum_{i=1}^{n_o} x_i(t)$. Let $y_i(t) = x_i(t) - \bar{x}(t)$. The y_i are also Brownian sample paths, but are no longer completely independent in that

$$\sum_{i=1}^{n_o} y_i(t) = \sum_{i=1}^{n_o} x_i(t) - \sum_{i=1}^{n_o} \bar{x}(t)$$

$$= n_o\bar{x}(t) - n_o\bar{x}(t)$$

$$= 0 \, .$$

Let

$$p_t(m,l) = p_o(m,l) + y_i(t) \, ,$$

where $i = i(m,l)$ is determined beforehand. Then p_t is continuous in $[0,T]$ with probability 1. We must ensure that $p_t(\cdot,\cdot)$ is a probability distribution.

$$\sum_{m\in M_t}\sum_{l\in m} p_t(m,l) = \sum_{m\in M_o}\sum_{l\in m} p_o(m,l) + \sum_{i=1}^{n_o} y_i(t)$$

$$= \quad 1 \qquad\qquad + \qquad 0$$

$$= \quad 1$$

It is not necessarily true, however, that $p_t(m,l) \geq 0$, since $y_i(t)$ may be negative. To adjust for this, let

$$\tau = \sup\{t \mid p_s(m,l) > 0,\ l\in m,\ m\in M_o, \forall s,\ 0\leq s<t\}$$

In other words, all the $p_t(m,l)$ are positive before τ . Then with probability 1, there is a unique $m'\in M_o$, $k\in m'$, such that

$$\lim_{t\to\tau} p_t(m',k) = p_\tau(m',k) = 0 .$$

But this is reminiscent of case A or C in Thm. 11 (see Fig.5), where one word in a meaning loses its ability to be grouped with the others. Then all the $p_\tau(m',l)$ should drop to zero and all the $p_\tau(m'-\{k\},l)$ jump to compensate. According to whether $m' - \{k\} \in M_o$ or not, we have case A or case C respectively. Then it is a simple matter to determine M_t. Now, change the definition of all the $p_t(m,l)$ for $t > \tau$, by calculating

$$n_\tau = E_\tau[|m|]\cdot|M|$$

and starting over as for $t = 0$.

Continuing this way until $t = T$, we ensure that no $p_t(m,l)$ ever drops below zero.

We now have a word-meaning process, but not a very healthy one, in that $|M_t|$ decreases monotonically with t.

To counteract this, we superimpose another process on our construction. We select points in $[0,T]$ at random as follows: The probability of no points being selected in an interval $[t,t+\Delta t]$ is

$$1 - \mu \Delta t + h$$

where $h/\Delta t \to 0$ as $\Delta t \to 0$). At each point τ selected, randomly choose $m \in M_\tau$ and $k \in L$, $k \notin m$. If $f_\tau(m + \{k\}) > 0$, the system undergoes a change as in case B' of Thm. 11. If $f_\tau(m + \{k\}) = 0$, the system undergoes an A'-type change with probability π, and a C'-type change with probability $1 - \pi$. In each of these cases an element $l \in m$ must be selected at random so that $p_t(m + \{k\},k)$ + $p_t(m + \{k\},l)$ and $p_t(m,l)$ are discontinuous but their sum is continuous. The size of the discontinuity is uniformly distributed between 0 and $p_t(m,l)$. In case A', we assume that after this latter step is done, each element in m loses a random (but fixed) proportion of its probability weight to the corresponding element in $m + \{k\}$.

This ends the construction. Note that case B of Thm. 11 does not occur in this example.

Had we not insisted on the extra discontinuities (in $p_t(m + \{k\},k)$) in the definition of a word-meaning system history, we would not have been able to use the Brownian motion. If $p_\tau(m + \{k\},k) = 0$, and if we add a Brownian motion $y_i(t)$, $p_{\tau+t}(m + \{k\},k)$ will be zero again for arbitrarily small t. Hence we must start $p(m + \{k\},k)$ at a finite value, i.e. discontinuously.

Stability

The first thing we would like to know about our system is whether or not it is degenerate. Does it tend to degenerate into a single word-meaning relationship with $p(\{1\},1) = 1$? Does the number of meanings $|M_t|$ or word-meaning relationships n_t tend to grow without bounds as T and $|L|$ increase?

By increasing μ to a high enough value, we can increase the rate at which new word-meaning relationships are created, and hence reduce the time during which n_t is at low values. At the same time, n_t cannot increase without bound, since as the number of word-meaning relationships increases, the probability weight attached to each must decrease, on the average. Hence a higher proportion of relationships tends to be annihilated per unit time, as in cases A and C of Theorem 11 . A rigorous proof that n_t is neither too large nor too small most of the time does not seem easy to achieve, simply because of the complication of the model and the importance of the initial conditions. In any case, such a result would be rather weak. It seems likely, and we will present evidence from sampling experiments to support this, that as $t \longrightarrow \infty$, n_t tends to vary about an equilibrium mean value according to an equilibrium distribution, depending only on the system parameters μ and π .

Regularity of change in $(M_t, d_{t,t})$

For each α, a word-meaning system (relatively complicated) was associated with a relatively simple metric space $(M_\alpha, d_{\alpha,\alpha})$. The meanings corresponded to points in the metric space and the distance between meanings varied continuously almost everywhere with respect to α.

The same remarks hold true, of course, for the analogous metric spaces $(M_t, d_{t,t})$. As t increases each meaning moves continuously except at certain points where it can split into two or merge with another meaning. At such times there are discontinuities in $d_{t,t}$, but these are not usually very large. This regularity of motion ensures that we have some sort of correspondence between the sets of meanings at two distinct times. In the Swadesh model, a well defined correspondence is assumed, in terms of the universal set of meanings. If we do not postulate anything of this nature, since it must necessarily refer to cultural universals, not linguistic universals, it becomes more difficult to make word-meaning comparisons at two points in time. Indeed, if after a point in time, s, a meaning loses a lexical representation (as in case C in Thm. 11), it ceases to exist, in our technical sense, and others close to it take up its semantic load - and we must, at the very least, assume some rule for choosing a related or close meaning, for all later points in time, if we are to make lexical comparisons. The intuitive use of the term "close" gives a clue as to the appropriate choice - the

meaning n which minimizes

$$d_{s,t}(m,n) .$$

This has one important desirable property for such a rule. For t very close to s, in most cases n will, of course, be m itself. $d_{s,t}(m,m) = d_{s,t}(m,n)$ will then be the sum of the absolute values of quantities approximately proportional to Brownian motion (see definition of $d_{\alpha,\delta}$) and hence will, on the average (or in expectation) increase monotonically. $1 - d_{s,t}(m,m)$ will decrease monotonically. After a discontinuity $1 - \min_{n \in M_t} d_{s,t}(m,n)$ will continue to decrease. Since it is the processes of lexical loss and lexical replacement which are responsible for this decrease, $\frac{1}{|M_s|} \sum_{m \in M_s} (1 - \min_{n \in M_t} d_{s,t}(m,n))$ is a likely candidate to replace Swadesh's $\frac{1}{|M|} \sum_{m \in M} \delta(k,l)$ as a lexicostatistic indicator. We will so use it, keeping in mind that it does not involve any pan-cultural or pan-linguistic method of selecting universal meanings to compare. If such a method existed (and it does, approximately speaking, e.g. the Swadesh list) our indicator must necessarily provide an upper bound for any indicator of the form $1 - d_{s,t}$.

Simulating word-meaning processes

A complete, purely mathematical treatment of the Brownian-based word-meaning system would be difficult, and no results analogous to Theorems 1 - 3 are yet available. On the other hand, by choosing a set of $p_o(m,l) = p_{\alpha_o}(m,l)$ from a word-meaning system,

and fixing μ and γ it is possible to simulate the behaviour of the bivariate functions $p_t(m,l)$. A sample from a number of simulated histories might produce some hint of what the corresponding theorems might be. The remainder of this chapter consists of an account of such an experiment.

A simulation program

A computer program (see Fig. 7) was written to provide word-meaning histories sampled from the Brownian-based process (actually an approximation of this process).

The program accepts as initial data T (the length of the simulation), parameters I (from which μ can be calculated), γ and θ ; and two matrices N(i,j) and P(i,j) with |M| rows and 20 columns. The row index i identifies the meaning being considered, and the non-zero N(i,j) identify the words connected to that meaning (up to 20). P(i,j) then, represents $p_0(m_i,l_k)$ of the system where $N(i,j) = l_k$. (It is more economical to store two $|M| \times 20$ matrices than one $|M| \times |L|$ matrix if $|L| > 40$.)

To approximate the Brownian motion from time t=0 to t=I, one part of the program adds a normal random variable to each of the non-zero P(i,j). These variables have mean zero and variance I and their sum is zero, as specified in the model. Each of these P(i,j) is then examined to see whether it has dropped to zero or below. If it has, the rest of the non-zero P(i,k) are set to zero as in cases A and C of Thm.11 and P(h,g) are increased by compensating amounts where h and g are the appropriate meanings and words for the cases.

Another part of the program picks an integer according to a Poisson random variable, with mean 10, and this variable represents the number of cases A', B' and C' which have occurred during the time increment I. Hence $\mu \approx 10/I$. For each of these occurrences the program then allows a choice of whether the word (see Thm 11) is to be a new word (borrowing) or a word that is already used for another meaning (this choice is made at random with probabilities θ , $1-\theta$). The meaning m and the word $l \in m$ (again as in Thm 11) are chosen at random. If necessary (not in case B') a random choice is made between A' and C' according to parameter π , and if necessary (case A') the allocation of probabilities between m and m + {k} is decided by choosing a random number (uniformly distributed between 0 and p(m,l)).

The program then provides for the examination of the system to calculate the resulting values of $|M|$, $|L|$, $N(\cdot,\cdot)$, $P(\cdot,\cdot)$ and n_t and it prints these out. From this point it returns to the Brownian motion section and sets $t = 2I$ and adds another batch of normal variables with variance I, etc.

The above is only a summary of the program. Other routines relabel words or meanings so that they may be stored and examined economically, and others allocate any "negative probability" from Brownian paths going <u>below</u> zero during a time increment (when in theory they are only allowed to go as far as zero) among the other word-meaning relationships of the meaning involved. Finally, in the version represented in Fig 7 there is a routine which compares the word-meaning system at time t with the initial word-meaning system (at

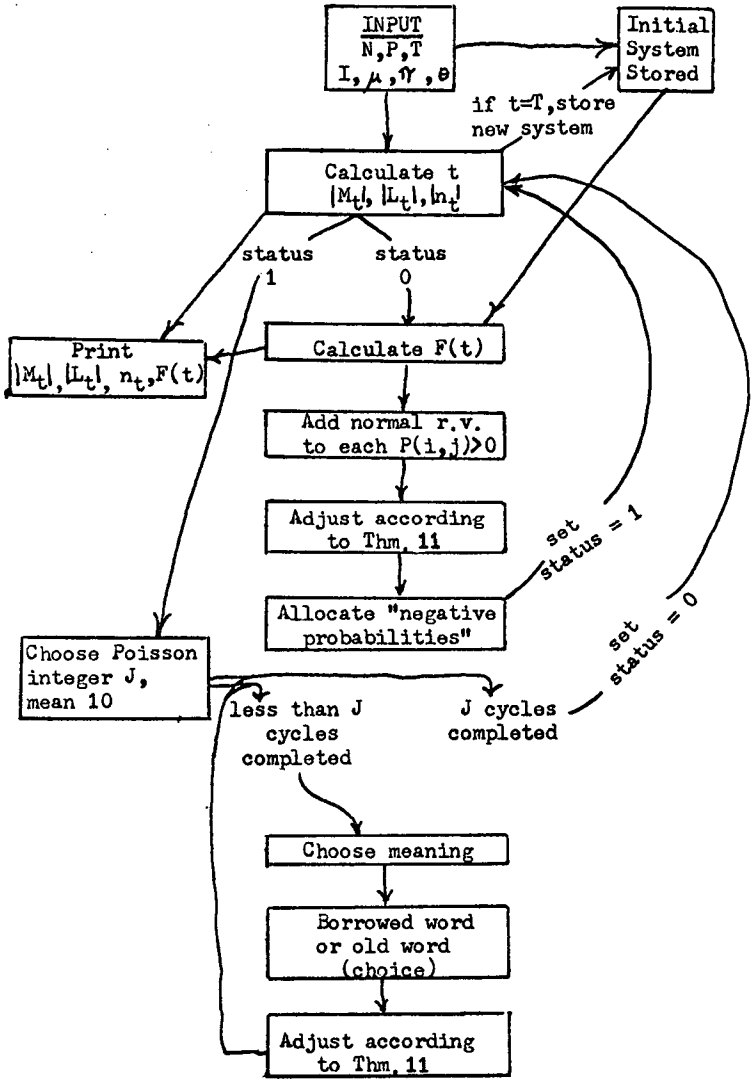


Fig. 7. Flow chart for simulation program.

time t=0) according to our lexicostatistic indicator

$$F(t) = \frac{1}{|M_0|} \sum_{m \in M_0} (1 - \min_{n \in M_t} d_{0,t}(m,n)) .$$

Results of a simulation experiment

To illustrate the properties of a Brownian-based process, we will present the results on 12 sample histories of a simulated process with the parameters fixed.

These histories were obtained as follows. For the first, the initial system was represented as in Fig. 8 .

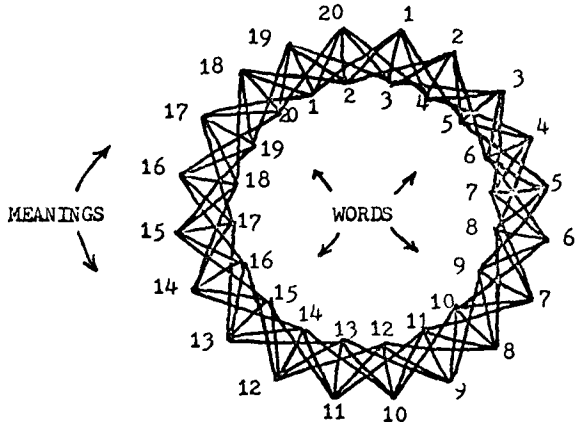


Fig. 8. Initial word-meaning system.

where each line between an m and an l represents $p_0(m,l) = .01$. Here $|M|=20$, $|L|=20$, $n_0=100$. $[0,T]$ was divided into 100 increments, and details of the system were extracted at time T

and these were used to provide the initial system for the second history. This general procedure was followed thereafter with the final status of some of the systems serving as the initial systems for others.

Stability and equilibrium distributions

As we conjectured earlier, the system moves rather quickly to equilibrium and we can trace this in the first history. Fig. 9 shows how $|M_t|$, $|L_t|$ and n_t tend to approach and then oscillate around an equilibrium value.

The "equilibrium" distributions in Fig. 10 are calculated from all the values of the system characteristics, at all points in time, of the last 11 histories (since the first history started with a non-equilibrium state).

Zipf's Law

It is a property of natural languages that, aside from the few most frequent words, the frequency of occurrence of a word G(l) and the rank order of this frequency, H(l), are related approximately as

$$G(l) = Ce^{-KH(l)}$$

where C and K are constants.

Our word-meaning systems do not have as many words as natural languages. Nevertheless, it is possible to calculate the probabilities (not frequencies) g(l) from

$$g(l) = \sum_{\substack{m, \\ l \in m}} p_t(m,l) \ .$$

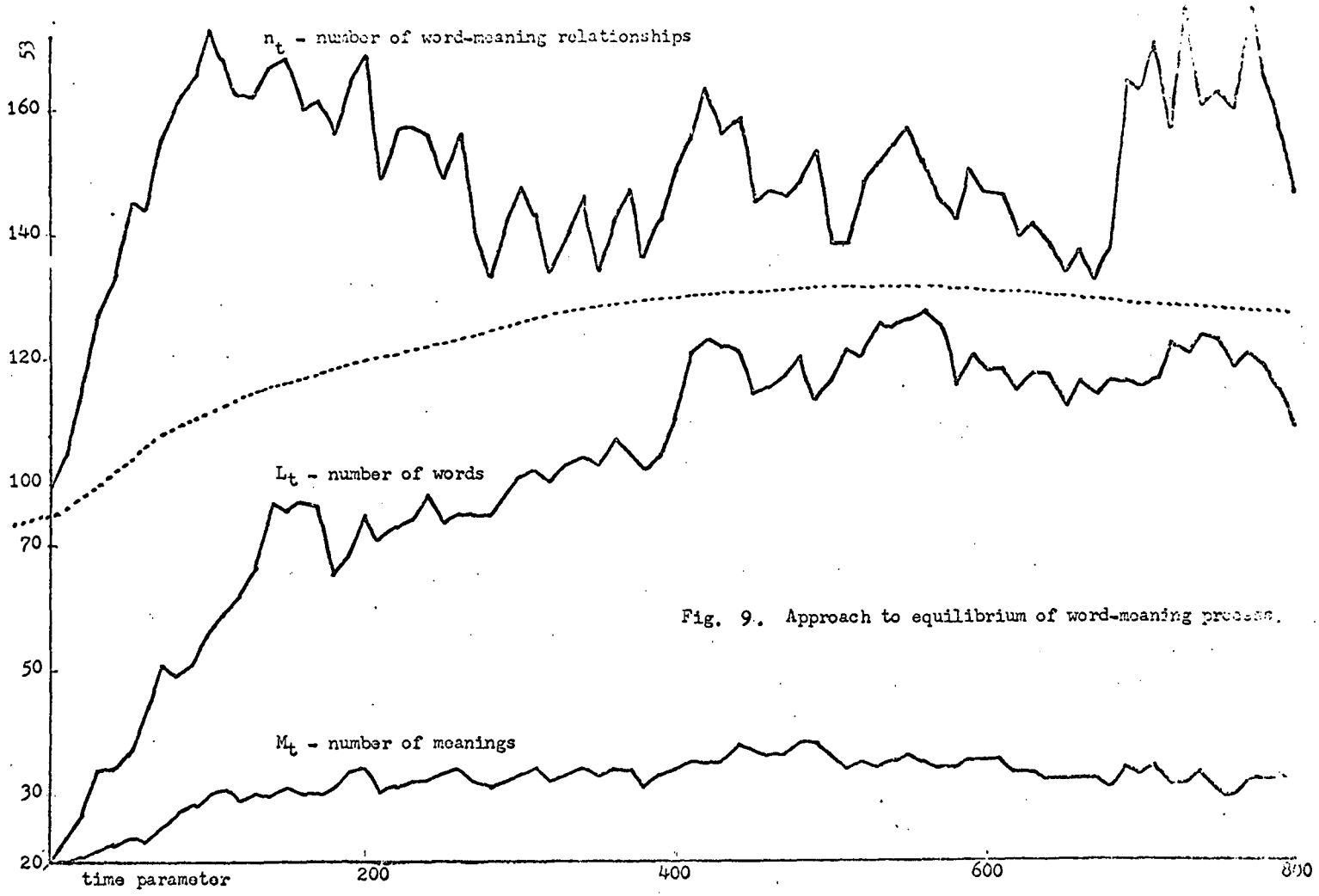
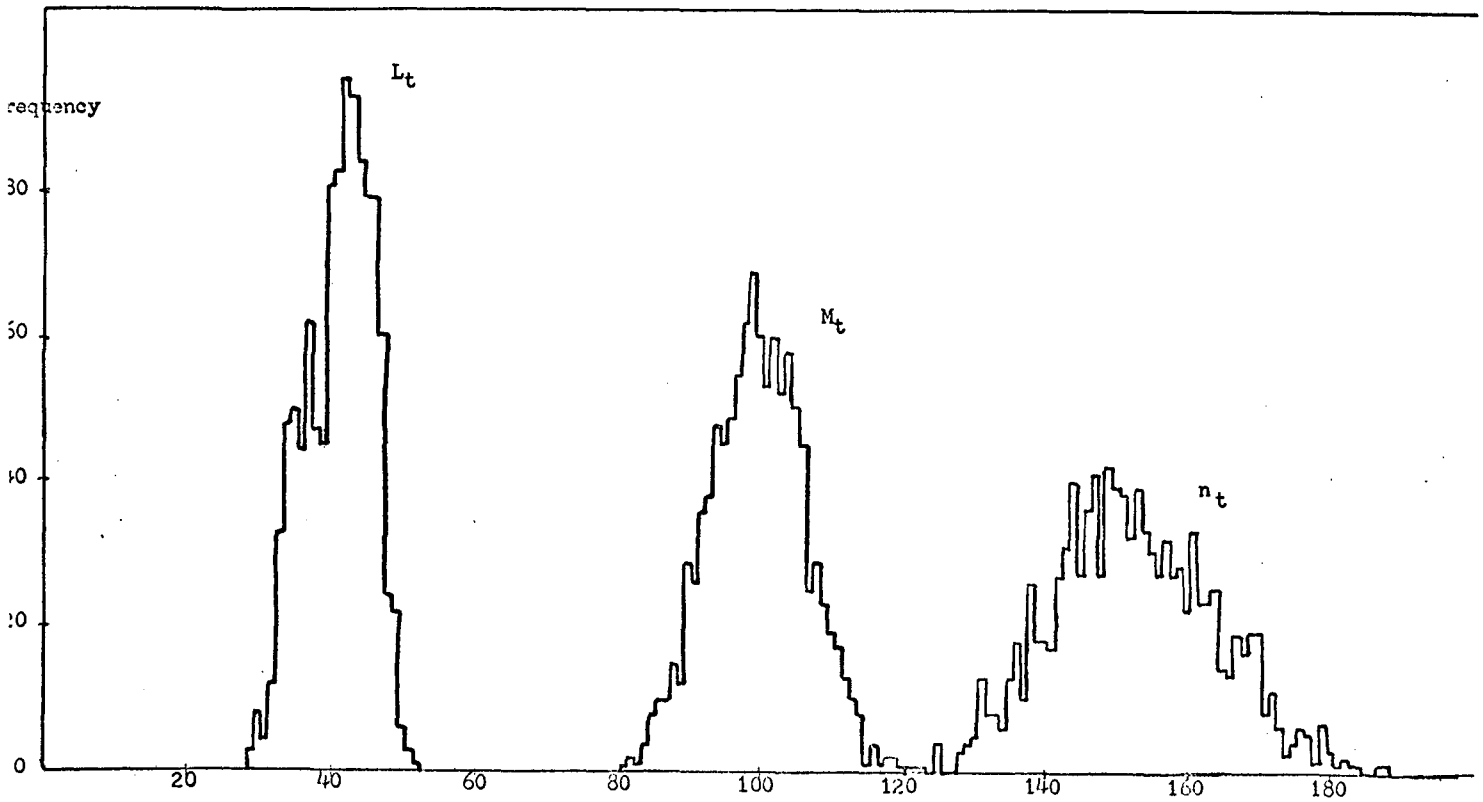


Fig. 9. Approach to equilibrium of word-meaning process.

Fig. 10 . Equilibrium distribution of word-meaning process characteristics (sample size approx. 1000).



This was carried out for eight of the terminal word-meaning systems of our simulation and the g(1) were then ordered to give H(1). Plotting these (Fig. 11), it is clear that a Zipf's law can be stated which holds for the majority of the words in the system, excepting the first few and the last few. The "tailing off" effect can perhaps be ascribed to the homogeneity of the Brownian process - any word, whose total probability fluctuates close to zero, is very likely to hit zero and be absorbed. By introducing an inhomogeneous diffusion, where the variance of the displacement of p(m,1) after time Δt is an increasing function of $p_t(m,1)$, this effect could be removed, and the total number of words and meanings could increase as well.

One interesting comparison can be made between the g(1) vs. H(1) curves for the initial and the terminal states of the first history (see Fig. 8). In the initial, non-equilibrium state all words have equal probability g(1) =.05 . The terminal state has shifted to a typical Zipf's law.

Lexicostatistics

Finally, we present the results of the lexicostatistic survey of the 11 equilibrium system histories. These are displayed in Fig. 12 and the mean behaviour is extracted and is displayed in Fig. 13 . These diagrams speak for themselves - after an initial sharp drop, the index

$$\frac{1}{|M_0|} \sum_{m \in M_0} (1 - \min_{n \in M_t} d_{0,t}(m,n))$$

undergoes an unmistakeably exponential decline.

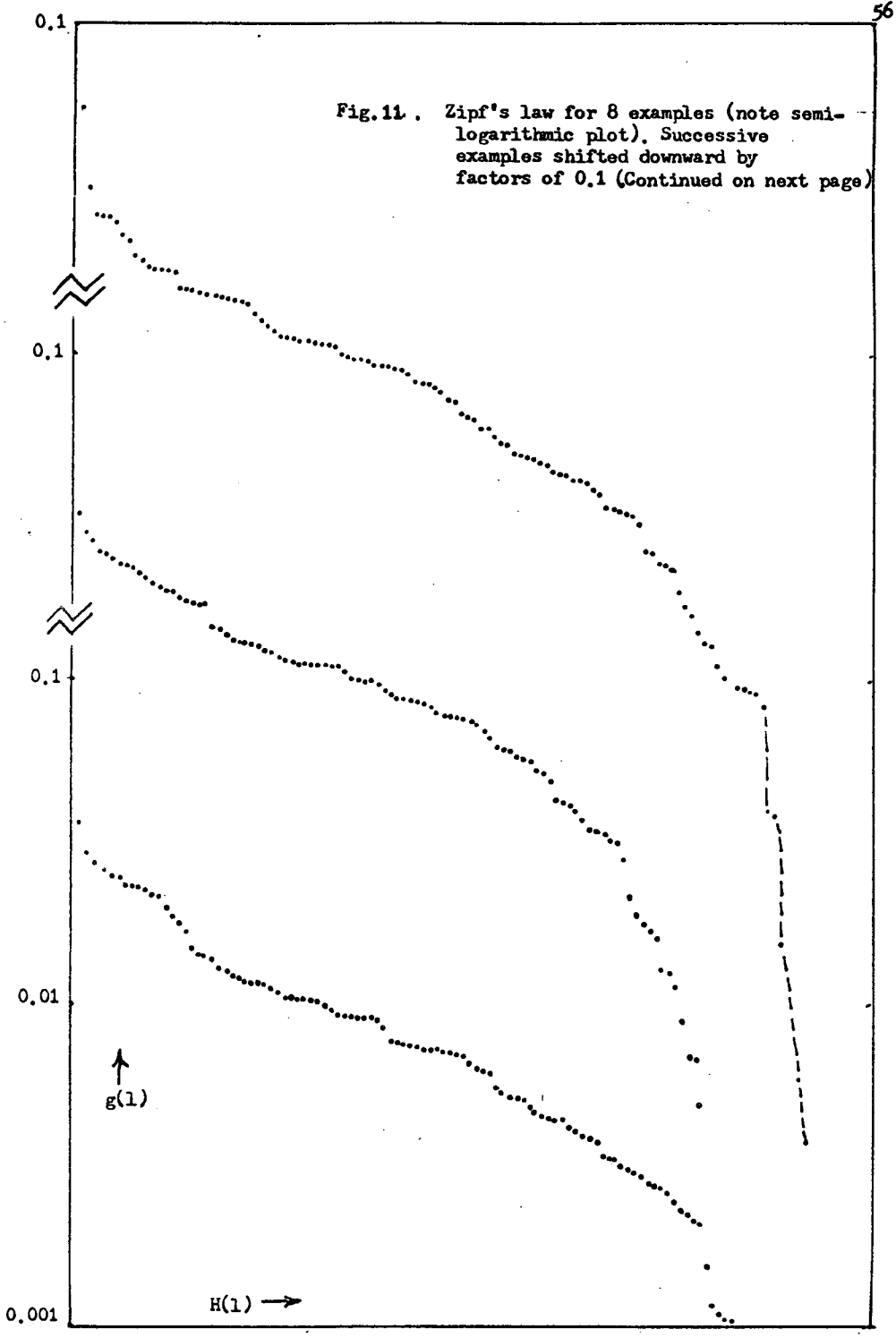


Fig. 11. Zipf's law for 8 examples (note semi-logarithmic plot). Successive examples shifted downward by factors of 0.1 (Continued on next page)

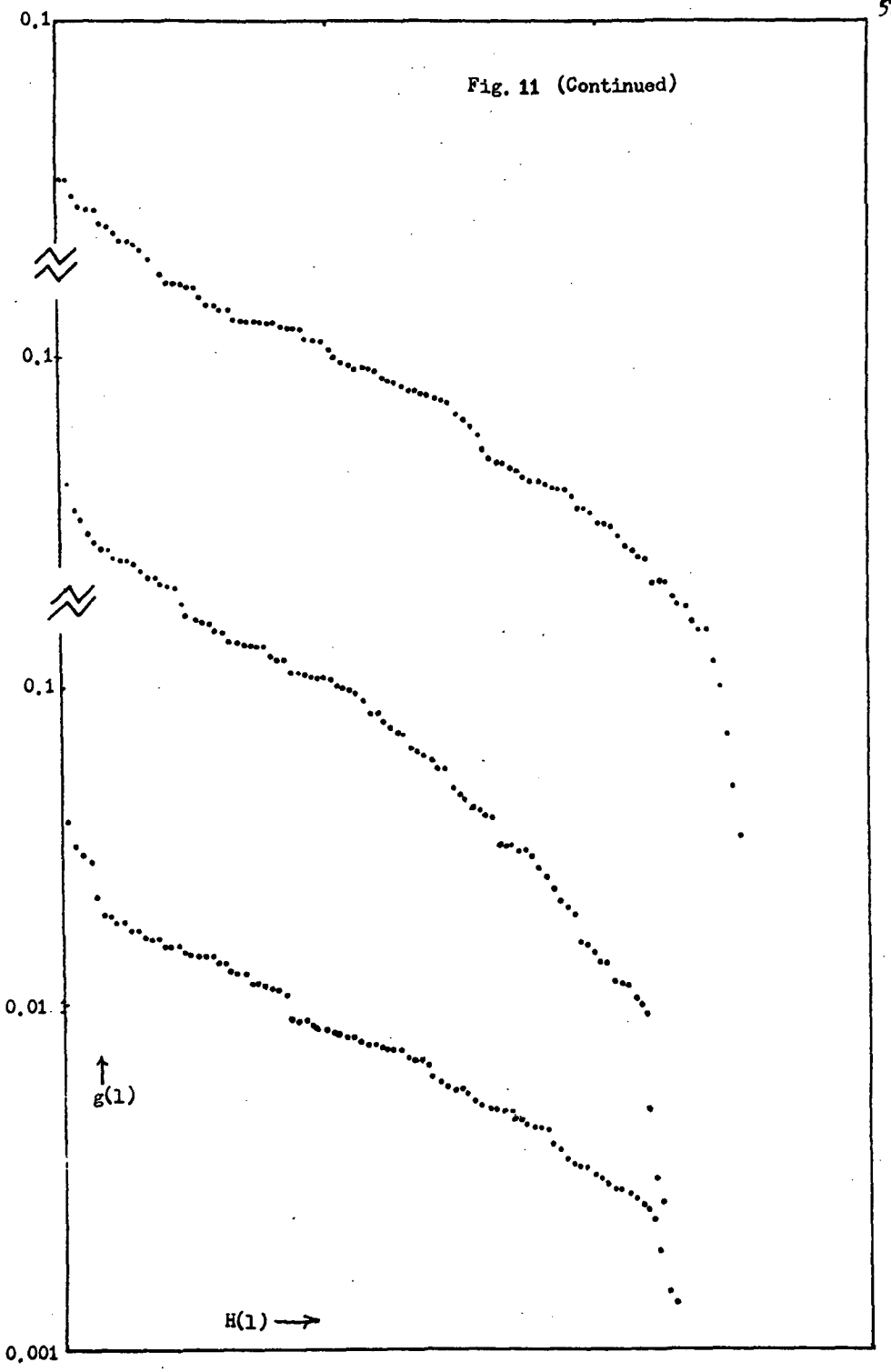


Fig. 11 (Continued)

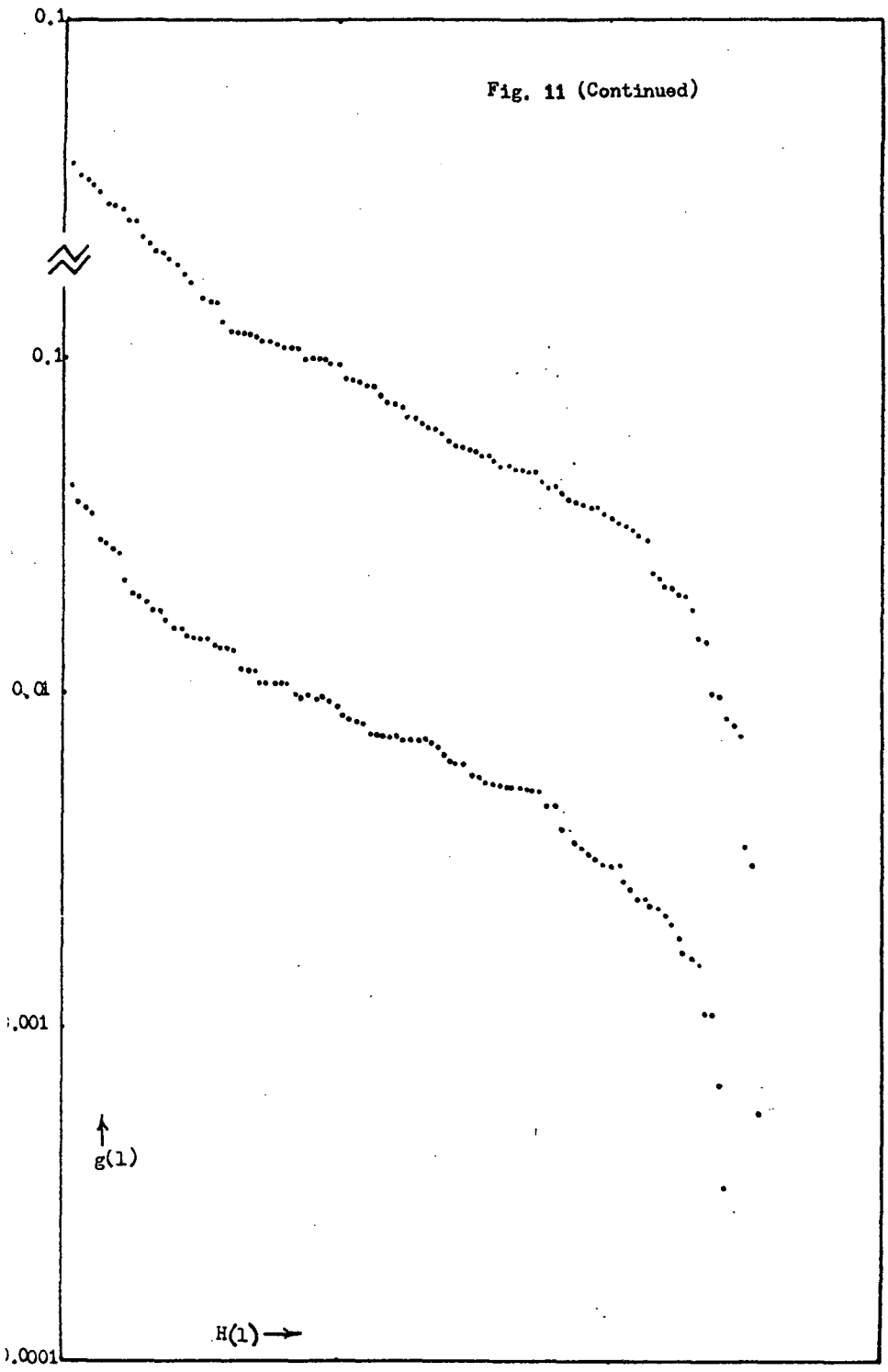


Fig. 11 (Continued)

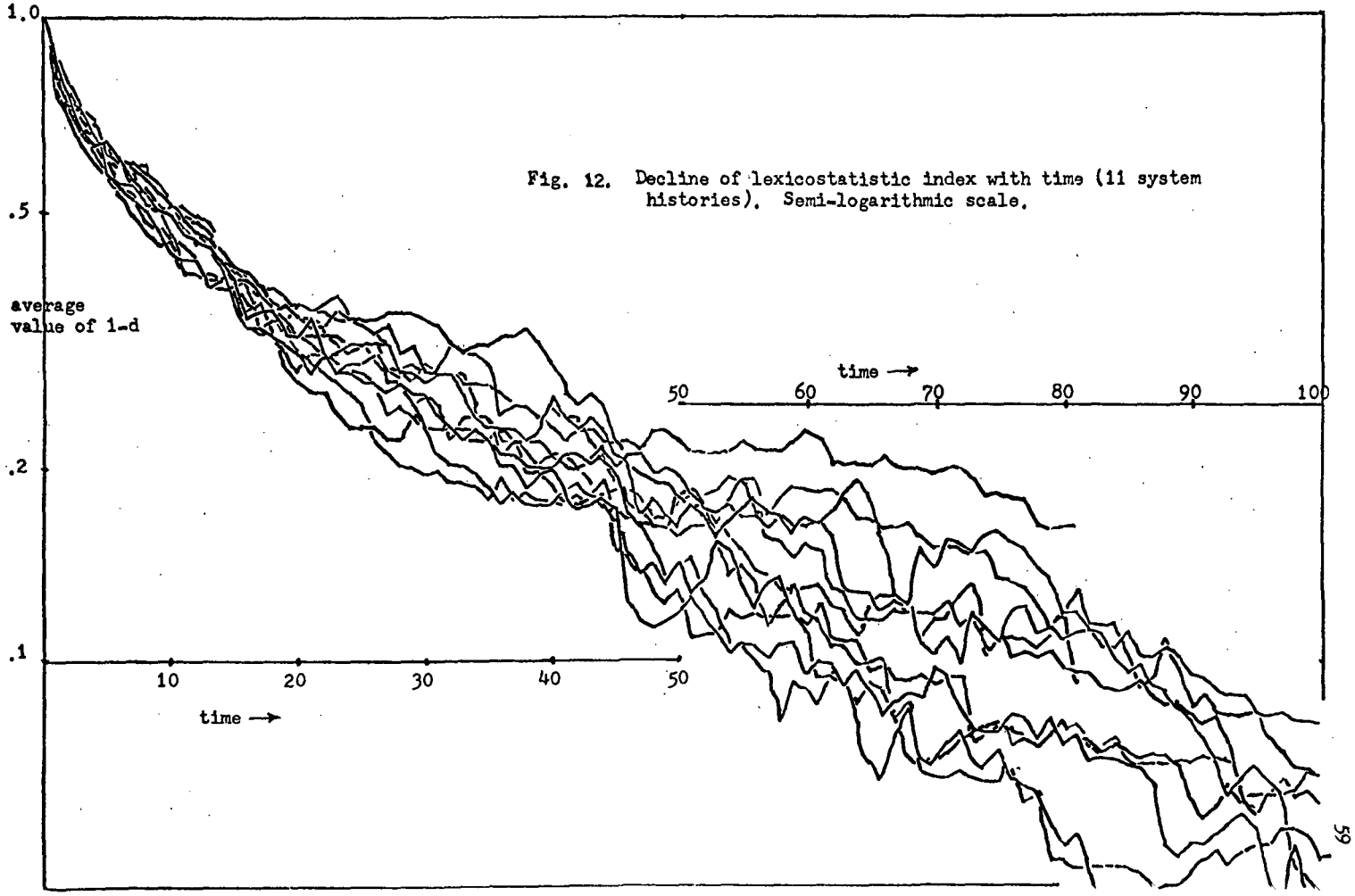


Fig. 12. Decline of lexicostatistic index with time (11 system histories). Semi-logarithmic scale.

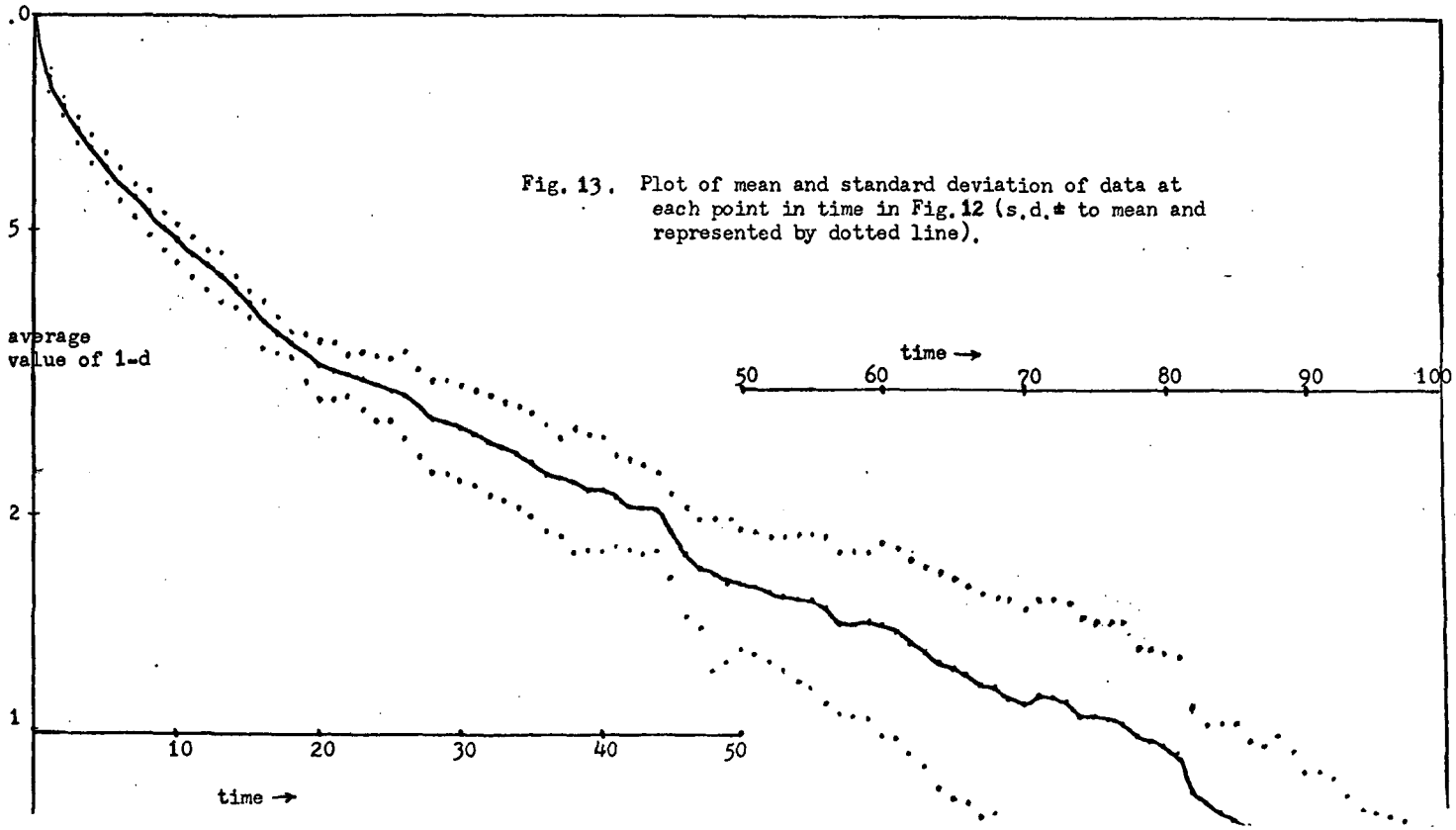


Fig. 13. Plot of mean and standard deviation of data at each point in time in Fig. 12 (s.d.* to mean and represented by dotted line).

To what extent the initial drop is a property of the particular metric being used and to what extent it is an inevitable consequence of the Brownian motion, must await further study. In any case, it does not seem to be a simple consequence of a Zipf's law distribution of word probabilities or the analogous effect for meaning, since it also occurs for very symmetrical initial systems such as the one in Fig. 8.

Without coming to any specific conclusions, it is appropriate to end this chapter by pointing out that both Swadesh's relatively simple model of lexical loss, using a universal meaning set to compare language stages; and our more complicated model, in which comparisons between stages of languages are made in terms of internal properties of the lexicon; concur in the very similar behaviour of their lexicostatistic indexes.

Bibliography

Andreyev, N.D.

1962 "Comment" on Bergsland and Vogt, Current Anthropology 3:130.

Bergsland, Knut & Hans Vogt

1962 "On the validity of glottochronology", Current Anthropology 3: 115-153.

Bloomfield, Leonard

1933 Language. New York.

Brainerd, B.

n.d. "A stochastic process related to language change"

Chretien, C.D.

1962 "The mathematical models of glottochronology", Language 38: 11-37.

Cohen, David

1964 "Problèmes de lexicostatistique sud-sémitique", Proceedings of the Ninth International Congress of Linguists, H. Lunt, ed., The Hague, pp.490-496.

Dyen, Isidore

1960 "Comment" on Hymes, Current Anthropology 1: 34-39.

1964 "On the validity of comparative lexicostatistics", Proceedings of the Ninth International Congress of Linguists, H. Lunt, ed., The Hague, pp.238-252.

Dyen, I., James, A.T., & J.W.L. Cole

1967 "Language divergence and estimated word retention rate", Language 43: 150-171.

Eaton, Helen S.

1940 Semantic frequency list for English, French, German and Spanish. Chicago, University of Chicago Press.

Ellegard, A.

1962 "Comment" on Bergsland & Vogt, Current Anthropology 3:130-131.

Fairbanks, G.H.

1955 "A note on glottochronology", International Journal of American Linguistics 21: 116-120.

Fodor, Istvan

1962 "Comment" on Bergsland & Vogt, Current Anthropology 3:132-134.

1965 The rate of linguistic change: limits of the application of mathematical methods in linguistics. University of Budapest.

Gleason, H.A., Jr.

1960 "Comment" on Hymes, Current Anthropology 1:20.

Gudschinsky, S.C.

1956 "The ABC's of lexicostatistics", Word 12: 175-210.

1960 "Comment" on Hymes, Current Anthropology 1: 39-40.

Hattori, S.

1953 "On the method of glottochronology and the time-depth of proto-Japanese", Journal of the Linguistic Society of Japan, no's.22,23, pp.29-77 (English summary pp.74-77).

1957 Kiso goi chosahyo (A test list of basic vocabulary).

Hirsch, David I.

1954 "Glottochronology and Eskimo and Eskimo-Aleut prehistory", American Anthropologist 56: 825-838.

Hockett, C.F.

1958 A course in modern linguistics. New York, Macmillan.

Hoijer, Harry

1956 "Lexicostatistics: a critique", Language 32: 49-60.

Hymes, D.H.

1960 "Lexicostatistics so far", Current Anthropology 1: 3-43.

Josselson, H.

1953 The Russian word count.

Juilland, Alphonse, & E. Chang-Rodriguez

1965a Frequency dictionary of Spanish words. Mouton, The Hague.

Juilland, A., P.M.H. Edwards & I. Juilland

1965b Frequency dictionary of Rumanian words. Mouton, The Hague.

Katz, J.J. & P.M. Postal

1964 An integrated theory of linguistic descriptions. MIT, Cambridge.

Labov, W.

1967 "Contraction, deletion and inherent variability of the English copula", paper given before the Linguistic Society of America, Chicago, December,1967.

1968 "Consonant cluster simplification and the reading of the '-ed' suffix", unpublished manuscript, Columbia University.

Lees, Robert B.

1953 "The basis of glottochronology", Language 29: 113-127.

Levin, Saul

1964 "The fallacy of a universal list of basic vocabulary", Proceedings of the Ninth International Congress of Linguistics, H. Lunt, ed., pp.232-236. Mouton, The Hague.

Lunt, H.

1964 "Comment" on Dyen, Proceedings of the Ninth International Congress of Linguistics, pp.247-252.

O'Grady, G.N.

1960 "Comment" on Hymes, Current Anthropology 1: 338-339.

Osgood, C.E., G.J. Suci & P.H. Tannenbaum

1957 The measurement of meaning. Urbana.

Parzen, Emanuel

1960 Modern probability theory and its applications. Wiley, New York.

Sankoff, D.

1969 Historical linguistics as stochastic process. Unpublished Ph.D. thesis. McGill University.

Satterthwaite, A.C.

1960 "Rate of morphemic decay in Meccan Arabic", International Journal of American Linguistics 26.

Swadesh, Morris

1950 "Salish internal relationships", International Journal of American Linguistics 16: 157-167.

1952 "Lexico-statistic dating of prehistoric ethnic contacts", Proceedings of the American Philosophical Society 96: 452-463.

Swadesh, Morris

1955 "Towards greater accuracy in lexicostatistic dating", International Journal of American Linguistics 21:121-137.

1962 "Comment" on Bergsland & Vogt, Current Anthropology 3:143-145.

Teeter, Karl V.

1963 "Lexicostatistics and genetic relationship", Language 39: 638-648.

Trager, G.L.

1966 "Comment" on van der Merwe, Current Anthropology 7: 497-498.

Ullman, Stephen

1957 The principles of semantics. Barnes & Noble, New York.

van der Merwe, N.J.

1966 "New mathematics for glottochronology", Current Anthropology 7: 485-500.

Zipf, G.K.

1945 "The meaning-frequency relationship of words", Journal of general psychology 33: 251-256.