Proc. 5th int. Congr. phon. Sci., Münster 1964, pp. 61-81
(S. Karger, Basel/New York 1965).

Suitable and Unsuitable Mathematical Models in Language Statistics, and their Consequences

By Gustav Herdan, Bristol

Originally the word 'mathematical model' meant a three-dimensional geometric structure in wood or cardboard representing visually the relation between three variables; later, it came to be used in the sense of hypothesis or theory, by which to explain an observed relationship between variables. More precisely, and more ambitiously, it is the name for the differential equation set up on the basis of an hypothesis about the behaviour of two or more variables, which by integration would lead to the empirically established relation between those variables. There is nothing to be said against the use of the term 'model', provided one is clear about the sense in which it is to be understood. But from the explanation given above, it is difficult to see what advantage it is to call an hypothesis, or a differential equation, or a theory, a 'model'. In the original sense, a model was very effective in making us visualize an observed mathematical relationship; in its metaphorical sense, it does nothing of the kind, but may even obscure the matter if it makes us forget that 'model' stands here for hypothesis or theory. If so, it tends to make us construct such models independently of one another, and regardless of the wider implications of the observed relation. This could be a great disadvantage, considering that it is the virtue of a theory to reveal hitherto unrecognized relationships between often widely different parts of a field of knowledge. If, therefore, in this address, I use the term 'model', it should be understood according to the above interpretation as hypothesis or, better, theory.

I.

1. For mathematical models to be of real value it is necessary that (1) the relation of events of which the mathematical structure

Table V

Macauley: (1) the vocabulary "expected" on the basis of the *Bunyan* distributions, table II, cols. 2 and 3, and (2) the observed distribution from table II, cols. 4 and 5 (summed).

1 Initial letter of noun	2 Expected vocabulary	3 Observed vocabulary	1 Initial letter letter of noun	2 Expected vocabulary	3 Observed vocabulary
A	207.1	245	O	69.0	70
B	169.8	156	P	339.6	329
C	370.0	377	Q	13.3	14
D	256.2	235	R	227.4	188
E	118.8	162	S	358.2	372
F	155.6	145	T	164.4	172
G	95.6	100	U	15.8	21
H	129.9	107	V	84.1	63
I	139.3	169	W	97.7	86
J	40.9	23	X	–	–
K	17.1	19	Y	4.8	8
L	108.8	117	Z	2.1	4
M	197.6	205			
N	53.8	50	Total	3436.9	3437

difference between the *Bunyan* and *Macaulay* distribution of vocabulary simply in terms of the respective proportions of OE and LR words. By simply varying the proportions in the *Bunyan* vocabulary, we can account for $r^2 = 0.9727$, or roughly for 97 % of the variance of the *Macaulay* distribution.

Regarding the general linguistic aspect of language contact, we have learnt from the above analysis that the relative distributions of nouns in literary texts of English according to the initial letter for each, the Teutonic and the Romance component of vocabulary, can be regarded as random samples of the corresponding probability distribution in the language. As random samples, they are independent of text length and content; moreover the distribution of each component, OE and LR, is independent of, and without influence upon, the other. It is these findings which provide the theoretical basis for drawing the inference of different proportions of the OE and LR components in the vocabularies of two English writers from an observed difference in the alphabetic distribution of their vocabularies.

2a. An obvious and interesting question is whether the stability of the alphabetic distribution of nouns of a given component, say the LR component, is the manifestation of the form of that type of distribution in the parent language, Latin in our case. If this could be established it would throw quite unexpected light upon the *mechanism of 'borrowing'* and the *mixing* of languages. Unexpected, because the current view in this matter among linguists is that change and not stability is the effect of such language contact.

Table VI gives the answer to this question. Col. 2 gives the mean *Bunyan-Macaulay* ranking of the alphabet for nouns of LR origin: col. 3 gives the ranking for Mediaeval Latin nouns (2454 altogether) from the 'De Imitatio Christi', together with samples from miscellaneous works by *Thomas à Kempis* and theological writings of *Gerson*. There is evidently good general agreement between the two, in spite of some differences in rank for certain letters. As a summarizing measure for the agreement, we use again the rank correlation coefficient which for our two series results as 0.960, which means a highly significant correlation. In fact, it is not appreciably different from the ranking correlation coefficient between the two LR series from *Bunyan* and *Macaulay*, which results as 0.986.

Table VI

Comparison of the ranking of initials (1) in nouns of Latin-Romance origin in *Bunyan* and *Macaulay*, col. 2; (2) in Latin nouns occurring in samples from the Imitation, *à Kempis* and *Gerson*, col. 3.

1 Rank	2 *Bunyan* and *Macaulay*, RL component	3 *à Kempis* and *Gerson*, Mediaeval Latin	1 Rank	2 *Bunyan* and *Macaulay*, RL component	3 *à Kempis* and *Gerson*, Mediaeval Latin
1	C	C	14	G	O
2	P	P	15	L	N
3	S	S	16	O	H
4	A	A	17	H	G
5	D	I	18	N	B
6	R	D	19	J	J
7	M	M	20	U	U
8	I	F	21	Q	Q
9	E	R	22	W	Z
10	T	T	23	Z	K
11	F	L	24	K	W
12	V	V	25	X	X
13	B	E	26	Y	Y

The high correlation between the alphabetical distribution of the LR component in both writers, and the corresponding distribution from works written in Mediaeval Latin provides the explanation for the stability discussed above. It is the *tenacity of the functional burdening* of particular sounds (represented by letters in our illustration) when used as noun initials which has made the same probability distribution persist from mediaeval times through the subsequent development of English.

A new and important item of knowledge we derive from the investigation in this respect is that the borrowing of Mediaeval Latin by Old English and the initimate mixing of the two main components of English has left the original alphabetical distribution of nouns in Mediaeval Latin unaltered. It follows that *such alphabetical distributions are to a very high and significant extent independent from one another, in spite of the intimate mixing of the components.*

2b. A point of methodological interest to which I should like to draw attention is that the investigation shows how what started as research of an apparently literary nature only, namely as stylo-statistics, could lead to *highly relevant linguistic results.* This, I submit, is one of the most valuable concomitants of literary or stylo-statistics, to use its conventional name. That this feature should have escaped the attention of *W. Plath*, the author of the chapter on 'Mathematical Linguistics' in "Trends in European and American Linguistics 1930 to 1960" (Utrecht 1961) is most deplorable and must be regarded as a severe fault of his presentation of the subject, in particular where it deals with 'Statistical Linguistics'. According to *Plath*, there has been no important development of the statistical study of language since *Yule*, apart from my new derivation and interpretation of the predominant parameter of vocabulary distribution, *Yule's* Characteristic K, as the coefficient of variation of the mean.

Mr. *Plath* has here missed the essential point that whereas *Yule* was only concerned with the statistical study of vocabulary for purposes of characterizing *individual* style in an objective manner, I have been mainly concerned with extending the analysis of word frequency, and of frequency of linguistic forms in general (phonemes, letters, morphemes, syllables, etc.) to language as such, which led to *statistical linguistics* as the quantitative interpretation of *de Saussure's* 'langue-parole' dichotomy, and with it to a new branch of linguistics as a science in the sense in which *L. Bloomfield* uses the term. That this point should have been missed so completely in a publication which is primarily meant to acquaint linguists with the development of this branch of knowledge is highly regrettable.

3. The question now arises whether the calculus of statistical conditions in language in the form of a multinomial law also applies on the vocabulary level. It is quite conceivable that this formula being only part of a general calculus of linguistic observations, may not be the appropriate one on the higher levels of language.

As it turns out, the difference between the comparatively small number of phonemes in a language (say between 20 and 50) and the very large number of vocabulary items (of the order of magnitude of 50,000) is such as to make it practically impossible to apply the multinomial law if the variable is the occurrence probability of the individual vocabulary items. This would mean a series of as many p_i as there are vocabulary items in a comprehensive dictionary of the language.

However, the multinomial law becomes applicable if instead we let the p_i denote not the probabilities of particular vocabulary items – against which there might also be objections of a theoretical nature –, but the probability of a vocabulary item belonging to one part of the text, to two parts, . . . to all parts into which the original text, or complex of texts, has been equally divided. This means that instead of with vocabulary occurrence frequency, we work with what is known as *vocabulary partitioning*, or *vocabulary connectivity*, a most useful and important characteristic of the vocabulary structure in a language.

In place of the p_i in the statistical universe of phonemes, we have here the probabilities of vocabulary connectivity according to the Random Partitioning Function (R.P.F.)*. Their series serves as a yardstick against which to judge an observed vocabulary connectivity with a view to arriving at a decision whether the differences between theory and observation are compatible with regarding the observed series as a random sample of the universe, represented by the theoretical series*.

3a. To examine, the vocabulary connectivity in say four samples from *Macaulay's* Essay on *Bacon*, we construct by the random partitioning function the chance model of vocabulary connectivity as consisting of all possible combinations in the group, in order to compare the observed connectivity in the members with what would obtain by chance. This enables us also to decide whether a particular member differed significantly from the chance model, and hereby, from the

* *G. Herdan:* Type-Token Mathematics, esp. chapter 3, section 6.3 and chapter XVIII A (Mouton & Co., The Hague 1960).

rest, and if so, in what points, i.e. with regard to which particular classes of combination. A significant χ^2 would mean that the member in question was significantly different from the other members – all or some – in the group, as regards vocabulary and/or occurrence frequency.

The calculation of the required probabilities from first principles soon becomes very troublesome. I have therefore prepared a table from which they can be read off directly for combinations up to the order 30, and occurrence frequencies up to 100 (Type-Token Mathematics, pp. 341–412).

The four cell frequencies for the chance model resulted as 277.51, 61.60, 71.58, 273.66: total 684.35. For the χ^2-test they are adjusted to the observed total 681.50, as in the second row of table VII. This provides our yardstick of chance with which we compare the observed class frequencies from the essay on *Bacon*, as shown in the first row of the following table.

Table VII

Vocabulary connectivity (averaged) in four samples from essay on *Bacon**.

	(ABCD)	(ABCδ)	(AB$\gamma\delta$)	(A$\beta\gamma\delta$)	Total
Observed	271.00	59.25	67.50	283.75	681.50
Calculated	276.40	61.35	71.29	272.57	681.61

$\chi^2 = 0.839$

Since χ^2 remains far below the value required for significance (7.815) for the given degrees of freedom, we conclude that the vocabulary connectivity between one sample (averaged) and the rest is, by and large, that expected on random partitioning within the same universe.

In a comparison 7 samples from works by *Bunyan*, the calculated cell frequencies were 449.17, 75.64, 76.02, 259.53: total 860.36. For the χ^2-test they were proportionally adjusted to the observed total 718.58, as shown in the second row of the table VIII.

Contrary to the previous comparison, the value of χ^2 substantially exceeds that required for significance, and we conclude that the vocabulary connectivity between the samples (averaged) and the rest differs significantly from what is expected by random partitioning. The reason is evidently that we are here dealing with samples from three different works by *Bunyan*, which statistically must also be regarded as different universes.

* The presence of a vocabulary item in the four samples is denoted by the capital Latin letters A, B, C, D, and their absence by the Greek letters α, β, γ, δ.

Table VIII

Vocabulary connectivity (averaged) in four works by *Bunyan*.

	(ABCD)	(ABCδ)	(AB$\gamma\delta$)	(A$\beta\gamma\delta$)	Total
Observed	259.00	72.75	74.33	312.50	718.58
Calculated	375.15	63.18	63.49	216.76	718.76

$\chi^2 = 81.527$

4. As if to disprove the view held by some linguists that literary statistics could not contribute to our knowledge of language, or – as the magic formula goes – were "linguistically not relevant", the Italian Semiticist, *P. Franzaroli*, has adapted the method of vocabulary connectivity to provide a method for the investigation of philological or, as we say, linguistic phenomena. His problem was the classification of six semitic languages: Babylonian, Ugaritic, Hebrew, Syrian, Arabic and Ge'ez (abbreviated: Ba, Ug, Heb, Sy, Ar, and Ge).

For the investigation he selected certain phonological and morphological characteristics, altogether 217 isoglosses, as a representative sample of such features, and recorded for each of the six languages whether a particular characteristic was present or not. The following is a sample from the fundamental table listing the results of the investigation.

Table IX

	Ba	Ug	Heb	Sy	Ar	Ge
1. *p conservato*	+	+	+	+	–	–
2. *p > f*	–	–	–	–	+	+
3. *ṯ conservato*	–	+	–	–	+	–
4. *ṯ > s (s in Ge'ez)*	+	–	+	–	–	+
5. *ṯ > t*	–	–	–	+	–	–

He then chose the characteristic of vocabulary connectivity – which here becomes that of connectivity in phonological and morphological features, though, in order to avoid the longer form, we shall continue to use the term vocabulary connectivity – and recorded the numbers of characteristic features which each language had in common with 1, 2, ... 5 other languages (table X).

Table X

Languages	1	2	3	4	5	Total
Ba	33	11	18	15	8	85
Ug	3	18	17	17	18	73
Heb	9	22	22	20	16	89
Sy	12	14	16	21	17	80
Ar	12	28	14	20	16	90
Ge	14	21	12	11	15	73
Average	13.8	19.0	16.5	17.3	15.0	

For comparing one language with the rest, we require again a chance model of connectivity. We can, however, not directly use the numerical table of the random partitioning function because the basic data are here given in a different form from how they were provided in the illustration under 3. There the basic information was the frequency distribution of vocabulary according to the occurrence number of the vocabulary items. These numbers could vary from 1 to anything. Here every one of the 217 phonological or morphological items can only be present or absent, and in symbols have thus the occurrence numbers 1 or 0, in each language. However, the chance model can here be obtained directly from the column sums of table X by dividing the column sums by the number of rows, that is, as the average vocabulary connectivity.

Illustration

For the purpose of comparison, the theoretical figures are reduced to the observed totals, e.g. for Babylonian:

Table XI

	1	2	3	4	5	Total
Ba	33	11	18	15	8	85
Chance model*	14.3	19.8	17.2	18.0	15.6	84.9

$\chi^2 = 32.59$

* Last row of table X adjusted to a total of 85.

which for the given degrees of freedom means that a χ^2 like this or greater could on pure chance occur less often than 1 in 100. We, therefore, conclude that Babylonian is exceptional in its vocabulary connectivity with the other languages.

* * *

For the sake of completeness, and to avoid confusion, it should be remembered that just as our model is not the whole of the Calculus of Statistical Conditions in language, so that Calculus is not the whole of what I have called the Calculus of Linguistic Observations, which in addition to statistics comprises Combinatorics and certain branches of Geometry.

References

Franzaroli, P.: Prospettive di Metodo Statistico nello classificazione delle lingue Semitiche. Atti della Accademia dei Lincei, Serie Ottava., XVI, fasc. 1–12, Roma 1961.

Herdan, G.: Type-token mathematics (Mouton & Co., 's-Gravenhage 1960).

Herdan, G.: The calculus of linguistic observations (Mouton & Co., 's-Gravenhage 1962).

Mandelbrot, B.: An informational theory of the statistical structure of language. *W. Jackson*, Communication Theory (Butterworth, London 1953).

Plath, W.: Mathematical linguistics. In Trends in European and American Linguistics 1930–1960 (Spectrum Publ., Utrecht 1961).

Yule, G. U.: The statistical study of literary vocabulary (Cambridge University Press, Cambridge 1944).

Zipf, G. K.: Selected studies of the principle of relative frequency in language (Harvard University Press, 1932).

Zwirner, E. und *Zwirner, K.:* Die Häufigkeit von Buchstaben- und Lautkombinationen. F. u. F. *12:* 286–287 (1936).

Autor's address: Gustav Herdan, University of Bristol, *Bristol* (Great Britain).