LINGUISTICS AND MATHEMATICS: MIX WITH CARE

Robert Abernathy

University of Colorado

Views about what the relationship between mathematics and linguistics should be range all the way from that which holds the latter to be an in some sense "non-mathematical" discipline, to that which would make of linguistics essentially no more than a branch of mathematics. Both extremes represent, I think, misconceptions, resting in the one case on an indefensibly narrow notion of the nature and scope of mathematics--say the popular idea of this as beginning and ending with real-number arithmetic--and in the other case on a likewise unjustifiable restriction of the linguist's domain of interest: an empirical science does not, save perhaps in a context of sheer Pythagorean mysticism, reduce to its mathematical models.

Aside from the hopefully safe generality of rejecting such rash generalities, there seems to be little of a purely theoretical character to be said about the title subject (little, at any rate, which would not simply belong to the philosophy of science in general, without being specific to the problem of linguistics' situation vis-à-vis mathematics). It is surely too early for apodictic pronouncements on a topic which future developments, in either or both of the fields concerned or in such related areas as laboratory phonetics and computer technology, may still cause

to assume new and perhaps surprising aspects.

At the same time, there is already no dearth of published essays in the application of mathematical methods to linguistic problems, and at this stage the critical analysis of explicit proposals of limited scope--whether successful or not--may be of more value than attempting a global survey of an inchoate subject. The example which I mean to consider in some detail here is that presented in the self-contained eighth chapter of Gustav Herdan's Type-token mathematics ('s-Gravenhage 1960, this chapter earlier in Language and speech, 1958), under the heading "The relation between the functional burdening of phonemes and their frequency of occurrence." This is, be it noted at once, one of the unsuccessful attempts, which does not mean it is without interest--indeed, there is often much to be learned by examining a construction to see just how it has gone wrong, and the present one has certain positive merits: it contains in compact form the essential ingredients of statement of an empirical problem and development and testing of a mathematical model, and both the facts and the suggested model are reasonably simple and straightforward. Not so much can be said for the presentation, which seems calculated to confuse the reader, particularly if he is one of those who find mathematical formulas as such heavy going. But the book is (or so one seems to gather from the author's introduction) at least partly intended for such a reader, and that with reason. The former of the two extreme views which I began above by discounting no doubt contains its grain of truth in that, whether or

not linguistics or any science can be truly non-mathematical, it is still the case that many linguists are "non-mathematical". By this I mean that we tend to be the sort of people who, confronted by a text containing alternate sections of discursive explanation in ordinary language and of mathematical exposition replete with symbols and algebraic expressions, are inclined to skim or skip over the latter and study the former the more closely--supposing, of course, that the subject matter is of some intrinsic interest--in the not always unjustified hope of being able to follow the general idea while taking the specifically mathematical developments more or less on faith.

Let us here, as a heuristic device, first go over the cited chapter from the standpoint of such a "non-mathematical reader", then undertake to examine its formal part in order to locate the sources of the perplexity to which this first reading must almost surely give rise.

Herdan begins by identifying the problem as one first set forth by Trubetzkoy, who observed in his Grundzüge that linguistic "statistics" are often of two quite disparate kinds at the basic data-gathering and processing level: numbers of instances of items of interest (say phonemes) are sometimes found from sample texts in a language, sometimes from lists such as the usual word-dictionaries. Trubetzkoy prudently stressed the importance of not confusing these two kinds of "frequency"; indeed, in practice the distinction is sometimes ignored or at least not made clear in contexts of informal discussion. Herdan, however,

sees Trubetzkoy's caveat as reflecting primarily just an unsatisfactory situation in linguistic theory, in that "the relation between the two distributions of relative frequencies has so far not been reliably established" (p. 127), and he proposes to remedy this.

Some empirical data of the two kinds are adduced (from a list count reported by Kramský and a text count by Fowler). In each case, the facts to be used for theory testing in the sequel are aggregate frequencies for classes of English consonants, grouped by their manner of articulation (labial, dental, palatal, velar): the corresponding relative frequencies from the two counts mentioned are, respectively, .244, .598, .015, .143 and .235, .649, .010, .106. Clearly, as Herdan points out, these figures taken pairwise exhibit what looks like a more than accidental similarity. This he sees as suggesting that the text frequencies (the second four numbers above) are, in some fashion yet to be determined, dependent on the list frequencies (the first four numbers), and that it should be possible to express this dependence mathematically, "to derive the frequencies of phonemes in speech by a stochastic process from the pattern of functional burdening in the dictionary," as he puts it (pp. 129f). More precisely (p. 130):

> "The stability of the distribution of frequency of use, its independence of the type of text, points to a general function according to which phonemes are produced in speech. Insofar as the phoneme distribution from texts is similar to that from the dictionary, it might well be considered to be a random sample of the latter, and a law of chance might present such a

> general function. We shall therefore use the Poisson law of rare events, which is most likely to fit the data in question if they are governed at all by chance, and calculate the numbers of phonemes belonging to the different categories... If the calculated numbers sensibly agree with these actually present in the sample, and if the proportions are also like those observed, we would conclude that the frequencies of use of phonemes may be regarded as random samples of their functional burdening, and the distributions of the former as random samples of the distribution of the latter."

So far, the non-mathematical reader should have been able to follow the drift of the argument with no particular difficulty. (We must not, after all, imagine him so naive as to boggle at such notions as that of relative frequency--in that case, he would surely not be reading such a text in the first place!--or that of a procedure designed to yield "calculated", "theoretical", or "predicted" frequency figures for eventual comparison with the results of actual counting.) In the present instance, the idea would seem to be to find a formula which, given dictionary frequencies, will "predict" text frequencies. This is a decidedly startling idea: if we had such a formula and it worked, that would mean among other things that a good deal of labor which has gone and still goes into compiling text-frequency counts (of phonemes, say) for various natural languages is work wasted, since it would suffice to analyze the dictionary once and for all and from the results of doing this compute, if desired, text frequencies to be expected, on the average, in texts of any specified length.

Now follows the "math part". This is merely two short paragraphs, presenting the proposed formula with glosses on the symbols used in writing it and with a brief sketch of its derivation. Our

non-mathematical reader will presumably pass over this lightly and go on to the subsequent account of results obtained by using the formula, trusting (or mistrusting, according to temperament) that it will explain itself in use.

These results are presented in a table headed "Calculated frequencies for samples of different sizes," giving (in each case both as absolute and as relative frequencies) two sets of figures, one for "text length" 1000 and the other for "text length" 5000 corresponding in each instance to the four consonant-classes that were recognized in connection with the empirical data. The accompanying discussion calls attention to two aspects of this table for which significance seems to be claimed:

(1) For text lengths 1000 and 5000 respectively the absolute frequencies calculated total 967 and 4191. As Herdan interprets this, it signifies "that the sample occurrence of 1000 consonant phonemes is accounted for to 96%, and that of 5000 to 84% by the theoretically expected number of phonemes... This means that the number of different consonantal phonological oppositions in the text samples are, to that extent, accounted for by those in the dictionary, representing the functional burdening of phonemes. The remaining 4% and 16% are repetitions."

(2) "Moreover, there is a striking agreement between the actually observed relative frequencies in running texts and those calculated from the dictionary probabilities, using a law of chance, in the proportion of phonemes accounted for by the four phoneme groups." It is easily verified by inspection

that this "striking agreement" exists in fact, e.g. for text length 1000 the four relative frequencies found are .236, .643, .010, .111 (cf. the results from Fowler's count: .235, .649, .010, .106). At text length 5000 the resemblance is a little less close, but close enough.

The hypothetical reader will surely be rather at a loss with regard to point (1)--it is not clear here just what is being "accounted for" and in what sense, nor how the "phonological oppositions" got into the picture. To find these things out, one must either go back and work through the mathematical exposition or, failing that, can only agree that, in a statistical frame of reference, 96 or even 84 per cent may represent a pretty good score in "accounting for" things, and go on in search of something more familiar.

With point (2), one does seem to find the continuation of the line of thought embarked upon previously: here, it appears, are the "calculated numbers" which were to confirm or disconfirm a theory according to whether they did or did not agree reasonably well with observed frequencies. The numerical results, obviously, do agree quite closely, and what Herdan goes on to say about this (under "Conclusions" at the end of the chapter) seems designed to reinforce the impression that a strong substantive claim--which, as noted above, would have immediate practical consequences--is being made: "Functional burdening thus appears as the dominant factor in the use of phonemes in speech... The mutual relation between phonemes as regards functional burdening

determines that of the categories of phonemes in speech output." (p. 131). Hence, he says: "Trubetzkoy's 'double relativity' is an illusion in this case: the functional burdening of a phoneme group in use being only a random sample of that in the dictionary." (p. 132).

On closer inspection, however, one may find room to wonder whether any claim as strong as that suggested by the language of both prospectus and conclusions is really made here. Perhaps the relation formulated is only supposed to hold for certain classes of phonemes, not for their individual members? Or perhaps it only applies to particular classes of consonants, as in the exemplary data, or only to these classes of consonants in English? In the last case, it would have no general consequences at all. (Note the reservation "in this case" to the concluding statement, also the again puzzling restriction to data "accounted for" in connection with the comparison of observed and calculated frequencies. It is at least apparent that Herdan is using this frequently slippery expression "accounted for" in an idiosyncratic and unexplained fashion, as is also true of his employment of some other terms--notably "functional burdening", which however seems to explain itself (cf. quotations above) as simply a synonym for "relative frequency" (either in the "dictionary" or the "text" sense).)

The linguistically sophisticated reader will, indeed, readily think of some instances for which the existence of any simple relationship, or any regular relationship at all, be-

tween dictionary and text frequencies of particular phonemes seems on the face of it most unlikely, e.g. the notorious case of English /ŋ/. And he will, if he is wise, conclude that--whatever may have been shown here in some abstruse sense--the formula offered is not something which he can safely adopt and apply in general to empirical data in the expectation of obtaining reliable empirical predictions.

Now let us go back and examine systematically the two paragraphs we skipped. These are as follows:

> For each of the four phoneme groups the number, n, of consonant phonemes to be expected in a text of specified length, N, is calculated according to the Poisson law of rare events as
>
> $$n = L(1 - e^{-Np/L})$$
>
> where L equals the number of the phoneme group in question from Table 39, column 3, and p equals the probability of such phonemes from Table 40, column 5 (average). More precisely, L is the number of occasions on which a phoneme, belonging to a specified category, appears in phonological opposition, disregarding the possibility that in a number of cases the opposition may not be phonologically relevant (Trubetzkoy's "Aufhebungsstellung").
>
> The argument implied in the above formula is as follows: according to the law of rare events, the probability of one of L phonological oppositions not occurring in a sample of one phoneme is e-1/L; the probability of one of L phonological oppositions not occurring in a proportion of Np of a text of length N is e-Np/L, and consequently the probability of its occurrence is 1 - e-NP/L. The probable number of phonological oppositions of the group in a text length N is then L(1-e-Np/L).

(Tables 39 and 40 are the list count and text count data from Kramský and Fowler respectively; the reference to an "average" in connection with the latter is because the table gives also figures for a sequence of subsamples.)

One thing quickly becomes obvious: the formula given cannot be what the accompanying discussion led one to expect, namely an expression for text frequency as a function of dictionary frequency. This one is of the form $f(L, N, p)$, where L is a numerical value obtained in practice from a list count, N and p are ones obtained from a text count (or N is hypothetical, as in the computations which follow, but p is a text-count ratio). I.e., one must put empirical figures of both kinds into the formula before one gets anything out of it, so what it yields cannot possibly be a prediction of results of one kind, given only those of the other kind.

What Herdan's n really is is not quite so clear, but becomes so if one reconstructs the derivation he informally sketches. The line of reasoning here runs via consideration of the events 'particular alphabet element at a particular place in a particular word' (this Herdan calls a "phonological opposition", which is peculiar usage, but there is no point to quibbling over terminology). In a proper dictionary, i.e. one which contains each word in the language exactly once (whereas a text may contain repetitions of the same word) evidently each "opposition", in this sense, will likewise occur exactly once. For a given alphabet element (phoneme) a, let L be the number of "oppositions", i.e. dictionary

occurrences, of a. Now consider any occurrence of a in text: this will be (on the natural assumption that the text consists of words which are also in the dictionary, is a text in the language which the dictionary is a dictionary of) also an occurrence of some particular "opposition" of a. Herdan assumes, apparently on the basis of a classic principle of indifference, that in each instance there is a probability L^{-1} for the given text occurrence of a to be an instance of any particular "opposition" of a. We are, as it were, to think of the text as composed, in part, by a sequence of random drawings (sampling with replacement) from the collection of "oppositions" of a. (N.B. that we are not entitled to reject the model at this point on the grounds that it would be absurd as a model of how utterances, say, are actually produced by a speaker--rejection would be justified if such an interpretation were intended, but in fact no such demand is imposed upon the model, and one cannot legitimately object to its having a counterintuitive "as if" character at this stage. Note, also, that nothing has been said so far about what fraction of the text is made up of occurrences of the element a.)

Let the number of occurrences of a in a given text be M. In terms of the suggested model, these amount to M independent drawings (with replacement) from an urn containing L different objects (the "oppositions" of a). In such a case, the probability that any specific one of the set of objects--i.e., here one and the same "opposition" is drawn exactly k times

($0 \leq k \leq M$) is expressed by the binomial formula:

$$B(k, M, L^{-1}) = \binom{M}{k} L^{-k}(1 - L^{-1})^{M-k} \qquad (1)$$

and, if both M and L are large--conditions which may be conceded for the intended application here--then a good approximation to $B(k, M, L^{-1})$ is given by the Poisson probability:

$$P(k, M, L^{-1}) = (k!)^{-1}(ML^{-1})^k e^{-ML^{-1}} \qquad (2)$$

where e is the universal constant (base of natural logarithms, $e = 2.71828...$).

When $k = 0$ --i.e., when we consider the probability that a given "opposition" fails to turn up at all in M drawings--this probability reduces to

$$P(0, M, L^{-1}) = (0!)^{-1}(ML^{-1})^0 e^{-ML^{-1}}$$
$$= e^{-ML^{-1}} \qquad (3)$$

The complementary probability that a given "opposition" occurs at least once in the course of M drawings is, of course,

$$\sum_{k=1}^{M} P(k, M, L^{-1}) = 1 - e^{-ML^{-1}} \qquad (4)$$

This probability corresponds to the event that the given "opposition" occurs once or twice or ... or M times. But Herdan calls this loosely "the probability of its occurrence," which is dangerously ambiguous.

If we now consider that the M occurrences of a are part of a text made up altogether of N symbols, we can write $p = MN^{-1}$ and $M = Np$, this p being the relative frequency of the symbol a in the text.

Since there were altogether L different"oppositions" of a and for each of them the probability of occurring at least once after M drawings is given by (4) above, the expected number of distinct oppositions which will have turned up at least once apiece after M drawings is

$$n = L(1 - e^{-ML^{-1}}) \tag{5}$$

or, making the substitution above indicated,

$$n = L(1 - e^{-Np/L}) \tag{6}$$

which is of course Herdan's formula. It does not represent "the number, n, of consonant phonemes to be expected" in either of the two possible senses of this ambiguous expression (number of distinct phonemes : number of phoneme-occurrences); rather, it is a number of "oppositions" (in the first sense, number of different oppositions). That it is mislabeled as quoted seems to be a consequence of the author's having lost sight of the distinction he began by drawing between phonemes and oppositions, and also of that between the two senses of "number". This is confirmed by the fact that, on the basis of just this purely verbal equivocation, he goes on to attribute significance to numerical similarities between computed n values and observed frequencies (M values from Fowler's count). (As to why a "striking agreement" is found in this example, cf. below.)

As for testing calculated n values against empirical observations, this cannot be done on the basis of the data adduced, since these data include no information of the required kind.

The reason why the ratios $n_i / \sum_{j=1}^{m} n_j$ of quantities n_i calculated by means of the formula to the total of quantities so calculated so closely resemble the ratios $M_i / \sum_{j=1}^{m} M_j = p_i$ of empirical text frequencies M_i to the total of such frequencies is, of course, basically that, for each i, p_i is used to compute n_i. For the rest, how close this resemblance is depends on the relative size of M_i and the corresponding list frequency L_i. In Herdan's two specimen computations, $M_i = Np_i$ is fixed arbitrarily by the choice of N = 1000 and N = 5000 (involving the standard assumption that an observed relative frequency can be construed as a probability which will hold good for samples of all sizes--as Herdan points out in connection with the empirical data, there seems to be evidence that the assumption is viable for the kind of material involved here).

We have, to be exact,

$$\lim_{L \to \infty} L(1 - e^{-ML^{-1}}) = M \qquad (7)$$

i.e. as the list becomes large in comparison to the text, n tends to become identical with the text frequency M, until the dependence on L vanishes in the limit, and conversely also,

$$\lim_{M \to \infty} L(1 - e^{-ML^{-1}}) = L \qquad (8)$$

at the other extreme, as the text becomes much bigger than the dictionary, n becomes approximately L and dependence on M (or on p) vanishes.

Setting N = 1000 and N = 5000, while $\sum_{i=1}^{m} L_i = 13862$ (from the empirical list count, and with m = 4 in these examples) amounts to choosing the M_i small enough, in both cases, in relation to the corresponding L_i that the behavior of the formula approximates rather closely to that indicated by (7). In fact, since for small x a good approximation to e^x is given by 1 + x, when the ratios M_i/L_i are small it follows that

$$\frac{n_i}{\sum_{j=1}^{m} n_j} = \frac{L_i(1 - e^{-M_i L_i^{-1}})}{\sum_{j=1}^{m} L_j(1 - e^{-M_j L_j^{-1}})} \approx \frac{L_i(1 - (1 - M_i L_i^{-1}))}{\sum_{j=1}^{m} L_j(1 - (1 - M_j L_j^{-1}))} = \frac{M_i}{\sum_{j=1}^{m} M_j} \quad (9)$$

All that is accomplished by comparing the left- and right-hand expressions' numerical values (as Herdan does in his Tables 41 and 42) is to verify by computation a well-known property of the constant e.

It remains to clarify the obscurity of "phonological oppositions...accounted for" or "phonemes accounted for" (cf. above). The percentages quoted here are the ratios $\sum_{i=1}^{m} n_i/N$ times 100. With reference to (8) above, we note that

$$\lim_{N \to \infty} \sum_{i=1}^{m} n_i/N = \lim_{N \to \infty} \sum_{i=1}^{m} L_i/N = 0 \quad (10)$$

i.e., the percentage in question will approach 0 as the sample becomes large, while on the other hand we have

$$\lim_{N \to 0} \sum_{i=1}^{m} n_i/N = \lim_{N \to 0} \sum_{i=1}^{m} p_i e^{-\frac{Np_i}{L_i}} = \sum_{i=1}^{m} p_i = 1 \quad (11)$$

so that the proportion of the sample "accounted for" approaches 100% as the sample size approaches zero. (N.b., however, that in letting $Np_i = M_i$ become very small, we have overstepped one of the conditions which justified the use of the Poisson distribution as an approximation to the binomial (at formula (2) above), and consequently the model cannot be expected to make sense in this case in terms of the intended interpretation).

Recalling that the n_i represent expected number of distinct "oppositions" occurring in a text, while N represents the total number of opposition "occurrences"- in the text, including repetitions, evidently the interpretation of these ratios could be paraphrased by saying that, in a given text, an opposition-occurrence is "accounted for" if and only if it is the first occurrence of the given opposition, and otherwise it is not "accounted for"; or--in line with the author's alternative language--a phoneme-occurrence in a text is "accounted for" just in case it is the first occurrence of some one of its oppositions in the text. Obviously, then, a text of one element is necessarily 100% accounted for, one of 2 elements may be either 100% or 50% so according to whether the second element is or is not a repetition of the first, and so on, until any sufficiently long text will contain a negligible proportion "accounted for". The model behavior above noted is in line with the interpretation except at the lower end of the function's domain (as $N \to 0$), as was to be expected in view of the nature of the approximation it represents. (Contrast e.g. for M = 1 the approximation

$1 > L(1 - e^{-ML^{-1}}) > 1 - \frac{1}{2L}$ with the corresponding exact expression $L(1 - (1 - L^{-1})^{M}) = 1.$)

From the foregoing examination there results, for the problem originally stated, only a non liquet, which should not be mistaken for a negative conclusion, nor yet for a demonstration that the problem is either unsolvable or meaningless. This problem is, in necessarily somewhat vague intuitive formulation: Given certain objects which we are willing to call dictionaries of and texts in natural languages, and considering these just as consisting of sequences of symbols in some alphabet, is it possible to specify any regular relationship between the relative frequencies of symbols in the two instances (for the case, of course, of a dictionary of and a text or texts in one and the same language)?

A direct empirical approach to a problem thus broadly stated would promise little. On the one hand, from the standpoint of mathematical model-construction, this could lead to interminable and inconclusive attempts to fit more or less likely-looking functions, from among the indefinitely many available or constructible, to available or obtainable data about actual languages, subject always to the objection that an apparently valid relationship might fail in application to the next language investigated. And on the other hand, from the standpoint of empirical validation, objections to a proposed general theory on grounds of its failure in application to particular data might always in principle be met by denying that the given dictionary and/or text(s) were properly representative of the language in question--indeed, there can be a

great deal of uncertainty in such matters, cf. e.g. the discussions which have taken place over what sort of entries should figure in a word-dictionary of Chinese.

In such a case, it is much better, or even essential, to adopt an abstract approach, constructing a general model and deducing consequences from it. To this end we first need reasonably precise definitions of the notions which play an essential part in the informal statement of the problem.

Suppose given an alphabet $A = \{a_1, \ldots, a_m\}$, a dictionary $D \subset A^*$, $D = \{W_1, \ldots, W_n\}$, and a text $T \in D^*$. We now define the relative frequency of a_i with respect to D as follows:

$$f_D(a_i) = \frac{\sum_{k=1}^{n} N_{ik}}{\sum_{k=1}^{n} L_k} \tag{12}$$

where N_{ik} is the number of occurrences of a_i in W_k, and L_k is the length of (number of symbols in) W_k.

Similarly, define the relative frequency of a_i with respect to T by:

$$f_T(a_i) = \frac{\sum_{k=1}^{n} q_k N_{ik}}{\sum_{k=1}^{n} q_k L_k} \tag{13}$$

with N_{ik} and L_i as before, q_k the number of occurrences of W_k in T. (Note that (12) is the special case of (13) with $q_k = 1$ for each k: in other words, one can think of the dictionary for present purposes as some arbitrarily selected member of the class

of texts which contain each word in the language exactly once.)

Now the problem can be stated as that of finding, given D and T, a function φ such that $\varphi(f_D(a_i)) = f_T(a_i)$ for each i. It is easy to see that, for particular pairs (D, T), such a function exists just in case there are no a_i, a_j such that $f_D(a_i) = f_D(a_j)$ while $f_T(a_i) \neq f_T(a_j)$. If this condition is satisfied, we have $\varphi = \{(p_1, p_1'), \ldots, (p_r, p_r'), \ldots\}$, where p_s and p_s' are rational numbers between 0 and 1, r is the number of distinct pairs $(f_D(a_i), f_T(a_i))$, and the second ellipsis indicated the possibility of extending the domain of definition of φ to numerical arguments not represented among the $f_D(a_i)$ for the given D. φ might turn out to be (hopefully) some simple or familiar function of a kind amenable to mathematical handling (it might even have an inverse φ^{-1})..

We are not, of course, much interested in ad hoc functions valid just for particular pairs (D, T) and amounting (if they exist) just to lists of arguments and image points. Rather, what we would like to know is whether any useful generalizations can be made about such functions for a variety of possible dictionaries and texts--in particular, for classes of both which might appropriately figure in mathematical models for natural languages. In other words, it becomes natural to pose the question: what must a "language" (consisting, for present purposes, just of a dictionary and one or more texts associated with it) be like to guarantee the existence of a function φ? Knowing this, we should be able to assess reasonably the likelihood that natural languages

belong to a restricted class for which such a relationship between dictionaries and texts holds good.

Two possibilities suggest themselves: viz., imposition of formal constraints on D (so that only certain dictionaries can qualify as proper dictionaries), and imposition of such constraints on T (so that only some texts will be admissible as proper texts).

Case 1. Conditions on D. For unrestricted T, the function φ exists if and only if, for each i and k, D is such that

$$\frac{N_{ik}}{L_k} = C_1 \tag{14}$$

where C_i is a constant for each i.

Proof: (a) Sufficiency. By the hypothesis that the condition is satisfied, and by definitions (12) and (13) above, we have that $N_{ik} = L_k C_i$, and hence:

$$f_D(a_i) = \frac{\sum_{k=1}^{n} L_k C_i}{\sum_{k=1}^{n} L_k} = \frac{C_i \sum_{k=1}^{n} L_k}{\sum_{k=1}^{n} L_k} = C_i \tag{15}$$

and

$$f_T(a_i) = \frac{\sum_{k=1}^{n} q_k L_k C_i}{\sum_{k=1}^{n} q_k L_k} = \frac{C_i \sum_{k=1}^{n} q_k L_k}{\sum_{k=1}^{n} q_k L_k} = C_i \tag{16}$$

so $f_D(a_i) = f_T(a_i)$ for each i, and φ exists and is, in fact, the identity function $\varphi(x) = x$.

(b) Necessity. Assuming that the condition is not satisfied, this means that there is at least one pair of words, say W_r and W_s,

such that $N_{ir}/L_r \neq N_{is}/L_s$ for some i. Now consider T_1, T_2 such that for T_1 $q_k = 0$ for $k \neq r$, $q_r > 0$, and for T_2 $q_k = 0$ for $k \neq s$, $q_s > 0$. Clearly we have, by definition (13), $f_{T_1}(a_i) = N_{ir}/L_r \neq f_{T_2}(a_i) = N_{is}/L_s$, but $f_D(a_i)$ is the same in both cases. Hence no function φ of the kind sought can exist.

Case 2. Conditions on T. For unrestricted D, the function φ exists if and only if T is such that $q_k = q$ for all $1 \leq k \leq n$.

Proof: (a) Sufficiency. By definition (13) and the hypothesis that the stated condition is satisfied,

$$f_T(a_i) = \frac{\sum_{k=1}^{n} qN_{ik}}{\sum_{k=1}^{n} qLk} = \frac{q \sum_{k=1}^{n} N_{ik}}{q \sum_{k=1}^{n} L_k} = f_D(a_i) \qquad (17)$$

as defined by definition (12). Thus φ exists and is, again, the identity function $\varphi(x) = x$.

(b) Necessity. Suppose the condition is not satisfied. Then there is at least one pair of words, say W_r and W_s, such that $q_r \neq q_s$. Now consider, most simply, the case in which $W_r = a$, $W_s = b$ (words of one symbol each), and $D = \{W_r, W_s\}$. Then $f_D(a) = f_D(b) = 1/2$, but by the hypothesis

$$f_T(a) = \frac{q_r}{q_r + q_s} \neq \frac{q_s}{q_r + q_s} = f_T(b) \qquad (18)$$

so that, to one and the same dictionary frequency, there correspond two distinct text frequencies, and no φ of the kind required can exist.

The necessary and sufficient condition on dictionaries, then, amounts to requiring that every dictionary word contain occurrences of each of the m elements of the alphabet A in proportions fixed for the vocabulary as a whole. This would obviously mean, inter alia, that each word of the vocabulary must have a length which is a multiple of some fixed base length equal to or greater than m (e.g., if m = 26 for written English, then every word should be more than 26 letters long). This seems an unlikely requirement to be satisfied by anything which we would accept as a dictionary of a natural language; certainly it is not satisfied by existing dictionaries which purport to be such.

The necessary and sufficient condition on texts, on the other hand, requires in effect that every text be, essentially, just q repetitions of the whole dictionary in some order, possibly scrambled and interspersed in any of a great many possible ways, but such that eventually every dictionary word has exactly the same text frequency. (This would be satisfied, e.g., by a language in which no sentence was grammatical unless it included every word in the dictionary exactly once.) This, again, seems a thoroughly implausible expectation for natural-language texts, however selected.

Further conclusions of this nature could be elaborated. Consider, say, a partition P of any given D, such that if $D_r \in P$ and $D_s \in P$ then $f_{D_r}(a_i) = f_{D_s}(a_i)$ for each i, and suppose the class of admissible texts restricted to those in which, if $W_h \in D_j$ and $W_k \in D_j$, then $q_h = q_k$. Then a function φ of the required kind exists for "languages" thus constrained. (They are, in fact,

related as special cases to this more general one.) In other words, the vocabulary must fall into strata with respect to text frequency (in any given admissible text) in such a way that all the strata are homogeneous in terms of the aggregate segmental makeup of the items they contain. This is a somewhat more realistic picture, but not very: e.g. if English conformed even approximately to such a model, we should expect the relative incidence of /ϑ/ and /z/ to be the same in high- and low-frequency subvocabularies.

In short, while making due allowance for the vagueness and incompleteness of our knowledge of the statistical structure of natural languages, one must conclude that it is highly unlikely that these are of such a type as to display any general functional relationship between "dictionary frequencies" and "text frequencies" of alphabet elements. One must, on the contrary, acknowledge the correctness of Trubetzkoy's insight that the difference between these logically distinct notions cannot safely be ignored in linguistics. Herdan's claimed refutation of this thesis turns out, as was shown above, to be nothing but one more instance of just the kind of conceptual confusion which Trubetzkoy warned against.