R E P O R T R E S U M E S

ED 011 649 AL 000 108

THE QUANTIFICATION OF FUNCTIONAL LOAD--A LINGUISTIC PROBLEM.

BY- HOCKETT, C.F.
RAND CORP., SANTA MONICA, CALIF.

REPORT NUMBER RM-5168-PR PUB DATE OCT 66
EDRS PRICE MF-$0.09 HC-$1.43 33P.

DESCRIPTORS- *PHONOLOGY, *LINGUISTIC THEORY, FUNCTIONAL LOAD, ENTROPY, PHONEMES, ALLOPHONES, *MEASUREMENT, *MATHEMATICAL LINGUISTICS, INFORMATION THEORY, SHANNON THEORY OF COMMUNICATION, SANTA MONICA

MEASUREMENT CRITERIA ARE DEVELOPED FOR THE QUANTIFICATION OF THE FUNCTIONAL LOAD OF THE PHONEMES OF A LANGUAGE. THE CONCEPT OF FUNCTIONAL LOAD OR YIELD, FROM CERTAIN THEORIES OF LINGUISTIC CHANGE, STATES THAT SOME CONTRASTS BETWEEN THE DISTINCTIVE SOUNDS OF A LANGUAGE DO MORE WORK THAN OTHERS BY OCCURRING MORE FREQUENTLY AND IN MORE LINGUISTIC ENVIRONMENTS THAN OTHER CONTRASTS. THE CRITERIA DEVELOPED, WHICH UTILIZE SHANNON'S MEASURE OF ENTROPY, CAN BE USED TO MEASURE THE WORK OF PHONEMES, ALLOPHONES, OR COMPONENTS AND ARE THUS APPLICABLE IN ALL CURRENT THEORIES OF PHONOLOGY. IT IS SUGGESTED THAT THERE IS A BALANCE POINT IN THE PHONOLOGICAL SYSTEM OF ANY LANGUAGE BETWEEN THE FORCE OF LEAST EFFORT LEADING TO LOWEST REDUNDANCY AND THE NEED TO BE UNDERSTOOD, WHICH LEADS TO LOWER RELATIVE ENTROPY. (KL)

MEMORANDUM
RM-5168-PR
OCTOBER 1966

ED011649

THE QUANTIFICATION OF FUNCTIONAL LOAD: A LINGUISTIC PROBLEM

C. F. Hockett

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

PREPARED FOR:
UNITED STATES AIR FORCE PROJECT RAND

The RAND Corporation
SANTA MONICA • CALIFORNIA

AL 000 108

AL 000 108

MEMORANDUM
RM-5168-PR
OCTOBER 1966

THE QUANTIFICATION OF FUNCTIONAL LOAD: A LINGUISTIC PROBLEM

C. F. Hockett

This research is sponsored by the United States Air Force under Project RAND—Contract No. AF 49(638)-1700—monitored by the Directorate of Operational Requirements and Development Plans, Deputy Chief of Staff, Research and Development, Hq USAF. Views or conclusions contained in this Memorandum should not be interpreted as representing the official opinion or policy of the United States Air Force.

DISTRIBUTION STATEMENT
Distribution of this document is unlimited.

The RAND Corporation
1700 MAIN ST. • SANTA MONICA • CALIFORNIA • 90406

PRECEDING PAGE BLANK-NOT FILMED

PREFACE

The function of a phonemic system is to distinguish the utterances of a language. One concept that has appeared in certain theories of linguistic change is that some contrasts between the phonemes of a language do more work than others. This Memorandum suggests and discusses criteria for the quantification of this concept for three possible cases. It should be of interest to theoreticians and investigators in linguistics.

The author, Professor of Linguistics at Cornell University, is a consultant to The RAND Corporation.

PRECEDING PAGE BLANK-NOT FILMED

SUMMARY

Measures of the linguistic load carried by a contrast are developed for three cases, in which the contrasts are taken to be, respectively, phonemic, allophonic, and componential. The load carried by a contrast is non-negative and zero if the "contrasted" units are identical, or if neither occurs in any environment in which the other is found. The measure proposed is the change in entropy of the system if the contrasted phonemes are coalesced; some problems peculiar to the allophonic case are discussed. If each distinct bundle of components is an allophone, the entropy of a given system is independent of point of view (phonemic, allophonic, or componential).

PRECEDING PAGE BLANK-NOT FILMED

CONTENTS

THE QUANTIFICATION OF FUNCTIONAL LOAD

1. INTRODUCTION

Of the many problems in linguistics on which the work of A. Martinet has shed light, one of the most interesting is the notion of functional load (or yield or burden).[1] In simplest terms, the notion is this. The function of a phonemic system is to keep the utterances of a language apart. Some contrasts between the phonemes in a system apparently do more of this job than others. For instance, in English there are hundreds of pairs of words that differ only in that one has /p/ where the other has /b/ (pat : bat, nipple : nibble, cap : cab), but only a very few are kept apart by /š/ versus /ž/ (for some speakers mesher : measure; for some Asher : azure; for some Aleutian : allusion). Presumably, then, the contrast between /p/ and /b/ does more work even in complete utterances than does that between /š/ and /ž/. At least, it is easier to coin a pair of whole utterances such as Don't take that cap : Don't take that cab than it is to find one for /š/ and /ž/, simply because there are more minimally different words of the first type.

Martinet's concern with functional load has been with its possible relevance in linguistic change. Suppose, for example, that in a particular community the random drift of sound change[2] threatens to wipe out a contrast that carries a certain functional load. If that load is sufficiently high, is it possible that exigencies of communication would prevent the impending coalescence? How high must the load be for this effect? Or, indeed, are adjustments by paraphrase always made, so that the coalescence is free to proceed without impairing communication?

---

[1] Discussed in various essays, most of them included in André Martinet, Economie des Changements Phonétiques, Berne, 1955. Martinet cites various European predecessors, but I have not consulted their works.

[2] "Random" is a difficult word; in particular, we are discussing in this very paper a kind of factor that perhaps militates against completely random randomness. However, I do find it necessary to accept (as many contemporary linguists do not) the neogrammarian hypothesis of "regularity", in a certain modernized version, for which see Sec. 3.2 of this paper and my "Sound Change", Language, Vol. 41, No. 2, pp. 185-204, 1965.

We can imagine the language of some community undergoing a series of sound shifts that obliterate all distinctions and reduce all utterances to the same dull blur. But we can be quite sure that, if anything like this has ever in fact happened, it happened long ago in the very earliest stages of human evolution, and the communities in question ceased to be viable and left no mark on subsequent history. For all the languages of today, and for all known to us via written records or the comparative method, we can assuredly assert that a certain minimal fluency is always maintained. If contrasts carrying a certain functional load are lost, new contrasts develop to take over the load, or some of the contrasts not lost assume an additional share.

This does not help us very much, because we do not know what the required "minimal fluency" is--nor do we even know how to express such a "minimal fluency" in quantitative terms.

Another possible approach is to observe actual instances of lost contrasts. For example, almost all varieties of American English have lost the contrast between /t/ and /d/ after a stressed vowel before an unstressed vowel, so that such pairs as matter and madder, latter and ladder, sweetish and Swedish have become completely homophonous. True enough, most of us Americans can resort, in an emergency, to an artificial spelling pronunciation that restores the distinction; but most of the time we don't. To a speaker of British English, this particular coalescence is one of the most striking features of the "slurred" speech of Americans. Yet American English is clearly viable without the contrast. Now, if we could meaningfully quantify the functional load carried by this particular contrast before it was lost, we would know, at least, that that much load is not enough to prevent a coalescence--because, in fact, it didn't.

The present paper has limited aims. I shall not express opinions of Martinet's various suggestions about functional load. I believe he has carried the matter as far as it can be carried without actual quantification. His hunches are incisive and suggestive, and perhaps in part wrong; but they cannot be confirmed or disproved merely by someone else's hunches. The next step in this area of investigation

must be the development of quantitative methods. That is what will be undertaken here--but only in an abstract way; I have done no counting.[3]

We shall consider three cases: functional load in terms of (1) phonemic contrasts, (2) allophonic contrasts, and (3) componential contrasts. There is currently a very active debate as to the relative importance of these three different sorts of units in phonology. Although I take certain positions in this debate, I do not want to import them into the present paper. By dealing with all three cases, we can supply the requisite formal tools for quantification regardless of how the debate is eventually resolved.

2. CASE 1 -- PHONEMES

2.1. Algebra

2.1.1. Let $\underline{L}^{\underline{m}}$ be a phonemic system with $\underline{m}$ phonemes /1/, /2/,...; /m/. In the terminology of algebraic grammar[4] (which uses some words familiar from ordinary linguistics, but in potentially deceptive special senses), these $\underline{m}$ phonemes are the characters of a linear alphabet. This means that: (1) $\underline{m}$ is finite; (2) the characters can be anything at all, as long as they are pairwise distinguishable; and (3) every utterance of the language of which $\underline{L}^{\underline{m}}$ is the phonemic system consists, without residue, of a string of occurrences of characters of $\underline{L}^{\underline{m}}$. (On the other hand, of course not every string of occurrences of characters is necessarily an utterance of the language).

---

[3] An earlier and briefer effort of mine to quantify functional load will be found in my Manual of Phonology, Indiana University Publications in Anthropology and Linguistics, No. 11, 1955. This earlier effort was vitiated by a mathematical error, which will be pointed out below.

[4] See my "Language, Mathematics, and Linguistics", to appear in Current Trends in Linguistics, Vol. 3, 1966. The elements of a set may conveniently be called "characters" merely if they are pairwise distinguishable. This may seem redundant, but it is not: in some sets that must be discussed mathematically, the elements are not distinguishable. For example, one can tell the difference between an electron and a proton, or between one electron and two electrons, but not between one electron and another electron.

In Sec. 2 we ignore any variations in actual physical properties of a character from one occurrence to another; in Sec. 3 we shall pay systematic attention to such variations. Also, in Sec. 2 we ignore any partial resemblances between characters (such as the feature of bilabiality common to English /p/ and /b/); this is underscored by the inclusion of "linear" above. In Sec. 4 we shall deal with such resemblances.

For our first step, we forget (for the moment) that the elements of $\underline{L}^{\underline{m}}$ are phonemes, and take $\underline{L}^{\underline{m}}$ merely as a finite set of characters (that is, of pairwise distinguishable elements). Let us consider the system $\underline{S}^{\underline{m}}$, whose elements are all the <u>partitions</u> of the characters of $\underline{L}^{\underline{m}}$. We can illustrate what is meant by a "partition" by assuming some small value for $\underline{m}$, say $\underline{m} = 4$. Then each of the following lines displays one of the possible partitions of the four characters; we label them for subsequent cross-reference:

$\underline{L}^{4}$. /1/ /2/ /3/ /4/
$\underline{L}_1^{3}$. /12/ /3/ /4/
$\underline{L}_2^{3}$. /13/ /2/ /4/
$\underline{L}_3^{3}$. /14/ /2/ /3/
$\underline{L}_4^{3}$. /23/ /1/ /4/
$\underline{L}_5^{3}$. /24/ /1/ /3/
$\underline{L}_6^{3}$. /34/ /1/ /2/
$\underline{L}_1^{2}$. /123/ /4/
$\underline{L}_2^{2}$. /12/ /34/
$\underline{L}_3^{2}$. /124/ /3/
$\underline{L}_4^{2}$. /13/ /24/
$\underline{L}_5^{2}$. /134/ /2/
$\underline{L}_6^{2}$. /14/ /23/
$\underline{L}_7^{2}$. /234/ /1/
$\underline{L}^{1}$. /1234/

We see that a partition is an assignment of all the characters to classes, where each character is assigned to some class, and none is assigned to more than one. That is, /12//3/ is not a partition of our four characters because one has been left out; and /12/ /13/ /4/ is not a partition because one has been assigned twice. The total number of partitions

of a set of $\underline{m}$ characters is a function of $\underline{m}$. For $\underline{m} = 2$ there are just two partitions; for $\underline{m} = 3$, there are 5; for $\underline{m} = 4$, as shown above, there are 15; for $\underline{m} = 5$, there are 52. For still higher values of $\underline{m}$, the number of partitions increases very rapidly. The fifteen partitions listed above are the elements of the system $\underline{S}^4$.

Suppose that $\underline{L}$ and $\underline{L}'$ are two of the partitions of a system $\underline{S}^{\underline{m}}$; and suppose, further, that we can (so to speak) change $\underline{L}$ to $\underline{L}'$ by "coalescing" one or more of the classes of $\underline{L}$ into a single class of $\underline{L}'$. This means that, if two characters $\underline{x}$ and $\underline{y}$ belong to different classes of $\underline{L}$, they may or may not belong to the same class of $\underline{L}'$; but if $\underline{x}$ and $\underline{y}$ belong to the same class of $\underline{L}$, they must also belong to the same class of $\underline{L}'$. If this relation holds between a particular pair of partitions $\underline{L}$ and $\underline{L}'$, we say that $\underline{L} \geq \underline{L}'$. For example, in $\underline{S}^4$, whose elements are listed above, $\underline{L}_1^3 \geq \underline{L}_2^2$, $\underline{L}_6^3 \geq \underline{L}_2^2$, and even $\underline{L}_1^3 \geq \underline{L}_1^3$ (indeed, if $\underline{L}$ is any partition, then $\underline{L} \geq \underline{L}$); but, clearly, $\underline{L}_1^3 \not\geq \underline{L}_2^3$ and $\underline{L}_1^3 \not\geq \underline{L}^4$.

Figure 1 displays the system $\underline{S}^4$ graphically. The nodes represent the fifteen partitions, and are appropriately labelled. If, given two distinct partitions $\underline{L}$ and $\underline{L}'$, it is the case that $\underline{L} \geq \underline{L}'$, then, in the figure, it is possible to pass from $\underline{L}$ to $\underline{L}'$ along one or more connecting lines, moving generally from left to right (perhaps slanting upwards or downwards, but never backing up from right to left).

The system $\underline{S}^{\underline{m}}$ of all partitions of a set $\underline{L}^{\underline{m}}$ of $\underline{m}$ characters, with the relation $\geq$ defined as we have defined it, is known to be an exemplification of a formal mathematical system called a relatively complemented semimodular lattice (or matroid lattice).[5] Any property shared by all matroid lattices will, of course, hold for any system $\underline{S}^{\underline{m}}$, even when we put a different interpretation on our symbols. For the application we have in view, most of these properties are quite irrelevant. But we do need to note the following, all of which can easily be read from Fig. 1 for the specific example displayed there:

---

[5] Garrett Birkhoff, Lattice Theory, 2nd Ed. 1949, p. 107.

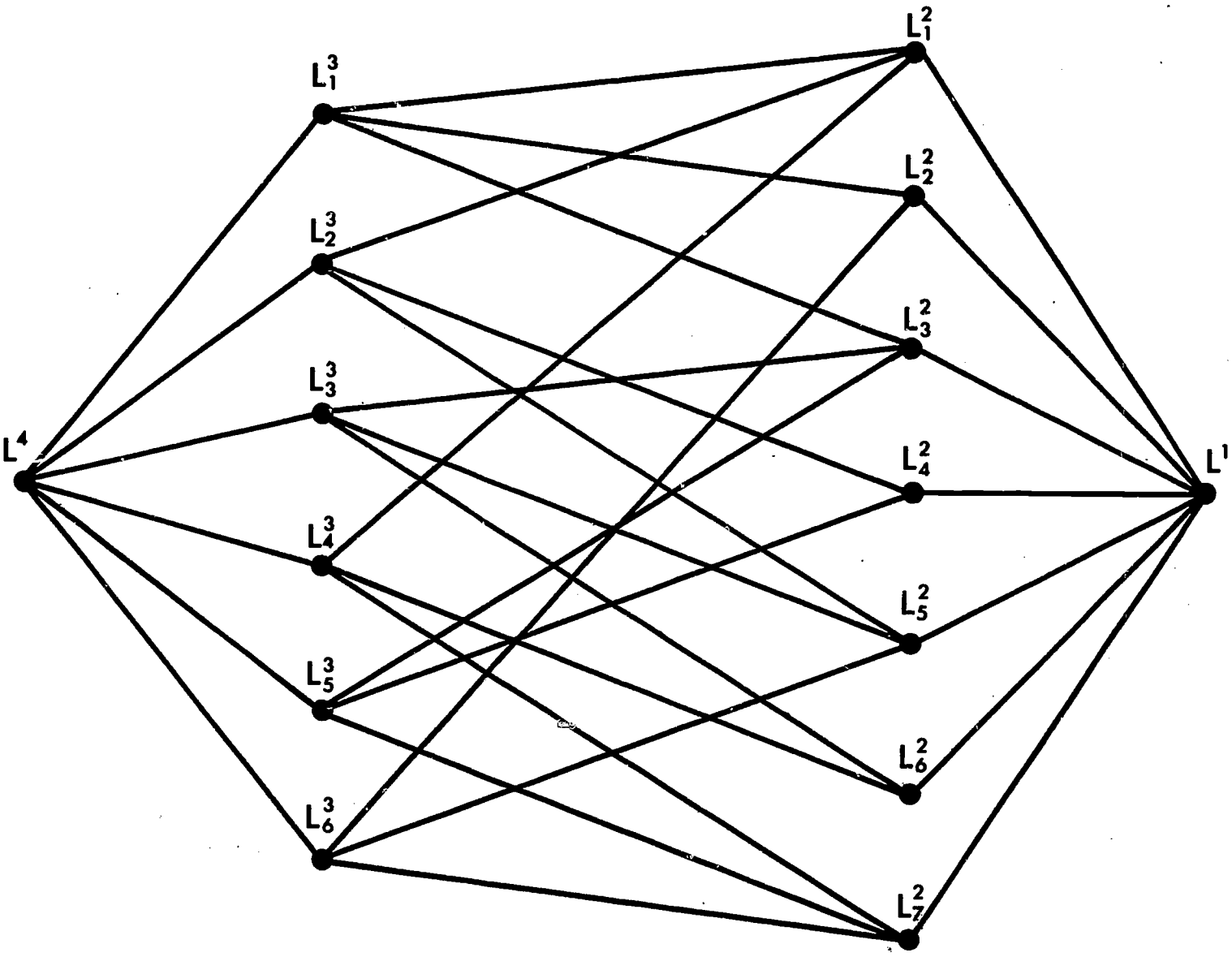

Fig. 1—The system S^4

(1) A system $\underline{S}^{\underline{m}}$ includes a unique universal upper bound $\underline{L}^{\underline{m}}$ such that, if $\underline{L}$ is any partition of $\underline{S}^{\underline{m}}$, then $\underline{L}^{\underline{m}} \geq \underline{L}$, and a unique universal lower bound $\underline{L}^{1}$ such that, if $\underline{L}$ is any partition of $\underline{S}^{\underline{m}}$, then $\underline{L} \geq \underline{L}^{1}$. In Fig. 1, these unique elements appear as the leftmost and rightmost nodes ($\underline{L}^{4}$ and $\underline{L}^{1}$).

(2) Given any element $\underline{L}$ other than the universal upper bound, it is possible to find an element $\underline{L}'$ such that $\underline{L}' \geq \underline{L}$ but such that there is no element $\underline{L}''$ for which $\underline{L}' \geq \underline{L}'' \geq \underline{L}$; we shall then say that $\underline{L}' \rightarrow \underline{L}$. Similarly, given any element $\underline{L}$ other than the universal lower bound, it is possible to find an element $\underline{L}'$ such that $\underline{L} \geq \underline{L}'$ but such that there is no element $\underline{L}''$ for which $\underline{L} \geq \underline{L}'' \geq \underline{L}'$; we shall say that $\underline{L} \rightarrow \underline{L}'$.

(3) A chain from $\underline{L}^{\underline{m}}$ to $\underline{L}^{1}$ is any set of elements $\underline{L}_1, \underline{L}_2, \ldots, \underline{L}_n$ of $\underline{S}^{\underline{m}}$ such that $\underline{L}^{\underline{m}} = \underline{L}_1 \rightarrow \underline{L}_2 \rightarrow \cdots \rightarrow \underline{L}_n = \underline{L}^{1}$. Clearly (see Fig. 1) the number of elements in any chain in $\underline{S}^{\underline{m}}$ is just $\underline{m}$. Two chains in $\underline{S}^{\underline{m}}$ are distinct if one of them contains at least one element of $\underline{S}^{\underline{m}}$ not in the other. The number of distinct chains in $\underline{S}^{\underline{m}}$ is

$$\frac{\underline{m}(\underline{m}-1)^2(\underline{m}-2)^2 \ldots 2^2 .1}{2^{\underline{m}} - 1} = \frac{\underline{m}!\,(\underline{m}-1)!}{2^{\underline{m}} - 1}.$$

2.1.2. Now we return to the interpretation of $\underline{L}^{\underline{m}}$ as a phonemic system. We shall imagine that we can operate on $\underline{L}^{\underline{m}}$ in the following way to produce a new phonemic system: we select any two phonemes /$\underline{i}$/ and /$\underline{j}$/ of $\underline{L}^{\underline{m}}$ and agree to ignore the difference between them, so that the new system will not contain either /$\underline{i}$/ or /$\underline{j}$/ but only a new phoneme /$\underline{ij}$/; otherwise the new system is just like the old one. This finds its diachronic analog in the coalescence of two phonemes of an earlier stage of a language to form a single phoneme at a later stage. Note that our notation "/$\underline{ij}$/" does not represent a string of two phonemes, as it would in ordinary linguistic usage. When we need to represent such a string, we will insert commas: "/$\underline{i},\underline{j}$/".

With this interpretation, the system $\underline{S}^{\underline{m}}$ consists of a basic phonemic system, $\underline{L}^{\underline{m}}$, plus just all those other phonemic systems that can be obtained from $\underline{L}^{\underline{m}}$ by one or more pairwise coalescences of the kind just described. Thus, in Fig. 1, the phonemes of the basic system

$\underline{L}^4$ are /1/, /2/, /3/, and /4/; if we now ignore the distinction between /1/ and /2/, we get the system $\underline{L}_1^3$ with phonemes /12/, /3/, and /4/; and so on. This change of interpretation does not in any way alter the fact that $\underline{S}^4$ is a matroid lattice, with all the formal properties of any such system. Of course, all of this is a matter of mathematical convenience; in particular, $\underline{L}^1$ obviously could not be a real phonemic system, and even $\underline{m} = 4$ is too low a value for a real one. But our application of the formal apparatus will be such that these departures from reality do not matter.

2.2. Measure

2.2.1. The notion of functional load is that a phonemic system $\underline{L}^m$ has a (quantifiable) job to do, and that the contrast between any two phonemes, say /a/ and /b/, carries its share. There is only one way in which the contrast between /a/ and /b/ can stop doing its share of the work: that is for /a/ and /b/ to coalesce, yielding a new phoneme /ab/ and hence a new system $\underline{L}_1^{m-1}$. It makes sense to infer that if the contribution of the contrast between /a/ and /b/ is thus withdrawn, one of two things must happen: (1) the job done by the whole system is rendered smaller; or (2) the total job remains the same, and the share no longer carried by the lost contrast is somehow divided up among the contrasts that remain. We shall first explore alternative (1), returning to (2) below in Sec. 2.2.7.

2.2.2. We assume that the load of work done by a whole phonemic system can be expressed in the form of a nonnegative real number (a "negative" load seems not to make sense). Let $\underline{f}(\underline{L})$ be the load carried by a system $\underline{L}$, and let $\underline{f}$ (/a/, /b/) be the share carried by the contrast between /a/ and /b/. Then, under alternative (1), we have

$$\underline{f}(\underline{L}_1^{m-1}) = \underline{f}(\underline{L}^m) - \underline{f}(/a/,/b/)$$

or, transposing,

$$\underline{f}(/a/,/b/) = \underline{f}(\underline{L}^m) - \underline{f}(\underline{L}_1^{m-1}).$$

The equation is more useful in this second form, because it suggests that if we can find an appropriate measure of the functional loads of whole systems of $\underline{S}^m$, then a suitable and suitably related measure of the functional loads of individual contrasts is immediately at hand.

A desirable property for $\underline{f}$ is that $\underline{f}(\underline{L}^1) = 0$. The reason is that a system with no contrasts can carry no information; as we have already said, $\underline{L}^1$ is only a mathematical convenience, not by any stretch of the imagination a phonemic system. Similarly, for any phoneme /x/ we must have $\underline{f}(/x/,/x/) = 0$: for if we "operate" on a system by agreeing to ignore the difference between a phoneme and itself, we have not changed the system at all.

Another desirable property is that the contribution of any contrast should be at least zero: that is, if $\underline{L}_1 \rightarrow \underline{L}_2$, then $\underline{f}(\underline{L}_1) \geq \underline{f}(\underline{L}_2)$. The justification for allowing a zero load, but not a negative one, requires some discussion of phonological theory.

Two phonemes may have nonintersecting distributions, in the sense that neither occurs in any environment in which the other is found. By one possible phonemicization of English, this is true of /h/ and /ŋ/. But if English /h/ and /ŋ/ have this distribution, then no pair of utterances can differ only in that one has /h/ where the other has /ŋ/. Consequently, a coalescence of /h/ and /ŋ/ (however difficult to imagine phonetically) would destroy no contrasts of whole utterances; the total load carried by the system would be undiminished. Therefore $\underline{f}(/h/,/ŋ/) = 0$.

If two phonemes are not in nonintersecting distribution, however, then there must exist at least one environment in which both occur. Now, for two allophones to be phonemically different, it is sufficient that they should be in direct contrast in a <u>small</u> environment. It is therefore possible for two phonemes to stand in contrast in small environments, and still not serve as the sole differentia of two whole words or two whole utterances. On the other hand, it is quite impossible for two words or two utterances to be kept apart by a single-phoneme difference unless the two phonemes involved also contrast in small environments. The inferences are as follows. Suppose we have a method of measuring functional load by successive approximations, and that the earlier approximations involve the inspection only of small environments. For these earlier approximations, any contrast between phonemes that are not in nonintersecting distribution will prove to carry a positive share of the total load. As the approximations

continue, and larger environments are taken into consideration, some of these shares may become vanishingly small, but none can become negative.

We ask next if the measure ought to be <u>additive</u>, in the sense that the sum of all loads carried by all contrasts between pairs of phonemes in $\underline{L}^{\underline{m}}$ would be just the load carried by $\underline{L}^{\underline{m}}$. The hasty answer is affirmative, but wrong.[6] We must remember the nature of a phonemic system. If a system $\underline{L}^{\underline{m}}$ were a set of $\underline{m}$ elements each of which <u>individually</u> made some contribution to a measure defined for the whole system, then additivity might be natural. We would assume, in such a case, that the elements of $\underline{L}^{\underline{m}}$ could be deleted, one by one, until all were gone, and that the measure would correspondingly diminish to zero. But a phonemic system is not composed of elements that can be deleted in this way, and the measure is not defined for single elements, only for pairs. The pairs are not independent. They cannot be "deleted"; they can only coalesce. If our first step is to coalesce /a/ and /b/ into a new phoneme /ab/, then it is no longer possible to perform a similar operation on any pair /a/, /x/, or /b/, /x/, since the phonemes /a/ and /b/ are no longer present.

2.2.3. To summarize: we want the measure $\underline{f}(\underline{L})$ to have the following properties, all but the last of which have now been discussed:

(P1) $\underline{f}(\underline{L}) \geq 0$ for all $\underline{L}$ in $\underline{S}^{\underline{m}}$.

(P2) $\underline{f}(\underline{L}^{1}) = 0$.

(P3) If $\underline{L}_1 \rightarrow \underline{L}_2$, then $\underline{f}(\underline{L}_1) \geq \underline{f}(\underline{L}_2)$.

(P4) If $\underline{L}_1 \rightarrow \underline{L}_2$ and $\underline{L}_3 \rightarrow \underline{L}_4$, and if $\underline{L}_1$ and $\underline{L}_3$ contain /a/ and /b/ while $\underline{L}_2$ and $\underline{L}_4$ contain /ab/, then $\underline{f}(\underline{L}_1) - \underline{f}(\underline{L}_2) = \underline{f}(\underline{L}_3) - \underline{f}(\underline{L}_4)$.

Property P1, of course, follows from P2 and P3.

Property P3 guarantees that the load carried by the universal upper bound of $\underline{S}^{\underline{m}}$ is the upper bound of the loads carried by the systems of $\underline{S}^{\underline{m}}$.

---

[6] This hasty answer was the error in my earlier discussion, cited in footnote 3. The error was called to my attention by William S.-Y Wang.

Property P4 guarantees that the load carried by a contrast does not vary depending on where within $\underline{S}^{\underline{m}}$ we choose to measure it.

If a measure has all four of these properties, then we can extend it to cover any subset of the phonemes of any system $\underline{L}$ of $\underline{S}^{\underline{m}}$ as follows: Let $\underline{L}$ be a system in which the phonemes /a/, /b/,..., /i/ are all distinct, and let $\underline{L}'$ differ from $\underline{L}$ only in that /a/, /b/,..., /i/ have all coalesced into a single phoneme /ab...i/. Then we define

$$(D1) \quad \underline{f}(/a/, /b/,.../i/) = \underline{f}(\underline{L}) - (\underline{L}').$$

When there are only two phonemes in the set, /a/ and /b/, then this reduces to the second form of the equation in Sec. 2.2.2., and gives us the appropriate measure of the functional load of a single contrast. Further, if /a/ = /b/, then $\underline{L}' = \underline{L}$ and $\underline{f}(\underline{L}) - \underline{f}(\underline{L}') = 0$, so that, as desired, $\underline{f}(/x/, /x/) = 0$ for any phoneme /x/.

Any measure with the first three of these properties shows what we may call <u>additivity along a chain</u>. Suppose we move along any chain of $\underline{S}^{\underline{m}}$ from $\underline{L}^{\underline{m}}$ to $\underline{L}^{1}$. Each step involves coalescing a single pair of phonemes /x/, /y/ of the predecessor into a single phoneme /xy/ of the successor. The sum of $\underline{f}(/x/,/y/)$ for all pairs coalesced in passing from $\underline{L}^{\underline{m}}$ to $\underline{L}^{1}$ is just $\underline{f}(\underline{L}^{\underline{m}})$. This sum is obviously the same regardless of choice of chain. The individual addends need not be the same. But if the measure also has property P4, then the addends along any chain are a permutation of those along any other chain that involves just the same coalescences.

2.2.4. A measure that meets the requirements proposed in Sec. 2.2.3. is Shannon's entropy $\underline{H}$ (in binits per symbol).[7]

Let $\underline{p}$ be a relative frequency (or a probability), $0 \leq \underline{p} \leq 1$. Then we define:

$$\underline{I}(\underline{p}) = -\underline{p} \log_2 \underline{p} \qquad \underline{p} \neq 0$$
$$= 0 \qquad \underline{p} = 0.$$

---

[7] Claude E. Shannon, "The Mathematical Theory of Communication", in Claude E. Shannon and Warren Weaver, <u>The Mathematical Theory of Communication</u>, Urbana, 1949.

Now let $p/i/$ be the relative frequency of phoneme /i/, and let $I(p/i/) = I/i/$. Then the first-order approximation to the entropy of L^m is defined as:

$$H_1 = \sum_{i=1}^{m} I/i/.$$

Similarly, if $p/i_1, i_2,\ldots,i_n/$ is the relative frequency of the string indicated, then the nth-order approximation is

$$H_n = \frac{1}{n} \sum_{i_j=1}^{m} I/i_1, i_2,\ldots,i_n/ ,$$

where each i_j ranges independently from /1/ to /m/.

If there is reason to believe that the proper limit exists, then we can define the entropy of L^m to be

$$H(L^m) = \lim_{n\to\infty} H_n (L^m).$$

Otherwise we can define $H = H_n$ for some suitably large n; this is discussed below in Sec. 2.2.6.

2.2.5. For any system L in S^m, we define $f(L) = H(L)$. This measure has the four properties set forth in Sec. 2.2.3. Most of the proof is simple. We need only consider property P3.

The discussion of Sec. 2.2.2. shows that all we need demonstrate here is that two phonemes, both of which occur in some small environment, cannot make a zero or negative contribution to a sufficiently low-order approximation to the entropy. Let $p/a/ = p$ and $p/b/ = q$ be the relative frequencies of /a/ and /b/ in the particular environment. Then both p and q lie in the open unit interval, as does their sum $p + q$, except that $p + q = 1$ just if /a/ and /b/ are the only phonemes that occur in the given environment.

We now show that, for all possible values of p and q with the constraints just given, $I(p) + I(q) > I(p+q)$. The proof is direct (rather than contrapositive):

$$\underline{p} < \underline{p} + \underline{q} \quad \text{and} \quad \underline{q} < \underline{p} + \underline{q}$$

$$\log \underline{p} < \log (\underline{p}+\underline{q}) \quad \text{and} \quad \log \underline{q} < \log (\underline{p}+\underline{q})$$

$$\underline{p} \log \underline{p} < \underline{p} \log (\underline{p}+\underline{q}) \quad \text{and} \quad \underline{q} \log \underline{q} < \underline{q} \log (\underline{p}+\underline{q}).$$

Adding these two inequalities, we get

$$\underline{p} \log \underline{p} + \underline{q} \log \underline{q} < (\underline{p}+\underline{q}) \log (\underline{p}+\underline{q})$$

or

$$- \underline{p} \log \underline{p} - \underline{q} \log \underline{q} > - (\underline{p}+\underline{q}) \log (\underline{p}+\underline{q})$$

which, by the definition of $\underline{I}$, is the proposition to be proved.

2.2.6. Although the mathematically most tempting definition of $\underline{H}$ involves a limiting process, this raises the question as to whether the desired limit exists. In one practical sense, this does not matter: any actual computations of functional loads based on our formalism are going to settle for relatively low-order approximations.

But there is also an empirical reason why we should perhaps not worry about the limit even if, mathematically, it does exist. People do not speak in indefinitely long utterances. Furthermore, truly long utterances (such as a political harangue or a university lecture) are broken into successive segments each of which has some sort of unity and cohesiveness about it. In some languages, words (when properly defined) have such unity and cohesiveness; in others, phonemic phrases of some sort do. It may perhaps be suggested that an appropriate definition of $\underline{H}$ is $\underline{H} = \underline{H}_{\underline{k}}$, where $\underline{k}$ is the length in phoneme-occurrences of the longest cohesive unit of whatever type is chosen; or, perhaps, we should let $\underline{k}$ be the average length in phoneme-occurrences of such units, where the averaging is based on text-frequency, not list-frequency.

2.2.7. Alternative Measure (1). The entropy $\underline{H}$ used in Sec. 2.2.4. is in binits per symbol-occurrence. If the average rate of emission of phoneme-occurrences is $\underline{r}$ per second, then $\underline{\hat{H}} = \underline{rH}$ is the entropy measured in shannons (binits per second).

A way to achieve alternative (2) of Sec. 2.2.1. is to assume that if $\underline{H}$ is decreased, $\underline{r}$ must increase enough to keep $\underline{\hat{H}}$ unchanged. That is, $\underline{\hat{H}}$ becomes a constant for any basic system $\underline{L}^{\underline{m}}$ and its derivatives within $\underline{S}^{\underline{m}}$. Then $\underline{H}(/a/,/b/)$, as already defined, can be regarded as an indirect

measure of the increase of r required to compensate for the loss of the contrast between /a/ and /b/. Let L' be the system derived from L^m by the coalescence of /a/ and /b/; let r' be the required new rate of emission; and let s (/a/,/b/) = r'/r. Then

$$s/a/,/b/) = \frac{H(L^m)}{H(L')} = \frac{H(L^m)}{H(L^m) - H(/a/,/b/)} .$$

It is obvious that as we pass along a chain of S^m towards L^1, s and r' both increase without limit.

Since s is defined only for contrasts, we cannot test it for properties P1-P3 of Sec. 2.2.3. unless we somehow extend it to systems as well as contrasts, but there seems to be no natural way of doing this. We can test for property P4, and it turns out that s does not have this property: in general, an "early" loss of a particular contrast (that is, a loss closer to L^m on some chain from L^m to L^1) entails a smaller s than a "late" loss of the same contrast. However, we could always agree to measure s starting with L^m. In any event, s is related so simply to H that information about s, if wanted, requires only trivial computation beyond that for H.

The possible empirical significance of s is not clear. Offhand, one might guess that all human languages have just about the same amount of work to do. No language is spoken always at the same rate, but there do seem to be variations from language to language as to "normal" or "average" rate, perhaps also as to maximum intelligible rate. If the guess just mentioned is valid, one might suspect that the average rate, or the maximum rate, is higher for a language with a relatively more complicated phonemic system. Impressionistically, Japanese seems to be spoken faster than German or Russian, and Hawaiian perhaps faster than Japanese. We need accurate measurements rather than impressions; perhaps some have been made, but I do not know about them. We do know that the Lord's Prayer is longer (in total number of phoneme occurrences) in those Chinese dialects that have less phonemic differentiation than in those that have more;[8] but this would attest to our guess only if

[8] Vide Y. R. Chao.

the time of delivery for the prayer were about the same for all dialects, and on this we have no information.

2.2.8. Alternative Measure (2). A different sort of measure is supplied by Shannon's relative entropy $\underline{C}$.[9] For any system $\underline{L}^{\underline{m}}$, $\underline{C}(\underline{L}^{\underline{m}}) = \underline{H}(\underline{L}^{\underline{m}})/\log_2\underline{m}$. The denominator in this expression is the entropy of a system with $\underline{m}$ phonemes in which all phonemes are constantly equiprobable, and is the maximum entropy achievable (neglecting channel noise) with $\underline{m}$ elements. Since both numerator and denominator are in binits, $\underline{C}$ itself is an absolute number; it is also independent of time, since $\underline{\hat{H}}(\underline{L}^{\underline{m}})/\underline{t}\log_2\underline{m} = \underline{t}\underline{H}(\underline{L}^{\underline{m}})/\underline{t}\log_2\underline{m} = \underline{C}(\underline{L}^{\underline{m}})$. Since for $\underline{L}^1$ the formula reduces to the indeterminate form 0/0, we must specify that $\underline{C}(\underline{L}^1) = 0$. Definition (D1) of Sec. 2.2.3. now says that the load carried by any set of contrasting phonemes is the loss in relative entropy entailed by the loss of the contrasts.

This measure has properties P1 and P2, but not P3 or P4. For, consider an artificial system $\underline{L}^4$ with four phonemes /a/, /b/, /c/, and /d/, each with constant probability ¼. Let $\underline{L}'_1$ contain /ab/, /c/, and /d/; let $\underline{L}'_2$ contain /a/, /b/, and /cd/; and let $\underline{L}''$ contain /ab/ and /cd/. Then $\underline{C}(\underline{L}^4) = \underline{C}(\underline{L}'') = 1$; but $\underline{C}(\underline{L}'_1) = \underline{C}(\underline{L}^1_2) < 1$. But $\underline{L}'_1$ differs from $\underline{L}^4$ only in that /a/ and /b/ have coalesced, and $\underline{L}''$ differs from $\underline{L}^1_2$ only in the same way. Thus (1) the load carried by the contrast between /a/ and /b/ in $\underline{L}'_2$ is negative, contrary to property P3, and (2) the load carried by the contrast between /a/ and /b/ depends on where in $\underline{S}^4$ it is measured, contrary to property P4.

These facts, in my opinion, constitute serious defects in $\underline{C}$ as a measure of functional load. On the other hand, Shannon has shown that approximations to $\underline{C}$ converge more rapidly than do those to $\underline{H}$. For real phonemic systems, which are much more complicated, of course, than our artificial example, it may be that $\underline{C}$ affords a more easily computable and sufficiently accurate approximation to $\underline{H}$; but this should be tested empirically.

---

[9] This is the measure used by William S.-Y. Wang and James W. Thatcher in "The Measurement of Functional Load," Report No. 8, Communication Sciences Laboratory, The University of Michigan, April 1962.

3. CASE 2 -- ALLOPHONES

3.1. For application to historical linguistics, functional load in terms of phonemes will not usually suffice. In the course of history, it is in the first instance not phonemes but allophones that change as to physical properties, and it is certain of these allophonic changes that entail restructurings of the phonemic system.

Let $\underline{L}^{\underline{m}}$ involve phonemes /1/, /2/,...,/m/; and let the allophones of phoneme /$\underline{i}$/ be $[\underline{i}_1]$, $[\underline{i}_2]$, ..., $[\underline{i}_{\underline{r}_i}]$, where $\underline{r}_{\underline{i}} \geq 1$ for every $\underline{i}$. Also, let:

$$\sum_{\underline{i}=1}^{\underline{m}} \underline{r}_{\underline{i}} = \underline{r}.$$

Then $\underline{L}^{[\underline{r}]}$ is the same system as $\underline{L}^{\underline{m}}$, but viewed as composed of allophones rather than phonemes.

3.2. In the kind of linguistics that uses allophones and phonemes, we have the first two of the following assumptions about allophones; in any kind of linguistic theory we have the second two:

(A1) A given allophone in a given environment always represents the same phoneme.

(A2) If two allophones belong to the same phoneme, they are in nonintersecting distribution.

(A3) In course of time, two allophones may coalesce.

(A4) In course of time, a single allophone may split into two, but only if the two new allophones are in nonintersecting distribution.

Assumption A1 guarantees that we can know what phoneme an allophone represents without knowing anything about the grammar or semantics of the utterance in which it occurs (separability of phonology from grammar). Assumption A4 is the modern form of the neogrammarian principle of regularity of sound change. From A2, two other facts immediately follow:

(A2.1) If two allophones are in intersecting distribution, they belong to different phonemes.

(A2.2) If a phoneme occurs in an environment, it is represented there always by the same allophone.

3.3. Suppose, now, that we examine a system $\underline{L}^{[r]}$, related to a given $\underline{L}^{m}$, but that we disregard the phonemic affiliations of the allophones and attempt to measure the functional load of the system directly in terms of allophones and their distribution. Let our measure be the $\underline{H}$ of Sec. 2.2.4. We have the following

Theorem 1. $\underline{H}(\underline{L}^{[r]}) = \underline{H}(\underline{L}^{m})$.

This says that the entropy of a system is the same whether we measure it in terms of allophones or of phonemes; also, that the entropy is invariant from one phonemicization to another as long as all phonemicizations accord with assumptions A1 and A2 above.

Proof. A pair of allophones contribute nothing to the load unless they are in contrast. If they are in contrast, then, by A2.1, they belong to different phonemes, and, by A2.2, each phoneme is represented by just this allophone in any environment in which the two allophones contrast. Thus the relative frequencies of the allophones, in any such environment, are just the relative frequencies of the phonemes they represent. Since these relative frequencies of phonemes in environments are just the determinants of $\underline{H}(\underline{L}^{m})$, exactly the same (nonzero) relative frequencies determine $\underline{H}(\underline{L}^{[r]})$.

3.4. An allophonic split or coalescence can affect a phonemic system in various ways. In some instances, the only change is in what we might call the "internal economy" of one or more phonemes: that is, a phoneme gains or loses an allophone, but continues to be represented in the same environments as before; or an allophone switches its affiliation from one phoneme to another, but without changing the number of phonemes and without altering the contrasts in any environment. For our purposes, any alterations of the kinds just described are irrelevant. An allophonic change is system-changing if and only if it does one of the following: (1) changes the number of phonemes in the system, or (2) alters the contrasts in some environment. From A4, an allophonic split cannot be system-changing. From the other assumptions, a coalescence cannot be system-changing if the two allophones belong to the same phoneme before the coalescence. This leaves two types of coalescence that are, or may be, system-changing:

(1) Suppose /a/ and /b/ are not in contrast, and that $[a_1]$ and $[b_1]$ coalesce. No new contrasts are introduced, nor are any lost, so that the load of the system is unchanged. But if, say, $[a_1]$ is the only allophone of /a/, and if the new allophone $[a_1b_1]$ belongs to /b/, then the number of phonemes has been reduced by one. Unfortunately, there is no way of stating (in purely formal terms) whether the coalescent allophone $[a_1b_1]$ will be assigned to /b/ or to /a/. This depends on "phonetic similarity", or on distribution of phonological components--or, indeed, on the individual linguist's taste and prejudices.

(2) Suppose /a/ and /b/ are in contrast, and that $[a_1]$ and $[b_1]$, the respective representatives of /a/ and /b/ in one of the environments in which both occur, coalesce. A coalescence under these conditions is always system-changing. But the exact consequences depend on further factors. We need to know whether one of the allophones, say $[a_1]$, is or is not the only allophone of its phoneme. And we need to know whether the coalescence is <u>compensated</u> or <u>uncompensated</u>.

Suppose $[c_1]$ and $[c_2]$, allophones of /c/, occur respectively in environments $\underline{E}_1$ and $\underline{E}_2$, and that the sole difference between $\underline{E}_1$ and $\underline{E}_2$ is that $\underline{E}_1$ involves an allophone [x] exactly where $\underline{E}_2$ involves a different allophone, [y]. By A2.1, then, [x] and [y] must belong to different phonemes, because they occur in identical environments--namely, what is left of either $\underline{E}_1$ or $\underline{E}_2$, plus /c/. We can therefore take $[x] = [a_1]$ and $[y] = [b_1]$. Now suppose that $[a_1]$ and $[b_1]$ coalesce. This coalescence renders $\underline{E}_1$ and $\underline{E}_2$ identical, so that $[c_1]$ and $[c_2]$, by A2.1, must now belong to different phonemes. A coalescence of $[a_1]$ and $[b_1]$ under these conditions is compensated. Under any other conditions, it is uncompensated.

The effect of coalescences of type (2) on phoneme-count is thus as follows: An uncompensated coalescence leaves the phoneme-count unchanged if both coalescing allophones belong to phonemes that have other allophones, but reduces the count by one if one of the coalescing allophones is the only allophone of its phoneme. For example, let the one-allophone phoneme be /a/; then the coalescent product $[a_1b_1]$ must belong to /b/--the indeterminacy of type (1) is not encountered. A compensated coalescence increases the phoneme-count by one if both

coalescing allophones belong to phonemes that have other allophones, but it leaves the count unchanged if one of the coalescing allophones is the only allophone of its phoneme--for though a phoneme is lost by merger with another, another phoneme is split in two.

3.5. The effect of coalescence on total load can be summarized by saying that the total load cannot be increased. An uncompensated coalescence may reduce it. A compensated coalescence cannot increase it: the new contrast between $[c_1]$ and $[c_2]$ can make exactly the same contribution made before the change by $[a_1]$ and $[b_1]$, but not more. The load loss is zero, for a compensated coalescence, if and only if the environments $\underline{E}_1$ and $\underline{E}_2$, involving respectively $[a_1]$ and $[b_1]$, account for all the occurrences of at least one of those two allophones.

This takes care of the only possibly nonobvious part of the proof of the following

Theorem 2. If [x] and [y] are any two allophones, not necessarily distinct, and /x/ and /y/ are the phonemes, not necessarily distinct, to which [x] and [y] respectively belong, then $\underline{H}([x],[y]) \leq \underline{H}(/x/,/y/)$.

4. CASE 3--COMPONENTS

4.1. We shall now speak of a system $\underline{L}^{|\underline{t}|}$ whose elements $|1|$, $|2|,\ldots, |t|$ are the characters of a componential alphabet. Each character is a "simultaneous bundle" (formally, merely an unordered set) $\{\underline{e}_1, \underline{e}_2,\ldots, \underline{e}_n\}$ of one or more distinct components, of which there is a finite stock $\underline{F} = \{\underline{c}_1, \underline{c}_2,\ldots,\underline{c}_p\}$; that is, in any character each $\underline{e}_i$, $1 \leq \underline{i} \leq \underline{n}$, is one or another of the $\underline{c}_j$ of the stock $\underline{F}$. At least some of the components, we assume, occur in more than one of the characters. From one character to another, $\underline{n}$ can vary, but we assume that at least one character contains more than one component. Two characters are distinct if and only if one contains at least one component missing from the other. The components are pairwise distinguishable; therefore so are the characters. Every utterance of the language of which $\underline{L}^{|\underline{t}|}$ is the phonological system consists, without residue, of a string of occurrences of bundles of components; but not every set of components constitutes a bundle, and not every string of occurrences of bundles is necessarily an utterance of the language.

Traditional phonemic theory attempted to set forth and to exploit the redundancy of any natural language by grouping allophones into units called "phonemes" on the basis of complementary distribution and phonetic similarity. The purposes of this operation are no longer entirely clear to me, except for the obvious but somewhat extraneous aim of achieving a simple yet accurate notation. The issue was confused by a desire to achieve, at the same time, as simple and nearly invariant as possible a notation for elements of a very different kind (morphemes).

The componential approach does not manipulate simultaneous bundles the way traditional phonemic theory manipulated allophones. Instead, limitations of distribution and co-occurrence are discovered and dealt with directly in terms of the components themselves, environments being simultaneous as well as successive. A regularly occurring and regularly observable feature of articulation or sound is not necessarily recognized as a component wherever it occurs; some or all of its occurrences may turn out to be <u>predictable</u> from the occurrences and arrangements of other features, provided the latter are formally recognized as components. Different practitioners of our craft go about this sort of analytic operation in different ways; there is little consensus as to the logic. What is even more troublesome, there are wild disagreements among analysts as to what they are willing to admit they hear in the utterances of one and the same language--even their own native language.

While it would be wrong to pass over these controversies in complete silence, we can stop now. Given care on one point, our formalism, as set forth in the first paragraph of this section (4.1.), stands ready to meet the empirical demands of whatever version of the componential approach emerges victorious.

The one point is that it would be awkward to have to talk about a component coalescing with nothing--that is, disappearing or appearing. Suppose we interpret English |p| as containing all the components present in |b|, plus "voicelessness". That is, we "zero out" the voicing of |b|, which is not at all to deny that |b| is voiced, but to choose to regard voicing simply as what one has except just when the "voicelessness" component is present.

A merger of |p| and |b| in some environment would then be difficult to describe within our formalism. The way to avoid this trouble is very simple. For other purposes we "zero out" as merrily as we please; but for the investigation of functional load (and perhaps, in general, for the discussion of linguistic change) we do not. English |p| and |b| may share a number of components, but each must be recognized as having a component missing from the other: |p| has voicelessness but not voicing; |b| has voicing but not voicelessness.

As a consequence, we must recognize a larger stock of components than may be necessary for certain other purposes. What is more, at least some components come in small sets that are mutually exclusive in occurrence, in that if one of such a set is present in a bundle, none of the others of the same set can be. A bundle cannot be both voiceless and voiced; in most languages, at least, it cannot be bilabial and at the same time apico-alveolar or dorso-velar. This is all commonplace, and makes no trouble. Indeed, for the present discussion we can now forget about it.

4.2. With $\underline{L}^{|\underline{t}|}$ defined as in Sec. 3.1., and $\underline{H}$ as in Sec. 2.2., it turns out that $\underline{H}(\underline{L}^{|\underline{t}|})$ is the entropy of a phonological system handled in terms of components and bundles, and that $\underline{H}(|a|,|b|)$ is the load carried by a particular pair of bundles of components.

Now, is there any simple relation between $\underline{H}(\underline{L}^{\underline{m}})$ and $\underline{H}(\underline{L}^{|\underline{t}|})$, assuming that we are dealing with the same language? Assuredly there should be. In fact, the two ought to be equal, since we hardly want the entropy of a phonological system to depend on the theoretical preferences of the analyst. What we can actually assert, however, is only as follows. Suppose a given language has $\underline{t} = \underline{r}$ distinct bundles of components, and that just these $\underline{r} = \underline{t}$ bundles are taken as the allophones of the system by someone who analyzes by the methods of traditional phonemics. From Theorem 1 (Sec. 3.3.) we know that $\underline{H}(\underline{L}^{[\underline{r}]}) = \underline{H}(\underline{L}^{\underline{m}})$. But clearly $\underline{H}(\underline{L}^{[\underline{r}]}) = \underline{H}(\underline{L}^{|\underline{t}|})$, since under the stated circumstances $\underline{L}^{[\underline{r}]} = \underline{L}^{|\underline{t}|}$. Hence, under just these conditions we know that $\underline{H}(\underline{L}^{\underline{m}}) = \underline{H}(\underline{L}^{|\underline{t}|})$.

This conclusion is a sort of Bill of Rights for the phonemicist. Of course, anyone can make mistakes of hearing or recording, and thus vitiate his results, but that is not the sort of thing we can deal with here. Setting this aside, our conclusion means that the phonemicist, as long as he operates within the constraints of assumptions A1 and A2 of Sec. 3.2. and takes simultaneous bundles of components as his allophones, can tinker with his data in any way he wishes, for any purpose he seeks (such as an elegant linear notation), with no fear that he is throwing information away.

4.3. The componential approach permits us to measure the functional load carried by a contrast between two components, either in a specific environment or in all environments, instead of only that carried by two whole bundles (or allophones).

When we want to measure the load carried by the contrast between two components in a specific environment, it is usually because there are two bundles |a| and |b|, which differ only in that |a| contains component $\underline{e}_1$ but not $\underline{e}_2$, while |b| contains $\underline{e}_2$ but not $\underline{e}_1$, and both of which occur in the same environment $\underline{E}$ (of preceding and/or following bundles). In this case, we define a "derived" system $\underline{L}'$ as identical with $\underline{L}^{|t|}$ except that in $\underline{L}'$ |a| and |b| have, in environment $\underline{E}$ (but not elsewhere), coalesced into a single bundle |ab|. In this particular environment, then, either $\underline{e}_1$ has been replaced by $\underline{e}_2$, or $\underline{e}_2$ has been replaced by $\underline{e}_1$, or both $\underline{e}_1$ and $\underline{e}_2$ have been replaced by some coalescent component $\underline{e}_1\underline{e}_2$ different from both. (It does not matter which of these is the case.) Then the very specific functional load being sought is $\underline{H}(\underline{L}^{|t|}) - \underline{H}(\underline{L}')$.

We might seek to determine the functional load carried, not by the contrast between single components in a given environment, but rather by the contrast between two sets of components in that environment, neither set necessarily being large enough to constitute a whole character. Let us say that |a| and |b| are the same except that |a| contains $\{\underline{e}_1, \underline{e}_2, \ldots, \underline{e}_q\}$ where |b| contains $\{\underline{e}'_1, \underline{e}'_2, \ldots, \underline{e}'_q\}$; we are to understand that, for $1 \leq \underline{i} \leq q$, $\underline{e}_i$ and $\underline{e}'_i$ cannot both be present in the same character. The definition and procedure are, of course,

exactly analogous to the preceding case. This would be the appropriate technique for determining the importance of the distinction of medial consonants in English matter and madder, and the like, present in British English but lost, as we noted earlier, from most American English dialects.

To measure the total load carried by a contrast between two components, we derive from $\underline{L}^{|\underline{t}|}$ a system $\underline{L}'$ in which the two components have coalesced in all environments, but nothing else has happened. Any pair of bundles which differ (in $\underline{L}^{|\underline{t}|}$) only in that one contains one of the components while the other has the other is thus coalesced into a single bundle in $\underline{L}'$. Then the desired load is $\underline{H}(\underline{L}^{|\underline{t}|}) - \underline{H}(\underline{L}')$. This would be the appropriate procedure, for example, for determining the importance of voicelessness versus voicing in English.

5. DISCUSSION

We have shown how functional load can be quantified in any of three different frames of reference: phonemic, allophonic, or componential. And we have described three interrelated measures, one based on $\underline{H}$, the entropy in binits, one based on $\hat{\underline{H}}$, the entropy in shannons, and one based on $\underline{C}$, the relative entropy.

A very small amount of empirical work has been devoted to the determination of the redundancy $\underline{R}$ of languages (usually in written rather than spoken form).[10] The redundancy is defined as $1 - \underline{C}$. If, in some system, every string of characters is a message, then the relative entropy is unity and the redundancy is zero.

---

[10] Some estimates for written English are given in C. E. Shannon, "The Mathematical Theory of Communication," in C. E. Shannon and W. Weaver, The Mathematical Theory of Communication, Urbana, 1949; see also Claude E. Shannon, "Prediction and Entropy of Printed English," Bell System Technical Journal, Vol. 30, pp. 50-64, 1951. I know of no printed data on spoken English, but a decade ago I attempted some determinations using phonemic transcription rather than standard orthography (and using test audiences familiar with the transcription); the results pointed towards a figure approximately the same (.50) as that for orthographic English. Clearly, little confidence should be placed in that figure; further empirical study is a desideratum.

In such a system, any change in a message between transmitter and receiver, brought about by channel noise, is an uncorrectable error. Since channel noise cannot be completely eliminated, redundancy plays a useful role. The empirical work referred to above suggests that the redundancy of natural languages hovers in the vicinity of .50. We shall use this figure instead of a completely arbitrary symbol, but pending further empirical study it should be taken as purely tentative.

One may now propose that, in the long run, the phonological system of the language of any community is governed by a law something like that that controls the behavior of a harmonic oscillator: if C deviates from its "neutral" value of .50, then there is a "restoring force" $\Phi = K(.50 - C)$, where K is a constant that presses C back towards the "neutral" value. The greater the displacement from neutral, the greater the restoring force. If Φ is positive, the redundancy is low and the relative entropy high: Φ tends to increase the former and decrease the latter. If Φ is negative, the redundancy is high and the relative entropy low; Φ tends to decrease the former and increase the latter.

This is perhaps more metaphor than mathematics, but let us see how it might work. Suppose, first, that the redundancy has become too low. Utterances are misunderstood oftener than usual. Ambiguous phrasings are therefore replaced or paraphrased by less ambiguous ones. For example, at that stage in the history of English when "Let him!" could be understood as a request either to leave him alone or to stop him, people began saying something like "Stop him!" if that was what they wanted done.[11] Also, people come to articulate more carefully.

---

[11] Leonard Bloomfield, Language, New York, 1933, p. 398. It is not implied, of course, that at that period in the history of English the overall redundancy had become too low. Indeed, perhaps it never does because perhaps adjustments in specific instances, such as the one cited in the text, are made too quickly for there to be any measurable diminution of redundancy for the whole language. This has to do with the magnitude of the constant K, discussed in the last paragraph of the paper.

But this really means the same thing, since typically a given sentence said rapidly and carelessly and the "same" sentence said slowly and carefully are not phonologically identical--the slow careful form retains stigmata of identity that are discarded in the rapid form. In general, then, utterances come to be distinguished from one another by larger numbers of occurrences of phonological units. This decreases the entropy and the relative entropy, and increases the redundancy.

Suppose, next, that the redundancy has become unnecessarily high. On the average, speakers are doing more work than necessary for intelligibility. Through laziness, "least effort," or whatever principle is actually involved here--clearly some principle of this sort is a reality--articulation becomes less careful. In such rapid careless speech, phonological units that are articulatorily similar can easily coalesce; and if there is little resort to slow-speech alternatives, then the fuller phonological structure of the slow-speech forms can be forgotten. In general, then, utterances come to be distinguished from one another by fewer occurrences of phonological units. This increases the entropy and the relative entropy, and decreases the redundancy.

Our "force" Φ, then, is actually the vector sum of two forces: one, which we might as well call "laziness," presses towards lower redundancy; the other, which is the practical need to be understood, presses towards lower relative entropy. Of course, both of these forces are statistical averages over whole communities of people and over many varied circumstances in which speech takes place--except in this gross statistical sense, we are not asserting that "people are naturally as lazy as they can be" or anything of the sort. At any one period, in any one community, the two forces have to operate via the actual linguistic system of the community, as it has been inherited, with all its arbitrary conventions. One could not venture, merely through the recognition of our two conflicting forces, to predict in any detail the near future of the language of any community. Even if one had considerably detailed information about the arbitrary conventions of the linguistic system, predictability would be severely limited, since so many different sorts of changes could equally well throw the two forces

out of balance, and so many different specific adjustments could restore the balance.

To complete the metaphor (if that is what it is), we may ask if anything might be empirically determinable about the constant K. If K is very small, then momentary deviations from balance--that is, from $\underline{R} = \underline{C} = .50$--might well be rather large. If K is large, then deviations are going to be small, and the restoration of balance is going to be more rapid. It may even be that K is an arbitrary constant, different from one language--or, perhaps, from one way of life--to another. It might even be that the balance point, which we have taken as .50, is different, say, between neolithic Polynesians and industrialized European-Americans. We have no information on these matters, but I see no reason why it could not be obtained if we want to obtain it.