TOWARD THE LOGICAL DESCRIPTION OF LANGUAGES IN THEIR PHONEMIC ASPECT

E. COLIN CHERRY
University of London

MORRIS HALLE
Massachusetts Institute of Technology

ROMAN JAKOBSON
Harvard University

Distinctive features occur in lumps or bundles, each one of which we call a phoneme. The speaker has been trained to make sound-producing movements in such a way that the phoneme-features will be present in the sound-waves, and he has been trained to respond only to these features and to ignore the rest of the gross acoustic mass that reaches his ears. LEONARD BLOOMFIELD (1933)

The number of different phonemes in a language is a small submultiple of the number of forms. LEONARD BLOOMFIELD (1926)

The logical demand that a science speak in quantitative terms is met by linguistics because it speaks in terms of phonemes. LEONARD BLOOMFIELD (1927)

1. INTRODUCTORY

This paper, an attempt to contribute to a logical description of the phonemic structure of a language, employs some of the elementary concepts of statistical communication theory.[1] Particular illustration is provided by a statistical analysis of colloquial Russian;[2] the material studied was the Russian urban conversations recorded by Peškovskij, comprising in the latter's phonetic transcription ten thousand sounds.[3]

In analyzing Russian or any other language, we must ascertain what and how many DISTINCTIVE FEATURES are needed to differentiate the meaningful units of its code, i.e. the smallest meaningful units, termed morphemes, and their combinations into words. Words are the maximum units that are expected to be entirely provided by the code. We must determine the minimum set of such features that the listener needs in order to recognize and distinguish all except homonymic morphemes, without help from context or situation. Once this set is determined, all other phonetic differences among morphemes or words of the given language can be shown to be predictable and therefore redundant.[4]

If we compare, for example, the Russian words (1) [bɨt] 'way of life', (2) [b,ɪt]

[1] See, in particular, C. E. Shannon and W. Weaver, *The mathematical theory of communication* (Urbana, 1949); D. M. Mackay, In search of basic symbols, *Cybernetics: Transactions of the eighth Conference* (New York, 1952); id., The nomenclature of information theory, ibid.

[2] This analysis was made as part of the research on contemporary Russian conducted by the Department of Slavic Languages and Literatures at Harvard University under a grant from the Rockefeller Foundation. The first volume of the description of contemporary Russian based on this research, dealing with the various aspects of Russian speech sounds, is now being prepared for publication. Grateful acknowledgment of help is also made to the Signal Corps, Air Materiel Command, Office of Naval Research, and the sponsors of the Fulbright program.

[3] A. Peškovskij, Desjat' tysjač zvukov russkogo jazyka, *Sbornik statej* 167–91 (Leningrad, 1952).

[4] For further information on distinctive features and their acoustic and articulatory correlates, see R. Jakobson, C. G. M. Fant, and M. Halle, *Preliminaries to speech analysis*, 2d printing (MIT Acoustics Laboratory, Technical report No. 13, 1952).

'beaten', (3) [bɨt,] 'be', and (4) [b,it,] 'beat',[5] we observe that words (1) and (2), or words (3) and (4), differ from each other in two respects: [ɪ] is farther forward than [ɨ] (i.e. has a higher second formant), and [i] is farther forward than [ɨ]; while [b,] is distinguished by its palatalization from [b]: it is produced with a flattening of the mouth cavity and a simultaneous widening of the pharyngeal channel which results in an upward displacement of energy along the frequency axis. Words (1) and (3), or words (2) and (4), also differ from each other in two respects: [ɨ] is closer than [ɪ], and [i] is closer than [ɪ]; while [t,] again differs from [t] in its palatalization.[6]

If, now, the presence as opposed to the absence of consonantal palatalization is viewed as the distinctive feature, this one feature (which we call sharp vs. plain) suffices to differentiate the four words in question. If, on the other hand, the distinctive function were to be assigned to the vowels, we would have to postulate two independent features, front vs. back and close vs. open. This superfluous multiplication of features is reason enough for rejecting the second solution. In Russian there is an additional reason for adopting the first solution, for in this language, in certain positions, the presence or absence of consonantal palatalization can by itself distinguish sequences otherwise identical, and must therefore in any case be considered an autonomous distinctive feature; compare [voʃt,] 'leader' with [koʃt] 'outlay', or [seļ,t,] 'herring' with [k,eļ,t] 'Kelt'. The distribution of advanced and retracted, on the other hand, as well as that of close and open varieties of vowels, is entirely predictable from the presence or absence of palatalized consonants before and after the vowel.

Proceeding consistently in this way, we find in the code of contemporary Standard Russian eleven distinctive features, grouped by superposition into forty-two phonemes.[7] These eleven distinctive features suffice to differentiate all but homonymic morphemes and words in Russian.

We leave aside here sound features that perform other functions, namely CONFIGURATIONAL features that signal the division of the utterance into grammatical units of different degrees of complexity, and expressive (or more precisely PHYSIOGNOMIC) features that signal solely the emotional attitudes of the speaker. Examples of configurational features signaling the division of the sound chain into word units: [dəvˈol,nɨj] /da vˈol,nij/ 'free besides' : [davˈol,nɨj] /davˈol,nij/ 'content'; [t,ˈẹ·n,ɨtˈam] /t,ˈen,i tam/ 'shadows are there' : [t,ˈẹn,ɪtˈam] /t,ˈe n,i-tˈam/ 'they are elsewhere'; [jixˈɨ·də jix,ˈi·də] /jˈix ˈida. jixˈida/ 'their Ida is

[5] Cf. A. Isačenko, *Fonetika spisovnej ruštiny* 177, 182 (Bratislava, 1947).

[6] We follow the IPA system of transcription, except in three respects: we use a comma after a letter to indicate palatalization; we place the accent mark immediately before the vowel letter; and we render the strident stop by the same letter as the corresponding constrictive with the addition of a circumflex.

[7] There are two competing varieties of contemporary standard Russian. The more conservative is codified especially in *Tolkovyj slovar' russkogo jazyka*, ed. by D. Ušakov (Moscow, 1935–40); the other is advocated in particular by S. Obnorskij, and is presented in *Slovar' russkogo jazyka*, ed. by S. Ožegov (Moscow, 1949). In general we accept Ušakov's norms; but in order to include all the phonemic discriminations possible in standard Russian, we add to his traditional repertory of phonemes a new phoneme /g,/ as distinguished from /g/. Such new gerund formations as /b,ir,ig,ˈa/ 'taking care', distinct from /b,ir,igˈa/ 'banks', are admitted into standard Russian by Obnorskij and his followers.

malicious'. Physiognomic features are illustrated in the different ways of pronouncing the word for 'yes' (simply [d'a] when unemphatic) according to the degree and kind of emphasis. These features convey subsidiary information similar to what is carried by such graphic equivalents of configurational features as spaces or punctuation marks, and such equivalents of physiognomic features as underlining or italicizing. The REDUNDANT features, on the other hand, operate in conjunction with the distinctive features, thereby facilitating the selective process on the part of the listener and lessening the burden on his attention.

For our computations, the text was split up into phoneme sequences consisting of two successive vowels and the consonants (if any) between them. In this way each vowel appears twice in our corpus, once as the initial and once as the final phoneme of a sequence. We chose these sequences 'from vowel to vowel' because phonemic conditioning is confined, in Russian, to consonantal clusters and to combinations of a vowel with preceding or following consonants; there is no apparent influence on consonants following a given vowel by those preceding it or vice versa. The compulsory syntactic pause (both initial and final) was denoted by a period and equated with a vowel.

Three sets of counts are of interest: (A) those that regard both the word boundaries (symbolized by a space) and the junctures between the immediate constituents of compound words[8] (symbolized by a hyphen); (B) those that regard only the word boundaries; and (C) those that regard neither the word boundaries nor the junctures, but break up a sequence only at the points of compulsory pause. The three ways of dividing a text into elementary sequences are illustrated in the accompanying table, based on the following passage: *Vot, na tebe na obed. Pojdëš'* ... /.v'ot. n'a t,ib,'e na-ab,'et. pa-jd,'oʃ./ 'Here, that's for your dinner. You'll go ...' The computations in this paper are made according to the first way of counting.

(A)	#v'o	'ot#	#n'a	'a#	#t,i	ib,'e	'e#	#na	a#	#a	ab,'e	'et#	#pa	a#	#jd,'o	'oʃ#
(B)	#v'o	'ot#	#n'a	'a#	#t,i	ib,'e	'e#	#na		aa	ab,'e	'et#	#pa		ajd,'o	'oʃ#
(C)	#v'o	'ot#	#n'a		'at,i	ib,'e		'ena		aa	ab,'e	'et#	#pa		ajd,'o	'oʃ#

2. THE FEATURE PATTERN AS A LOGICAL DESCRIPTION OF THE PHONEME

In the description that follows, language will be treated as a Markoff process.[9] The phonemes will be considered uniquely identifiable; but their order, in the sequences that compose our sample, can be described only statistically.

[8] Among Russian compound words we include all words with a non-initial root: words with more than one root, e.g. /adna-abr'aznij/ 'uniform'; words with prefixes, e.g. /za-astr,'it,/ 'to sharpen', /iz-vad,'it,/ 'to exhaust'; and words with preceding prepositions which are phonemically treated like prefixes, e.g. /za-akn'o/ 'behind the window', /iz-vad'i/ 'out of the water'.

[9] Cf. Shannon and Weaver 102: 'A system which produces a sequence of symbols ... according to certain probabilities is called a *stochastic process*, and the special case of a stochastic process in which the probabilities depend on the previous events, is called a Markoff process or a Markoff chain.' In his Essai d'une recherche statistique sur le texte du roman 'Eugène Onĕgin', illustrant la liaison des épreuves en chaine, *Bulletin de l'Académie Impériale des Sciences de St. Pétersbourg*, Vol. 7 (1913), A. A. Markov studied the distribution of vowel and consonant LETTERS in a part of Puškin's famous poem and showed that the transitional probabilities between the letters were not those of a random sequence but rather depended on the preceding letter or letters.

For the task of identifying one particular phoneme out of the set employed by the language, the distinctive features may be regarded as questions to be answered yes or no. Thus one may ask, Is the phoneme vocalic?—yes or no; Is the phoneme consonantal?—yes or no; and so on through the entire list of features. For the language under consideration here, a total of eleven such questions is necessary to identify any one phoneme uniquely. Table A illustrates these questions answered yes (+) or no (−); a zero (0) means either. This suggests that the logic is three-valued, a point that will be taken up again later.

A simple illustration of such a logical description is provided by Fig. 1, which shows a set of eight 'objects' A, B, ... H, to be identified by yes (+) or no (−) answers. Thus the group is first split in two, and we begin by asking, Is the object that we want on the right side (+) or not (−)? Successive subdivisions eventually identify any object in a set. If there are N objects in the set, and if N happens to be a power of 2, the number of yes-or-no answers necessary to identify each of the objects in the set is $\log_2 N$. The complete identification of any object is then a chain of plus and minus signs; thus, the object G in Fig. 1 is identified by the chain (+ + −).

A	B	C	D	E	F	G	H
−	−	−	−	+	+	+	+
−	−	+	+	−	−	+	+
−	+	−	+	−	+	−	+

FIG. 1. THE LOGICAL IDENTIFICATION OF OBJECTS IN A SET OF EIGHT

Even when N is not a power of 2, the quantity $\log_2 N$ can still be used as a measure. In such cases the fractional result must not be taken to imply a fraction of a question; it means, rather, that the N members of the set will not all necessarily require the same number of answers for identification. The fraction results from averaging.

The quantity $\log_2 N$ is conventionally expressed in BITS; the name for this unit is derived from BINARY DIGIT (i.e. yes-or-no choice).

In Fig. 1 the successive subdivision has been consistently into two equal subgroups; this method results in identification by the smallest possible number of answers, and so in the shortest chain of plus and minus signs. Subdivision into unequal subgroups requires, on the average, more questions and answers.

Let us now apply this process to the list of forty-two Russian phonemes listed in Table A. But first consider a purely hypothetical description of any one phoneme out of the forty-two, as though these were not phonemes but merely objects without linguistic significance. If they were successively subdivided as in Fig. 1, the description of any one object would require $\log_2 42$ questions, on the average, or 5.38 bits per phoneme. In our analysis of language we are concerned, however, not only with questions of logic but also with matters of fact; hence the answers yes or no in Table A are provided for us by considerations of the natural process of speaking.

One might ask, Why cannot a type of feature pattern be invented which

employs only 5.38 questions per phoneme, on the average, in a manner analogous to the hypothetical case discussed? This could perhaps be done; but the distinctive features used at present (Table A) serve other purposes and are intimately related to the physical production of speech. They number eleven, implying an average of 5.62 extra questions per phoneme (11 − 5.38). This means that redundant or extra plus and minus signs are brought in. Nevertheless these features, as they have been proposed for earlier linguistic analyses, fit into the logical

	k	k,	g	g,	x	c	ʃ	ʒ	t	t,	d	d,	s	s,	z	z,	ŝ	n	n,	p	p,
VOCALIC	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−
CONSONANTAL	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
COMPACT	+	+	+	+	+	+	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−
DIFFUSE	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GRAVE	+	+	+	+	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	+	+
NASAL	0	0	0	0	0	0	0	0	−	−	−	−	−	−	−	−	−	+	+	−	−
CONTINUANT	−	−	−	−	+	−	+	+	−	−	−	−	+	+	+	+	−	0	0	−	−
VOICED	−	−	+	+	0	0	−	+	−	−	+	+	−	−	+	+	−	−	−	−	−
SHARP	−	+	−	+	0	0	0	0	−	+	−	+	−	+	−	+	0	−	+	−	+
STRIDENT	0	0	0	0	0	0	0	0	−	−	−	−	0	0	0	0	+	0	0	0	0
STRESSED	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

	b	b,	f	f,	v	v,	m	m,	ˈu	u	ˈo	ˈe	ˈi	i	ˈa	a	r	r,	l	l,	j
VOCALIC	−	−	−	−	−	−	−	−	+	+	+	+	+	+	+	+	+	+	+	+	−
CONSONANTAL	+	+	+	+	+	+	+	+	−	−	−	−	−	−	−	−	+	+	+	+	−
COMPACT	−	−	−	−	−	−	−	−	−	−	−	−	−	−	+	+	0	0	0	0	0
DIFFUSE	0	0	0	0	0	0	0	0	+	+	−	−	+	+	0	0	0	0	0	0	0
GRAVE	+	+	+	+	+	+	+	+	+	+	+	−	−	−	0	0	0	0	0	0	0
NASAL	−	−	−	−	−	−	+	+	0	0	0	0	0	0	0	0	0	0	0	0	0
CONTINUANT	−	−	+	+	+	+	−	−	0	0	0	0	0	0	0	0	−	−	+	+	0
VOICED	+	+	−	−	+	+	−	−	0	0	0	0	0	0	0	0	0	0	0	0	0
SHARP	−	+	−	+	−	+	−	+	0	0	0	0	0	0	0	0	−	+	−	+	0
STRIDENT	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
STRESSED	0	0	0	0	0	0	0	0	+	−	0	0	+	−	+	−	0	0	0	0	0

TABLE A. THE PHONEMES OF RUSSIAN
showing their distinctive feature patterns as represented by the answers yes (+), no (−), either (0)

descriptive system, though apparently with some inefficiency. Can the efficiency of our empirical description be improved by simplification of Table A?

Table A shows the simplest possible description of the 42 phonemes in terms of the given eleven features. There are several points of difference between this table and Fig. 1. First, the successive questions have phonetic significance; they do not merely ask Right or Left? like those in Fig. 1. The answer to the first question (vocalic—yes or no?) does not split the 42 phonemes into two equal groups, but into 12 pluses and 30 minuses; Russian phonemes simply happen to have this characteristic. The second question (consonantal—yes or no?) again parts each of these groups into unequal subgroups; and so on.

Moreover, some of the questions in the list need not be answered at all for

particular phonemes, because the identification is complete without them. In Table A we use a zero to indicate 'either'—that is, either plus or minus. For example, the phoneme /t/ is represented by the chain (−+−0−−−−−−0). Each of the zeros can be replaced by either plus or minus without affecting the identification; in either case, the chain of symbols for /t/ remains unique. Since every zero may thus be regarded as either a plus or a minus, the total number of questions answered here is eleven per phoneme. This is a measure of the 'information' conveyed when the speaker selects any particular phoneme out of the 42, at least on the basis of the feature pattern here presented. But as we have seen, the true 'information' is rather to be expressed by an average of 5.38 questions (bits); the extra 5.62 bits represents the redundancy that would result from the replacement of the zeros by plus or minus signs. (It must be emphasized that our measure of 'information' has up to this point been based upon the assumption that all 42 phonemes have an equal probability of occurrence and that they are wholly independent units. Since language has, of course, a much more complex structure than this, our definition of 'information' will later have to be modified.)

The term 'redundancy' should not be taken to imply wastefulness; it is a property of speech, and in fact of every system of communication, which serves a most useful purpose. In particular, it helps the hearer to resolve uncertainties introduced by distortion of the signal or by disturbing noises. For example, the feature of nasality is marked 0 for all vowels. If these zeros were changed to pluses, the new symbols would not imply that a Russian speaker always nasalizes his vowels: normally he does not; but even if he did, the nasality would have no phonemic significance. In some cases a zero appears in a place where the substitution of plus or minus would imply an impossible articulation; but even here the point is that the phoneme is uniquely identified without this feature.

If the data given in Table A can be recast so as to eliminate the necessity of using the ambiguous symbol 0, then the number of questions needed to identify any one phoneme will, on the average, be reduced. That is, the description of phonemes in terms of features will be less redundant.

3. Removal of the Ambiguous Zero Signs

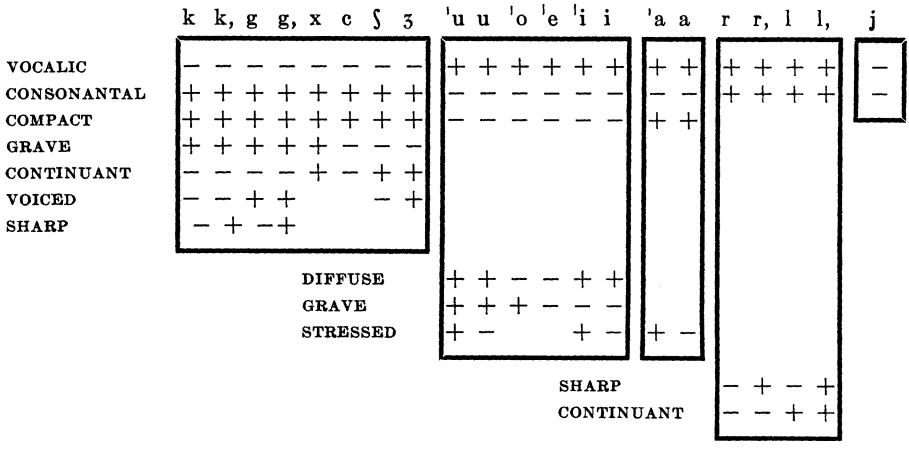
One might suppose that by re-ordering the feature questions, it would be possible to remove all the zero signs in Table A, or at least to shift them to the end of every phoneme column so that they could be omitted (the phoneme being identified then by the chain of plus and minus signs only). It turns out, however, that this cannot be accomplished by any simple re-ordering.

The whole problem may be changed by regarding the table of signs (+, −, 0) as a code book for identifying the various phonemes. In this view there is no reason why the order of the feature questions should not be different for different phonemes. In fact, the order could change during the identification of a particular phoneme, at certain stages depending upon the answers to earlier questions. Thus a sequence of different code books would be required. Table B shows the result of such a recoding.

As an example, consider the identification of the phoneme /ˈo/. The answers to the questions Vocalic? Consonantal? Compact? are respectively +−−,

which identifies the phoneme as belonging to the group /ˈu u ˈo ˈe ˈi i/. This requires that a new code book be used for the subsequent questions. These, as we see from Table B, are asked in the order Diffuse? Grave? Stressed? The code books are known a priori and represent here the independent phoneme structure of Russian; they themselves contain the 'information' provided by the zeros in Table A.

This process of recoding may be regarded as a TRANSFORMATION. The number of signs (bits) required to identify any phoneme uniquely is now less than before



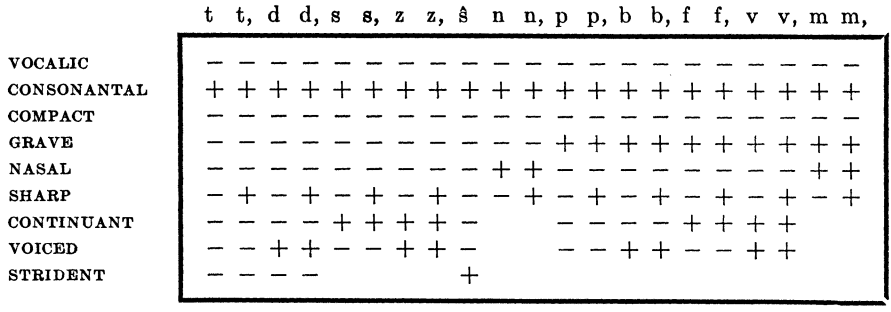


TABLE B. THE PHONEMES OF RUSSIAN
re-ordered to eliminate the ambiguous zero

by the number of zeros eliminated from Table A. Although it is different for different phonemes, on the average it is 6.5 bits per phoneme, a value considerably lower than our original 11 and nearer to the ideal value of 5.38. The description in terms of features has thus been made efficient.

4. CONSIDERATIONS OF PHONEME PROBABILITIES

(a) *Individual Frequencies of Occurrence*

The next step in our description of the language will be to consider the relative frequencies of the individual phonemes. The 'information' in bits per phoneme obtained previously has the hypothetical minimum value of 5.38 ($\log_2 42$), a

result obtained by successively subdividing the set of phonemes into two equal groups. When their frequencies of occurrence are unequal, however, the required average (bits per phoneme) is obtained by successively subdividing the set into two groups of equal total probability. The result then is that the average number of questions needed to identify a phoneme (in terms of bits per phoneme) is H_1, where

$$H_1 = -\sum p_i \log p_i \tag{1}$$

summed over all phonemes i. This is the 'expected' value of $-\log p_i$. (Remember that p_i is always less than 1).

a	*b*	*c*	*d*	*e*	*a*	*b*	*c*	*d*	*e*
a	1316	2.94	.387	4	d	177	5.81	.100	9
i	977	3.35	.328	6	l,	162	5.95	.096	4
t	602	4.05	.244	9	'u	153	5.96	.091	6
'a	539	4.23	.228	4	r,	133	6.20	.083	4
j	457	4.45	.202	2	z	130	6.25	.081	8
n	392	4.66	.183	6	d,	126	6.30	.080	9
'o	379	4.72	.179	5	b	119	6.39	.075	8
s	359	4.80	.172	8	x	102	6.60	.067	5
'e	343	4.86	.167	5	g	91	6.80	.062	7
k	284	5.14	.146	7	v,	89	6.84	.061	8
v	273	5.15	.140	8	ʒ	89	6.84	.061	6
'i	243	5.38	.131	6	f	85	6.86	.058	8
u	240	5.40	.129	6	s,	85	6.86	.058	8
p	232	5.42	.126	8	ŝ	59	7.40	.044	9
r	230	5.45	.125	4	m,	56	7.50	.043	6
n,	221	5.50	.121	6	b,	52	7.60	.039	8
l	212	5.55	.118	4	p,	50	7.64	.038	8
ʃ	207	5.56	.115	6	k,	36	8.10	.029	7
m	202	5.64	.114	6	z,	21	8.90	.018	8
c	197	5.65	.111	5	f,	8	10.30	.008	8
t,	196	5.65	.111	9	g,	7	10.50	.008	7

TABLE C

a = Phoneme (i); $b = p_i \times 10^4$; $c = -\log_2 p_i$; $d = -p_i \log_2 p_i$; *e* = number of features listed in Table B (i means 'any given phoneme'; p_i means 'the probability of a given phoneme')

The relative frequencies of the individual Russian phonemes have been counted from samples of the language, as described in Section 1; they are listed in Table C. From these frequencies p_i we may readily calculate the hypothetical 'information' H_1 given by (1). This is

$$H_1 = 4.78 \text{ bits/phoneme} \tag{2}$$

On the other hand we may calculate the average number of features, i.e. the binary choices per phoneme, knowing the probabilities p_i of the individual phonemes. If N_i is the number of features required to identify the *i*th phoneme in Table B, then the rate of feature choices which actually occurs is

$$\sum N_i p_i = 5.79 \text{ features/phoneme} \tag{3}$$

which may be compared to the ideal given by (2).

In a recent article,[10] Huffman has described a method for devising the most efficient code possible for a set of independent messages of known frequency distribution. In such a code 'the average number of coding digits per message is minimized'. If we regard the phonemes of our language as independent messages, we can apply Huffman's method and compute from the probabilities given in Table C the number of digits which in an optimal code would be necessary to identify each phoneme uniquely. This can be compared to the number of features necessary to identify each phoneme in Table B. It must be pointed out, however, that these are not strictly comparable: as we stated in Section 3, the description in terms of distinctive features presupposes that the digits are interpreted differently depending on the answers given in a preceding stage of the analysis, while in Huffman's code all digits have the same interpretation. In the following table we compare the number of phonemes having a given number of digits in the optimal code with the number of phonemes having the same number of distinctive features in Table B.

NUMBER OF DIGITS OR DISTINCTIVE FEATURES	2	3	4	5	6	7	8	9	10
In an optimal code	0	2	2	11	13	8	3	1	2
In the actual case (Table B)	1	0	6	4	10	4	12	5	0

Regarded purely as a descriptive process, then, the method of listing the distinctive features is rather efficient.

So far we have been regarding the phonemes of the language as independent. But the natural process of speech consists not merely of choosing a chain of independent phonemes; at the very least it consists of a succession of choices, where each choice is in part conditioned by the preceding phoneme chosen. It may be a truer description of the natural process of speech to say that phonemes are chosen in groups. Thus the simple analysis that we have made so far must be regarded as a somewhat artificial though quite efficient description of the language in its simplest aspect.

Before concluding this section on individual phonemes, it may be of interest to note a few statistical facts gathered from Table C.

Probability of a vowel occurring = 0.4190; of a liquid = .0737; of a glide /j/ = 0.0457; of a consonant proper = 0.4616.[11]

In the accompanying table, the plus and minus probabilities of each feature were calculated by adding the probabilities of all phonemes showing a plus for that feature in Table B and of all those showing a minus. Thus the probability

[10] David A. Huffman, A method for the construction of minimum redundancy codes, *Proceedings of the IRE* 40:9.1098–101 (1952).

[11] Markov, in his study of LETTER distributions in a Russian poem, obtained the value 0.4317 for vowels and 0.5683 for consonants. His figures are remarkably close to ours, especially if we make allowance for the fact that Markov counted some instances of /j/ as vowels and others not at all.

of a yes-answer to the question Voiced? is the sum of the probabilities of /g–g,–ʒ–d–d,–z–z,–b–b,–v–v,/, while the probability of a no-answer is the sum of the probabilities of /k–k,–ʃ–t–t,–s–s,–ŝ–p–p,–f–f,/. (We omit the data concerning nasality, stridency, compactness, and diffuseness; for here the pluses are much fewer than the minuses, and the lower probability of the former is obvious.)

	Probability of +	Probability of −
Voiced	.1174	.1920
Sharp	.1242	.3445
Stressed (vowels only)	.0935	.2533
Continuant	.1822	.2530
Grave (vowels)	.0772	.1563
Grave (consonants)	.1684	.2861
Totals	.2456	.4424

These figures are significant, especially since the pluses and minuses were assigned without considering their relative frequency, entirely on the basis of an examination of the features and their interrelations.[12]

But the phonemic structure of a language is not defined entirely by the total probabilities of feature occurrence; their distribution in time is also significant. These distributions measure what might be termed the continuity of each feature; they can be obtained from the analysis of joint probabilities presented below. Thus, if we know the probabilities p(a b c $\cdots$ n) of various chains of n phonemes, we can readily assess the probability that a certain distinctive feature exists uninterrupted for a duration greater than m phonemes, where m = 1, 2, $\cdots$ n. It is not our purpose here to execute such an analysis in detail, but rather to point out its potentialities as a basis for language description.

(b) *Phoneme Groups, Syllables*

In the preceding section we paid attention mainly to what may be termed phonemic monograms—that is, to individual phonemes, with some reference also to phoneme groups and to their joint probabilities of occurrence. These groups may be digrams, trigrams, and so on. Another type of probability which is of interest to the student of language structure is the TRANSITION PROBABILITY that a particular phoneme will follow a given phoneme or phoneme group. Thus, if p(a b $\cdots$ n) is the probability of the phoneme group (a b $\cdots\cdot$ n), then

$$\begin{aligned} p(a\ b \cdots n) &= p(a)p_a(bc \cdots n) \\ &= p(a)p_a(b)p_{ab}(cd \cdots n) \\ &= p(a)p_a(b)p_{ab}(c)p_{abc}(d \cdots n), \text{ etc.} \end{aligned} \tag{4}$$

In this way the joint probability of a group is related to the transition probabilities of the successive phonemes a, b, c, etc. occurring in the group.

Given a particular phoneme (a) of a language, or a possible group of phonemes

[12] For a fuller discussion, see R. Jakobson, *Sound and meaning* (to appear).

(ab ··· n), the phonemes (m) which can occur next in the chain have a set of probabilities $p_{ab\ldots n}(m)$. The fact that these probabilities vary according to the character of m implies that a certain degree of prediction is possible. This property provides another form of 'redundancy' in the language, a quality which is of great importance in aural recognition, as when we follow a conversation in a noisy room.

For instance, if one hears a palatalized /v,/ in a Russian utterance, one can be sure that no unstressed vowel except /i/ will follow. After a palatalized /b,/, the probability of an unstressed /a/ is extremely low; the sequence /b,a/, as in /gˈolub,a/ 'pigeon' (gen.-acc. sing.) and /galˈub,a/ 'fondling', is exceptional. In our count we have found the following phonemes after palatalized /s,/, with the indicated frequencies:

i	33	ˈa	6	u	1
ˈe	16	m	2	j	1
ˈo	9	ˈu	1	a	1
ˈi	8				

Note especially the almost complete absence of consonants and the very low frequency of unaccented /a/. On the other hand, after nonpalatalized /s/ the unaccented /a/ was the most frequent of all the vowels in our material, and consonants occurred very freely. Our figures for phonemes after /s/ are these:

t	76	p	9	k,	3
a	37	u	6	j	2
t,	30	v	6	ˈi	1
k	27	i	5	m,	1
l	20	m	5	r	1
ˈa	16	n,	5	ˈu	1
ˈo	11	p	5	v,	1
l,	10	x	3	r,	1
n	10				

Since the inequality of the transition probabilities makes possible a certain degree of prediction, the information conveyed by one phoneme in the chain of connected speech is less than that conveyed by one phoneme in isolation. Unless it is the first in the chain, we know something about it, so to speak, before it arrives. This information can be strictly defined, in the technical sense of the earlier sections; we can even derive formulae, analogous to equation (1), which will be applicable to connected groups of phonemes. Suppose, for example, that we have computed the probabilities p(ab) of all the phoneme digrams of a language; then the information conveyed by any digram of the language is, on the average, $H_{1,2}$:

$$H_{1,2} = -\sum p(ab) \log p(ab) \text{ bits/digram} \tag{5}$$

Similarly for trigrams:

$$H_{1,2,3} = -\sum p(abc) \log p(abc) \text{ bits/trigram} \tag{6}$$

But if, instead, we have computed the various transition probabilities $p_a(b)$, the information conveyed by the occurrence of each successive phoneme is $H_1(2)$:

$$H_1(2) = -\sum p(ab) \log p_a(b) \tag{7}$$

Again, if we know the transition probabilities $p_{ab}(c)$:

$$H_{1,2}(3) = -\sum p(abc) \log p_{a,b}(c) \tag{8}$$

Clearly these various information rates, based on different probability tables, are connected. To show this, consider equation (4); take logs of both sides and then average over all possible groups $(ab \cdots n)$:

$$-\sum p(ab \cdots n) \log p(ab \cdots n) =$$

$$-\sum p(ab \cdots n) [\log p(a) + \log p_a(b) + \log p_{ab}(c) \cdots] \text{ or}$$

$$H_n = H_1 + H_1(2) + H_{1,2}(3) + H_{1.2,3}(4) \cdots \text{ bits/n-gram} \tag{9}$$

This means that the information conveyed by groups of phonemes is, on the average, equal to the sum of the information obtained from each successive phoneme.

We have computed the values for the digrams and trigrams in our material according to the first count—the one that takes account of the boundaries between words and between the parts of compounds. The values were found to be 8.45 bits/digram and 9.15 bits/trigram. If the phonemes were independent, the corresponding values would be 9.54 bits/digram and 14.31 bits/trigram. As expected, the values are lower when the units in the chain are not regarded as independent.

Another very promising approach, which for the present must remain unexplored, is to calculate the distributions of the distinctive features in time, as already proposed in Section 4(a). Given a long sample of text transcribed phonemically, we write under each symbol a column of pluses, minuses, and zeros representing its distinctive features in some regular order (as in Table A). The horizontal sequences of pluses, minuses, and zeros produced in this way can then be used to measure the 'continuity' of the various features. The probabilities of such sequences may be written $p_+(m)$, $p_-(m)$, $p_0(m)$, where $m = 1, 2, 3$, etc. It is obvious that such distributions may provide a basis for statistical specification of the phonemic differences between one language and another.

The statistical analysis of the phonemes and their sequences in connected messages must be supplemented by a similar analysis of the dictionary, in order to understand the distribution of phonemes in the lexical code of the given language.[13] The comparison of the two sets of data is certain to be most instruc-

[13] In R. Carnap's terminology, the occurrences of phonemes, having been studied in the Russian word-EVENTS, are to be investigated in the word-DESIGNS, just as we have here studied the occurrences of distinctive features in the phoneme-DESIGNS; cf. *Introduction to semantics* 3 (Cambridge, Mass., 1946). Charles S. Peirce, the founder of modern semiotic, would say that besides the application of the phonemic LEGISIGNS within the lexical SINSIGNS, such an application must be scrutinized again within lexical LEGISIGNS; cf. his *Collected papers* 2.245–7 (Cambridge, Mass., 1932).

tive. The statistical analysis of the dictionary permits us to draw conclusions about the phoneme sequences peculiar to different types of morphemes and to words of different grammatical categories.[14] Furthermore, it forms the basis for definitive statements about phoneme combinations with probabilities of 1 and 0; for no phoneme sequence can occur in messages if it is not provided by the code.

Finally, among problems which remain to be investigated are those transitional probabilities which operate backwards, i.e. which depend not on earlier but on subsequent events, or, in linguistic terms, not on progressive but on regressive action of phonemes in a sequence. The comparison of these two sets of statistics is very important, because it is obvious that for different types of sequence the predictability is greater in one direction than in the other. Analysis of such data will provide the most solid basis for setting up a statistical model of the syllable as a recurrent link in the chain of speech.

[14] An exhaustive statistical analysis of the phonemic structure of Russian root morphemes is being prepared by Robert Abernathy within the framework of the research program mentioned in fn. 2.