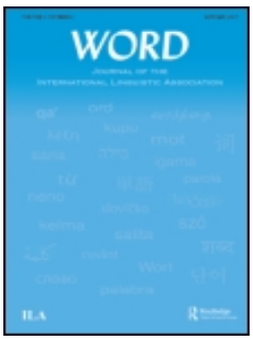

WORD

ISSN: 0043-7956 (Print) 2373-5112 (Online) Journal homepage: http://www.tandfonline.com/loi/rwrd20

# Distributional Structure

Zellig S. Harris

**To cite this article:** Zellig S. Harris (1954) Distributional Structure, WORD, 10:2-3, 146-162, DOI: 10.1080/00437956.1954.11659520

**To link to this article:** http://dx.doi.org/10.1080/00437956.1954.11659520

# DISTRIBUTIONAL STRUCTURE

ZELLIG S. HARRIS

**1. Does language have a distributional structure?** For the purposes of the present discussion, the term structure will be used in the following non-rigorous sense: A set of phonemes or a set of data is structured in respect to some feature, to the extent that we can form in terms of that feature some organized system of statements which describes the members of the set and their interrelations (at least up to some limit of complexity). In this sense, language can be structured in respect to various independent features. And whether it is structured (to more than a trivial extent) in respect to, say, regular historical change, social intercourse, meaning, or distribution—or to what extent it is structured in any of these respects—is a matter decidable by investigation. Here we will discuss how each language can be described in terms of a distributional structure, i.e. in terms of the occurrence of parts (ultimately sounds) relative to other parts, and how this description is complete without intrusion of other features such as history or meaning. It goes without saying that other studies of language—historical, psychological, etc.—are also possible, both in relation to distributional structure and independently of it.

The distribution of an element will be understood as the sum of all its environments. An environment of an element A is an existing array of its co-occurrents, i.e. the other elements, each in a particular position, with which A occurs to yield an utterance. A's co-occurrents in a particular position are called its selection for that position.

**1.1. Possibilities of structure for the distributional facts.**

To see that there can be a distributional structure we note the following: First, the parts of a language do not occur arbitrarily relative to each other: each element occurs in certain positions relative to certain other elements. The perennial man in the street believes that when he speaks he freely puts together whatever elements have the meanings he intends; but he does so only by choosing members of those classes that regularly occur together, and in the order in which these classes occur.

Second, the restricted distribution of classes persists for all their occurrences; the restrictions are not disregarded arbitrarily, e.g. for semantic needs. Some logicians, for example, have considered that an exact distributional description of natural languages is impossible because of their inherent vagueness. This is not quite the case. All elements in a language can be grouped into classes whose relative occurrence can be stated exactly. However, for the occurrence of a particular member of one class relative to a particular member of another class it would be necessary to speak in terms of probability, based on the frequency of that occurrence in a sample.

Third, it is possible to state the occurrence of any element relative to any other element, to the degree of exactness indicated above, so that distributional state-

ments can cover all the material of a language, without requiring support from other types of information. At various times it has been thought that one could only state the normative rules of grammar (e.g. because colloquial departures from these were irregular), or the rules for a standard dialect but not for "substandard" speech or slang; or that distributional statements had to be amplified by historical derivation (e.g. because the earlier form of the language was somehow more regular). However, in all dialects studied it has been possible to find elements having regularities of occurrence; and while historical derivation can be studied both independently and in relation to the distribution of elements,[1] it is always also possible to state the relative occurrence of elements without reference to their history (i.e. "descriptively").

Fourth, the restrictions on relative occurrence of each element are described most simply by a network of interrelated statements, certain of them being put in terms of the results of certain others, rather than by a simple measurement of the total restriction on each element separately. Some engineers and mathematicians (as also phoneticians and experimental psychologists) who have become interested in language have sought a direct formulation of the total restrictions on occurrence for each element, say for each sound.[2] This would yield an expression for how the occurrences of each element depart from equiprobability, and so would give a complete description of the occurrences of elements in the language. Now it is of course possible to enumerate the relative occurrences of a finite set of elements in finitely long utterances; but direct enumeration is of little interest because it yields no simple description of the over-all occurrences of elements, and because it does not order the restrictions in such a way that the larger restrictions get stated before the smaller ones. In contrast with this, it is possible to describe the occurrence of each element indirectly, by successive groupings into sets, in such a way that the total statements about the groupings of elements into sets and the relative occurrence of the sets are fewer and simpler than the total statements about the relative occurrence of each element directly.

We obtain then an ordered set of statements in terms of certain constructs—the sets at successive levels. Since the ordering of statements can be arranged so that the earlier ones will deal with the more inclusive sets, we can stop the process of setting up these statements at any convenient point, and accept the unfinished list of statements as an approximation to the distributional facts—knowing that the subsequent statements will only make subsidiary corrections to the earlier

[1] The investigation of historical regularity without direct regard to descriptive (synchronic) structure was the major achievement of the linguists of the late eighteen hundreds. There are incipient studies of historical-descriptive interrelations, as in H. M. Hoenigswald, Sound Change and Linguistic Structure, *Language* 22(1946). 138–43; cf. A. G. Juillald, A Bibliography of Diachronic Phonemics, *Word* 9(1953). 198–208. The independent study of descriptive structure was clarified largely by Ferdinand de Saussure's *Cours de linguistique générale*, the Prague Circle in its *Travaux du Cercle linguistique de Prague*, Edward Sapir in various writings, and Leonard Bloomfield's *Language*.

[2] These approaches are discussed by Martin Joos, Description of Language Design, *Journal of the Acoustical Society of America* 22(1950).702–8, and W. F. Twaddell, *ibid.* 24(1952).607–11.

statements. (This is not the case for the direct enumeration of restrictions, where the restrictions to be enumerated after a given point may be greater than those enumerated before.)

In view of this we may say that there is not only a body of facts about the relative occurrence of elements in a language, but also a structure of relative occurrence (i.e. of distribution). Hence the investigation of a language entails not only the empirical discovery of what are its irreducible elements and their relative occurrence, but also the mathematical search for a simple set of ordered statements that will express the empirical facts.[3] It may turn out that several systems of statements are equally adequate, for example several phonemic solutions for a particular language (or only, say, for the long vowels of a language). It may also be that different systems are simpler under different conditions. For example, one system may be adequate in terms of successive segments of sound (with at most stress and tone abstracted), while another system may be simpler if we admit the analysis of the sounds into simultaneous components of varying lengths. Or one system of stating distribution in respect to near neighbors (the usual environment for phonemic solutions) may be simple by itself, but if we are to imbed it in other statements about farther neighbors we may find that when we choose a modified system the statements covering the imbedding are simpler (i.e. a different phonemic solution may be more convenient for use in statements about morphemes). If the distributional structure is to be used as part of a description of speech, of linguistic behavior, then we will of course accept only such structures as retain a passably simple relation to the phonetic features. But for some other purpose, such as transmission or systemic analysis, phonetic complexity may be no serious objection. In any case, there is no harm in all this non-uniqueness,[4] since each system can be mapped onto the others, so long as any special conditions are explicit and measurable.

Various questions are raised by the fact that there can be more than one (non-trivial) structural statement for a given language. Can we say whether a particular item of structural analysis contributes to the simplicity of the system? It may be possible to do this: For example, if a given analysis involves a particular classification of elements (say, verbs), we may try some variation on this classification (say, by subdivision into transitive and intransitive—distributionally defined) and see whether the resulting analysis is simpler or not. Can we say what is invariant under all the possible distributional structures for a given body of data? For example, for all the phonemic solutions in a given language, there remains constant the minimal network of phonemically distinct utterance-pairs in terms of which we can distinguish every phonemically distinct utterance.

The various structural systems considered here all have this in common, that

[3] For a discussion of simplicity in this connection, see a forthcoming article by Noam Chomsky, Some Comments on Simplicity and the Form of Grammars.

[4] Y. R. Chao, The Non-uniqueness of Phonemic Solutions of Phonetic Systems, *Bulletin of the Institute of History and Philology, Academia Sinica* 4(1934).363–98. Cf. the two solutions of Annamese phonemes in M. B. Emeneau, *Studies in Vietnamese (Annamese) Grammar* 9–22.

they list items and their occurrences. There is at least one other type of structural statement which is essentially distributional but couched in different terms. This is the style which describes one linguistic form as being derived by some process (operation) from another. The item style says: Form *A* includes elements $e + f$ while form *B* includes elements $e + g$; and thus it describes all forms as combinations of elements. The process style says: Form *A* is derived from *B* by changing *f* into *g;* and thus it describes most forms as derived from certain base forms. The combinatorial or item style, which has a more algebraic form, is more parsimonious and representative for much of linguistic data. The process style, which is more similar to historical statements, is useful in certain situations, especially in compact morphophonemics.[5] Both styles are based solely on the relative occurence of parts, and are therefore distributional.

**1.2. Reality of the structure.**

Some question has been raised as to the reality of this structure. Does it really exist, or is it just a mathematical creation of the investigator's? Skirting the philosophical difficulties of this problem, we should in any case realize that there are two quite different questions here. One: Does the structure really exist in the language? The answer is yes, as much as any scientific structure really obtains in the data which it describes: the scientific structure states a network of relations, and these relations really hold in the data investigated.[5a]

Two: Does the structure really exist in the speakers? Here we are faced with a question of fact, which is not directly or fully investigated in the process of determining the distributional structure. Clearly, certain behaviors of the speakers indicate perception along the lines of the distributional structure: for example, the fact that while people imitate non-linguistic or foreign-language sounds, they "repeat" utterances of their own language[6] (i.e. they reproduce the utterance by substituting, for the sounds they heard, the particular corresponding variants which they habitually pronounce; hence the heard sounds are perceived as members of correspondence sets). There are also evidences of perception of sounds in terms of their morphophonemic memberships.[7]

A reasonable expectation is that the distributional structure should exist in the speakers in the sense of reflecting their speaking habits.[8] Indeed, responses

[5] This kind of formulation is best expressed in the work of Sapir and Newman; cf. reviews of *Selected Writings of Edward Sapir* (D. Mandelbaum, ed.), in *Language* 27(1951). 289–92; and of Stanley Newman, *Yokuts Language of California* in *International Journal of American Linguistics* 10(1944).196–211.

[5a] An opposition has sometimes been claimed between real facts and mathematical manipulation of structure. This claim ignores the fact that science is (among other things) a process of indicating much data by few general statements, and that mathematical methods are often useful in achieving this. Mathematical and other methods of arranging data are not a game but essential parts of the activity of science.

[6] As pointed out by Kurt Goldstein, *Language and language disturbances* 71, 103.

[7] E.g. in Edward Sapir, La réalité psychologique des phonèmes, *Journal de Psychologie Normale et Pathologique* 30(1933).247–65 (translated in David Mandelbaum, ed., *Selected Writings of Edward Sapir* 46–60).

[8] C. F. Hockett, review of *Recherches Structurales* in *International Journal of American Linguistics* 18(1952).98.

along the lines of distributional structure can be found in experimental psychology work.[9] However, different speakers differ in the details of distributional perception. One speaker may associate the stem of *nation* with that of *native*, while another may not: should the morpheme analysis be different for the two idiolects (individual dialects)? Even if we take the speaking habits to be some kind of social summation over the behaviors (and habits) of all the individuals, we may not find it possible to discover all these habits except by investigating the very speech events which we had hoped to correlate with the (independently discovered) habits.

If, as Hockett proposes, we measure the habits by the new utterances which had not been used in the structural description, we have indeed a possible and sensible measure; and this applies both to real productivity (the use of elements in environments in which they had not occurred before), and also to arbitrarily unused data (utterances which may have occurred before but which had not been used in deriving the distributional structure). However, even when our structure can predict new utterances, we do not know that it always reflects a previously existing neural association in the speakers (different from the associations which do not, at a given time, produce new utterances). For example, before the word *analyticity* came to be used (in modern logic) our data on English may have contained *analytic*, *synthetic*, *periodic*, *periodicity*, *simplicity*, etc. On this basis we would have made some statement about the distributional relation of *-ic* to *-ity*, and the new formation of *analyticity* may have conformed to this statement. But this means only that the pattern or the habit existed in the speakers at the time of the new formation, not necessarily before: the "habit"—the readiness to combine these elements productively—may have developed only when the need arose, by association of words that were partially similar as to composition and environment.

For the position of the speakers is after all similar to that of the linguist. They have heard (and used) a great many utterances among which they perceive partial similarities: parts which occur in various combinations with each other. They produce new combinations of these along the lines of the ones they have heard. The formation of new utterances in the language is therefore based on the distributional relations—as changeably perceived by the speakers—among the parts of the previously heard utterances.[10]

Concerning any habit, i.e. any predisposition to form new combinations along particular distributional lines rather than others, we know about its existence in the speakers only if we have some outside evidence (such as association tests), or if new formations of the type in question have been formed by these speakers. The frequency of slips, new formations, etc., is enough to make us feel that the bulk of the major structural features are indeed reflected in speaking habits—habits which are presumably based, like the linguist's analysis, on the distribu-

[9] As pointed out to the writer by A. W. Holt.

[10] This applies to the grammatical innovation involved in new formations; the selection of morphemes within a class is determined, not only by these "grammatical" associations but also semantically. Cf. the first paragraph of §1.1 above.

tional facts. Aside from this, all we know about any particular language habit is the probability that new formations will be along certain distributional lines rather than others, and this is no more than testing the success of our distributional structure in predicting new data or formations. The particular distributional structure which best predicts new formations will be of greatest interest from many (not all) points of view; but this is not the same as saying that all of that structure exists in the speakers at any particular time prior to the new formations.[11]

**2. Distribution and meaning.**

**2.1. Is there a parallel "meaning structure"?**

While the distinction between descriptive (synchronic) structure and historical change is by now well known, the distinction between distributional structure and meaning is not yet always clear. Meaning is not a unique property of language, but a general characteristic of human activity. It is true that language has a special relation to meaning, both in the sense of the classification of aspects of experience, and in the sense of communication. But the relation is not simple. For example, we can compare the structures of languages with the structure of the physical world (e.g. the kind of phenomena that are expressed by differentiation and integration in calculus), or with what we know about the structure of human response (e.g. association, transference). In either case, it would be clear that the structure of one language or another does not conform in many respects to the structure of physical nature or of human response—i.e. to the structure of objective experience from which we presumably draw our meanings. And if we consider the individual aspects of experience, the way a person's store of meanings grows and changes through the years while his language remains fairly constant, or the way a person can have an idea or a feeling which he cannot readily express in the language available to him, we see that the structure of language does not necessarily conform to the structure of subjective experience, of the subjective world of meanings.[12]

All this is not to say that there is not a great interconnection between language and meaning, in whatever sense it may be possible to use this word. But it is not

[11] Here we have discussed whether the distributional structure exists in the speakers as a parallel system of habits of speaking and of productivity. This is quite different from the dubious suggestion made at various times that the categories of language determine the speakers' categories of perception, a suggestion which may be a bit of occupational imperialism for linguistics, and which is not seriously testable as long as we have so little knowledge about people's categories of perception. Cf. for the suggestion, Benjamin L. Whorf, The Relation of Habitual Thought and Behavior to Language, *Language, Culture and Personality* (*Sapir Memorial Volume*, A. I. Hallowell, L. Spier, S. Newman, eds.) 75–93; Languages and Logic, *The Technology Review* 43–6 (1941); and against it, Eric H. Lennenberg, Cognition in Ethnolinguistics, *Language* 29(1953).463–71; Lewis S. Feuer, Sociological Aspects of the Relation Between Language and Philosophy, *Philosophy of Science* 20(1953).85–100.

[12] In E. G. Schachtel's On Memory and Childhood Amnesia, *Psychiatry* 10(1947).1–26 it is suggested that the experiences of infancy are not recallable in later life because the selection of aspects of experience and the classification of experience embodied in language, which fixes experience for recall, differs from the way events and observations are experienced (and categorized) by the infant.

a one-to-one relation between morphological structure and anything else. There is not even a one-to-one relation between the vocabulary and any independent classification of meaning: We cannot say that each morpheme or word has a single or central meaning, or even that it has a continuous or coherent range of meanings. Accidents of sound change, homonymity, borrowing, forgotten metaphors, and the like can give diverse meanings to a number of phonemic occurrences which we have to consider as occurrences of the same morpheme. Aside from this, if we consider the suggestion of Kurt Goldstein[13] that there are two separate uses and meanings of language—the concrete (e.g. by certain brain-injured patients) and the abstract—it would follow that the same grammatical structure and much the same vocabulary can carry quite different types of speaking activity.

The correlation between language and meaning is much greater when we consider connected discourse. To the extent that formal (distributional) structure can be discovered in discourse, it correlates in some way with the substance of what is being said; this is especially evident in stylized scientific discourse (e.g. reports on experimental work) and above all in the formal discourses (proofs) of mathematics and logic. However, this is not the same thing as saying that the distributional structure of language (phonology, morphology, and at most a small amount of discourse structure) conforms in some one-to-one way with some independently discoverable structure of meaning. If one wishes to speak of language as existing in some sense on two planes—of form and of meaning—we can at least say that the structures of the two are not identical, though they will be found similar in various respects.

**2.2. Are morphemes determined by meaning?**

Since there is no independently-known structure of meanings which exactly parallels linguistic structure, we cannot mix distributional investigations with occasional assists from meaning whenever the going is hard. For example, if the morphemic composition of a word is not easily determined, we cannot decide the matter by seeing what are the component meanings of the word and assigning one morpheme to each: Do *persist, person* contain one morpheme each or two? In terms of meaning it would be difficult to decide, and the decision would not necessarily fit into any resulting structure. In terms of distribution we have *consist, resist, pertain, contain, retain,* etc. (related in phonemic composition and in sentence environment), but no such set for *person;* hence we take *persist* as two morphemes, *person* as one.

Although rough indications of meaning are often used heuristically to guess at the morphemes of a word or utterance, the decision as to morphemic composition is always based on a check of what sections of that word or utterance are substitutable in a structured (patterned) way in that environment; as roughly indicated in the example above.

Where the meanings (in most cases, the translations) are not immediately suggestive, the analysis is laboriously distributional without any heuristic aids to

[13] *Human Nature in the Light of Psychopathology:* The William James Lectures for 1938–39, ch. 3.

test. For example in the Cherokee verb prefixes, we find scores of forms,[14] e.g. /agwalənáʔəgi/ 'I started', /sdəgadhénoha/ 'I and another are searching for you', /sdəgadhénohəgi/ 'I searched for you two'. These have obviously personal reference, but it is impossible to separate out a small set of phonemic segments which will mean 'I' or 'I as subject', 'I as object', etc. It is nevertheless possible to discover the morphemes distributionally. First we identify the words by their distributional relation to the rest of the sentence. We find that certain words with many different stems and a few different prefixes have certain types of environment in common. For example /zinəgali'a/ 'I am cleaning' and /agiyoseha/ 'I am hungry' occur in certain environments in which /uniyoseha/ 'they are hungry' does not occur. We take a set of words each with different stems but which have the same environment in the sense referred to above. We will assume that the sameness in this feature of the environment correlates with some morphemic part that is the same in all these words (and is obviously not the stem).[15] This means that the different prefixes of these words contain alternants of the same morpheme; and we try to state a morphophonemic relation between /z/, /(a)g/, etc., giving the environing conditions (in phonemic rather than morphemic terms if possible) in which each alternant occurs: we write the morpheme {z} and translate it 'I'. Another set, containing e.g. /ozinəgaliʔa/ 'I and others are cleaning', /ogiyoseha/ 'I and others are hungry', would thus be analyzed (in the same manner, but with the aid of {z}) as containing two morphemes, {o} 'others' and {z} 'I'. If we now turn to the set containing /osdinəgaliʔa/ 'I and another are cleaning', /oginiyoseha/ 'I and another are hungry', etc., our morphophonemic knowledge about {z} enables us to separate out /d/, /n/ etc. as alternants of some third morpheme {n}, with undetermined meaning. In /iginiyoseha/ 'you and I are hungry' our known morphophonemics enables us to analyze the prefix as an alternant of {z} plus an alternant of this same {n}, where it seems to have the meaning 'you'. However, in /hinəgaliʔa/ 'you (sg.) are cleaning' we are unable to fit the /h/ into the morphophonemic regularities of {n}, and thus set up a new morpheme {h} 'you'; and in /sdinəgaliʔa/ 'you two are cleaning' we can satisfy the morphophonemic regularities by saying that there are two morphemes: the /s/ alternant of {h} plus the /d/ alternant of {n}.

In this way we can divide each prefix into a unique combination of morphophonemic alternants of the following morphemes: {z} 'I', {h} 'you (sg.)', {a} 'third person sg.', {i} 'plural' (always including 'you', at least due to absence of {o}), {o} roughly 'person(s) excluding you', {n} roughly 'another person, you as first choice'. These morphemes were obtained as solutions to the environmental

[14] The following analysis can be fully understood only if one checks through the actual lists of Cherokee forms. The few forms cited here are taken from William D. Reyburn, Cherokee Verb Morphology II, *International Journal of American Linguistics* 19(1953). 259–73. For the analysis, see the charts and comments in Reyburn's work and in Z. S. Harris, Cherokee Skeletal Grammar, and Cherokee Grammatical Word Lists and Utterances, in the Franz Boas Collection of the American Philosophical Society Library.

[15] This assumption is based on the fact that each morpheme has a different distribution (§2.36), so that same feature of environment points to the same morpheme.

regularities of the prefixed phonemes. The translations offered above are an attempt to assign a single meaning to each on the basis of the meanings of all those words in which it occurs. If we write the prefixes morphophonemically, then the meanings of some of the occurring combinations are: {ozn} (phonemically /osd/ etc.) 'I and he', {oz} 'I and they', {zn} 'I and you (sg.)' {iz} 'I, you, and they', {h} 'you (sg.)', {hn} 'you two', {in} 'you (pl.)'. From this we can try to extract (as above) a single meaning contribution which {n} or {o} or {i} bring to each combination in which they are included. But it was not the isolation of these complicated central meanings (if that is always non-trivially possible) that led us to recognize {n} etc. as morphemes. We do not even know that these central meanings exist for the speakers: the speakers may be subjectively using two homonymous {n} morphemes, or they may be using these prefix combinations as fixed whole entities with only a vague impression of the phonemic and morphophonemic regularities.[16]

So far, we have not touched the great majority of verb forms, those which have objects together with the subjects. By using the morphophonemic relations established previously, we are able to extract the morphemes above from some of these new combinations, and small extensions of the morphophonemics reveal these morphemes in yet other combinations. Then we analyze the prefix in /gəiha/ 'I am killing you' as {z} + {n}, and in /sgwúsədohda/ 'you covered me' as {h} + {z}; and certain order statements about the two prefix components indicate the subject-object relation. The remaining phonemes of some of these prefixes can be grouped by rather simple morphophonemics into a few additional morphemes like {g} 'animate object'; and so we finally obtain a morphemic analysis of all the prefixes. This analysis does not necessarily correlate with any meaning units we may have in mind about person and number. For example, it gives the same morphemes {znn} for the prefix in /sdəgadhénoha/ 'I and another are searching for you (whether sg. or dual but not plural)' and in /sdəgadhénohəgi/ 'I searched for you two'. Even if we find different phonemes with different meanings, e.g. /izə-gow'diha/ 'I and he see you (pl.)' and /izəy-olighi/ 'I and they know you (sg.)' the analysis may say that these are alternants of the same morphemic composition {izn}; in that case both meanings can be obtained for each form.

The methods indicated so sketchily above suggest how the morphemic composition of a word or utterance can be determined by the occurrence of each phoneme sequence relative to others: e.g. *per*, *con* relative to *sist*, *tain;* or /z/ /gi/, /o/, etc. relative to various features of environment which are common to /z/ and /gi/ as against /o/. The final decision as to morphemic analysis always depends on this relative occurence of phoneme sequences, since the grammar then proceeds to state compactly the relative occurrence of the morphemes. That is, we set up as morphemes those phonemic sequences (or features) such that all utterances are compactly statable relative occurrences of them.

The chief difficulty with this is that it provides us only with a criterion that

[16] Since new formations of these combinations do not appear, we cannot apply the productivity tests of §2.1 to discover the speakers' morphemic recognition.

tells us whether a given phoneme sequence is a morpheme or not; more exactly, whether a particular segmentation of an utterance (once we propose it) divides it into morphemic segments. It does not provide us with a procedure which will directly yield a morphemic segmentation of an utterance. There is available, however, a procedure which yields most if not all of the morphemic segmentations of an utterance. In outline it is as follows: Given any test utterance, associate many utterances whose first phoneme is the same as that of the test utterance; and note how many different phonemes follow the first in these utterances. Then consider utterances whose first two phonemes are the same as the first two of the test utterance, and note how many different phonemes follow the first two in these. And so on. If after the first $n$ phonemes the number of different phonemes which follow the $n$th (in the associated utterances) is greater than the number after the first $n-1$ phonemes or the first $n + 1$, then we place a tentative morpheme boundary after the $n$th. Various operations are needed to correct and check the correctness of each result; but together with the final test of patterned relative occurrence, this yields the morphemes of a language without any reference to meaning or informant response.

**2.3. Meaning as a function of distribution.**

Distribution suffices to determine the phonemes and morphemes, and to state a grammar in terms of them. However, both (a) in determining the elements and (b) in stating the relations between them, it turns out that the distributional structure does not give ideal coverage. It must either leave many details unsaid, or else become extremely complicated. For example: (a) Morphemes are determined on the basis of a patterned independence (replaceability in utterances) in respect to other morphemes (or phoneme sequences); but not all morphemes have the same degree of independence: compare *hood* (*boyhood*) with *ness* (*bigness*). (b) The grammatical statements group morphemes into classes, and then say that certain sequences of these classes occur; but not every member of the one class occurs (in any actual body of data) with every member of the other: not every adjective occurs with every noun. Finally we may mention one other respect in which distribution fails to cover all the facts about speech occurrence: (c) We can state distributional regularities only within narrow domains—for phonology usually the immediately neighboring phonemes, for morphology usually the sentence or some part of the sentence.

At all these points where simple distributional regularities are no longer discoverable, people often revert to the position of our man in the street (§1.1) and say that here the only determinant is meaning: (a) *hood* has a meaning which ties it to certain few nouns; (b) with a given noun, e.g. *doctor*, there will be used those adjectives that make sense with it; (c) beyond the sentence there are no significant formal restrictions on what one says, and sentences are strung along purely according to meaning. Now meaning is of course a determinant in these and in other choices that we make when we speak. But as we make these choices we build a stock of utterances each of which is a particular combination of particular elements. And this stock of combinations of elements becomes a factor in the way later choices are made (in the sense indicated in the last two para-

graphs of §1.2); for language is not merely a bag of words but a tool with particular properties which have been fashioned in the course of its use. The linguist's work is precisely to discover these properties, whether for descriptive analysis or for the synthesis of quasi-linguistic system. As Leonard Bloomfield pointed out, it frequently happens that when we do not rest with the explanation that something is due to meaning, we discover that it has a formal regularity or "explanation." It may still be "due to meaning" in one sense, but it accords with a distributional regularity.

If we investigate in this light the areas where there are no simple distributional regularities, we will often find interesting distributional relations, relations which tell us something about the occurrence of elements and which correlate with some aspect of meaning. In certain important cases it will even prove possible to state certain aspects of meaning as functions of measurable distributional relations.

(a) There are different degrees of independence (§3.3). We find complete dependence in the various phonemes of one morpheme, or in the various parts of a discontinuous morpheme (including grammatical agreement). In *hood* we have sufficient independence to make it a separate morpheme, but it is limited to very few predecessors. In *ness* there is more independence. The degree of independence of a morpheme is a distributional measure of the number of different morphemes with which it occurs, and of the degree to which they are spread out over various classes or subclasses. The various members of a distributional class or subclass have some element of meaning in common, which is stronger the more distributional characteristics the class has. The major classes have the kind of common meanings that are associated, say, with the words "noun" or "adjective."

(b) The fact that, for example, not every adjective occurs with every noun can be used as a measure of meaning difference. For it is not merely that different members of the one class have different selections of members of the other class with which they are actually found. More than that: if we consider words or morphemes A and B to be more different in meaning than A and C, then we will often find that the distributions of A and B are more different than the distributions of A and C. In other words, difference of meaning correlates with difference of distribution.

If we consider *oculist* and *eye-doctor*[17] we find that, as our corpus of actually-occurring utterances grows, these two occur in almost the same environments, except for such sentences as *An oculist is just an eye-doctor under a fancier name*, or *I told him Burns was an oculist, but since he didn't know the professional titles, he didn't realize that he could go to him to have his eyes examined.* If we ask informants for any words that may occupy the same place as *oculist* in sentences like the above (i.e. have these same environments), we will not in general obtain *eye-doctor;* but in almost any other sentence we would. In contrast, there are many sentence environments in which *oculist* occurs but *lawyer* does not: e.g. *I've had my eyes examined by the same oculist for twenty years*, or *Oculists often have their prescription blanks printed for them by opticians.* It is not a question of

[17] This particular pair was suggested to me by Y. Bar-Hillel, who however considers that distributional correlates of meaning differences cannot be established.

whether the above sentence with *lawyer* substituted is true or not; it might be true in some situation. It is rather a question of the relative frequency of such environments with *oculist* and with *lawyer*, or of whether we will obtain *lawyer* here if we ask an informant to substitute any words he wishes for *oculist* (not asking what words have the same meaning). These and similar tests all measure the probability of particular environments occurring with particular elements, i.e. they measure the selections of each element.

It is impossible to obtain more than a rough approximation of the relatively common selection of a given word (with almost no indication of its rarer selection). But it is possible to measure how similar are the selection approximations of any two words (within various sets of data). If for two elements A and B we obtain almost the same list of particular environments (selection), except that the environment of A always contains some X which never occurs in the environment of B, we say that A and B are (complementary) alternants of each other: e.g. *knife* and *knive-*. If A and B have identical environments throughout (in terms of our data tests) we say that they are free variants: e.g. perhaps for /ekənamiks/ and /iykənamiks/ *economics*. If the environments of A are always different in some regular way from the environments of B, we state some relation between A and B depending on this regular type of difference: e.g. *ain't* and *am not* have frequent differences of a certain type in their environments (*ain't goin'* but *am not going*) which we would call dialectal. If A and B have almost identical environments except chiefly for sentences which contain both, we say they are synonyms: *oculist* and *eye-doctor*. If A and B have some environments in common and some not (e.g. *oculist* and *lawyer*) we say that they have different meanings, the amount of meaning difference corresponding roughly to the amount of difference in their environments. (This latter amount would depend on the numerical relation of different to same environments, with more weighting being given to differences of selectional subclasses.) If A and B never have the same environment, we say that they are members of two different grammatical classes (this aside from homonymity and from any stated position where both these classes can occur).

While much more has to be said in order to establish constructional methods for such a classification as above, these remarks may suffice to show how it is possible to use the detailed distributional facts about each morpheme. Though we cannot list all the co-occurrents (selection) of a particular morpheme, or define its meaning fully on the basis of these, we can measure roughly the difference in selection between elements, say something about their difference in meaning, and also (above and §4.1) derive certain structural information.

(c) If we investigate the relative occurrence of any part of one sentence in respect to any part of the neighboring sentences in the same discourse, we will find that there are certain regularities (§3.5 end). The sequence of sentences is not entirely arbitrary; there are even certain elements (e.g. pronouns) whose occurrence (and meaning) is specifically related to the grammatically-restricted occurrence of certain other morphemes in the neighboring sentences (§4.1, first paragraph). Such regularities (and meanings) will not extend from one discourse

to another (except to another related in some relevant way to the first, e.g. successive lectures of a series). A consecutive (or seriate) discourse of one or more persons is thus the fullest environmental unit for distributional investigation.[18]

**3. Distributional analysis.**

We now review briefly the basic analysis applicable to distributional facts.

**3.1. Element.** The first distributional fact is that it is possible to divide (to segment) any flow of speech into parts, in such a way that we can find some regularities in the occurrence of one part relative to others in the flow of speech. These parts are the discrete elements which have a certain distribution (set of relative locations) in the flow of speech; and each bit of speech is a particular combination of elements. The first operation is purely segmenting, arbitrary if need be. The first step of segmenting has to be independent of any particular distributional criterion, since we cannot speak of distributional relations until we have not only segments but also a similarity grouping of them (§3.2). After the first segmenting of utterances, each segment is unique and has a unique environment (completely different from every other one); after the segments have been compared, and "similar" ones grouped together, we find that various of these similarity groupings have partially similar and partially different environments. Hence we can speak about the distributional relations of these similarity grtoupings.

If we wish to be able, in the later operations (§3.3–4), to obtain elements (or classes of elements) whose distributions will have maximum regularity, we have to divide not only the time flow into successive portions, but also any single time segment (or succession of time segments) into simultaneous components (of one segment length, e.g. a tone, or longer, e.g. a pitch-stress contour). After we have set up the phonetically more obvious segmentations and simultaneities, and have studied their distribution, we may find that more regular distributions can be obtained if we change our original segmentation of elements, even to ones that are phonetically less obvious, and even if some of our adjusted elements become components which extend over various numbers of other elements.

**3.2. Similarity.** Another essential distributional fact is that some elements are similar to others in terms of certain tests; or are similar in the sense that if we group these similar elements into sets ("similarity groupings"), the distribution of all members of a set (in respect to other sets) will be the same as far as we can discover. This reduces ultimately to the similarity of sound segments under repetition, or in the pair test: $x_1$ is similar to $x_2$ but not to $y_1$ if, when one native

[18] It should be clear that only after we discover what kinds of distributional regularities there are among successive elements or sections in discourses can we attempt any organized semantic interpretation of the successions discovered. Various types of discourses have various types of succession (of sentences, clauses, or other intervals). In mathematics and the constructed "languages" of logic, certain conditions are imposed on what sentences can appear in succession in their connected discourses (proofs): each sentence (line in a proof) has to be a theorem or else derived from a preceding sentence in a particular way. This situation does not hold for natural languages, where the truth-value of logic is not kept constant through successive sentences, and where the types of succession are more varied.

speaker repeats $x_1z$, $x_2z$, $y_1z$, . . ., a second speaker can guess correctly whether $x_1z$ as against $y_1z$ is being said, but not whether $x_1z$ as against $x_2z$ is being said. We call $x_1$ and $x_2$ free variants of each other (or members of a similarity grouping). Note that the pair test involves discrimination of sound but not of meaning.

**3.3. Dependence (serial).** To obtain a least set of elements sufficient for description we join any elements which are completely dependent: if A is a set of similar elements (a similarity grouping) and so is B, and (in a particular type of environment) only AB occurs (not necessarily contiguously), never A or B alone, then we set up AB as a single element (a single set of similar elements).

Thereafter we don't have any two elements which are completely dependent upon each other in occurrence. But our elements have various degrees of dependence: for each element we can say that any utterance (or shorter domain) which contains it will also contain such and such other classes. For example, morpheme A may occur always close to (i.e. within a stateable distance from) any one of a few or many $B_1$, $B_2$, . . . If the sequence $B_1A$ occurs in environments X, it may be that $B_1$ by itself also occurs in X (e.g. *kingdom* and *king*), or that $B_1$ does not (e.g. *kingly* and *king*). The $B_1$ with which A occurs may all have the same types of environment when they occur without A (e.g. all predecessors of *dom* are nouns), or some may have one type and some another (e.g. *ish* occurs with both nouns and adjectives). These are a few of the various degrees and types of occurrence-dependence which an element can have to the elements that occur in the same utterances as it does.

**3.4. Substitutability (parallel).** It will in general appear that various elements have identical types of occurrence-dependence. We group A and B into a substitution set whenever A and B each have the same (or partially same) environments X (X being at first elements, later substitution sets of elements) within a statable domain of the flow of speech. This enables us to speak of the occurrence-dependence of a whole set of elements in respect to other such sets of elements. Some of the types of partial sameness of environment were listed in §2.3(b).

The elements of distributional structure are usually obtained by the operations of §3.1, §3.2 and the first paragraph of §3.3. The distributional relations are usually combinations of §3.3 and §3.4. For example, *hood* occurs after few morphemes $N_1$, $N_2$, . . . of a certain substitution set ("nouns"), *ish* after many of them, *s* and its alternants after all or almost all of them. $N_i$ + *hood* or $N_i$ + *s* occur in the same large environments in which $N_i$ occur alone. But $N_i$ + *ish* occur in different environments than $N_i$ alone; however *ish* also occurs after many members of another substitution set, $A_1$, $A_2$, . . . ("adjectives"), and both $N_i$ + *ish* and $A_i$ + *ish* occur in the larger environments of $A_i$ alone.

**3.5. Domains.** All the statements about dependence and substitutability apply within some specified domain, the domain being determined either by nature (e.g. silence before and after an utterance) or by the types of environment within which there is regularity (e.g. the narrow restriction of *hood* is only to what precedes it, and only to the first morpheme in that direction). It is often possible to state the co-occurrences of elements within a domain in such a way that that domain then becomes the element whose co-occurrences are regular within a larger

domain: e.g. the occurrences of stems and suffixes within word-length, and of words within phrases. Common types of domain are the word, phrase, clause. In many cases the stretches of speech covered by certain long pitch and stress components (or fixed sequences of short pitch and stress components) are identical with the domains of distributional relations: word, sentence.

Although grammar has generally stopped with the sentence, it is possible to find distributional regularities in larger domains. There are certain sentence sequences in which the second can be described as a fixed modification of the first (e.g., with certain restrictions, in the case of questions and answers in English). There are certain types of distributional relation (e.g. between English active and passive, between *buy* and *sell*) which have particular kinds of regularity in (not necessarily immediately) neighboring sentences. For example, if one sentence contains noun A + active (transitive) verb B + noun C, and a neighboring sentence contains C + verb + A, there is a certain likelihood that the verb will be the passive of B; or if the neighboring sentence contains C + the passive of B + some noun, there is a certain likelihood that the second noun will be A or some noun which elsewhere in that discourse has similar individual environments (selection) to those of A. And if one sentence contains A *buys* B *from* C, and a neighboring sentence contains C *sells* B *to* + some noun, there is a good likelihood that the noun will be A or an environmentally similar noun (and given C + some verb + B *to* A, we may expect the verb to be *sell* or some environmentally similar one).[19]

Finally, if we take a whole connected discourse as environment, we find that there are certain substitution sets of morphemes which occur regularly (relative to the other sets) throughout the discourse or some portion of it;[20] these are not the major substitution sets of the language (e.g. nouns) or its grammatical subclasses, but new groupings which are often relevant only to that one discourse. And there are certain sequences of these sets which constitute the subdomains of the discourse, i.e. such that the sets are regular within these intervals and the intervals are regular within the discourse; these intervals are not necessarily sentences or clauses in the sense of grammatical structure. The regularities in a discourse are far weaker and less interrelated than those within a sentence; but they show that occurrence-dependence (and the environment relevant for distribution) can extend throughout a whole discourse.

**3.6. Data.** The distributional investigations sketched above are carried out by

[19] Such relations as that of active to passive, or *buy* to *sell*, are essentially substitutability relations (§3.4), i.e. they show that certain elements have similar environments (e.g. partially inverted ones). The fact that they may appear in neighboring sentences is a serial relation (§3.3) which is a secondary characteristic of certain substitutabilities. Relations like that of active to passive are different from the essentially serial relations of successive intervals of a discourse, discussed at the end of §3.5.

[20] The fact that a discourse contains several or many occurrences of a given substitution class, often in parallel positions, brings out a rare relation in linguistics: the order of occurrence of various members of the same class. Something like this comes up in compound nouns, or in successions of two or more adjectives (sometimes with preferred order). Usually, if two members of a class occur in one domain, their order is not regular (e.g. in most cases of N *and* N); but in compound nouns, for instance, certain members are frequent in the first N position, and others in the second.

recording utterances (as stretches of changing sound) and comparing them for partial similarities. We do not ask a speaker whether his language contains certain elements or whether they have certain dependences or substitutabilities. Even though his "speaking habits" (§1.2) yield regular utterances, they are not sufficiently close to all the distributional details, nor is the speaker sufficiently aware of them. Hence we cannot directly investigate the rules of "the language" via some system of habits or some neurological machine that generates all the utterances of the language. We have to investigate some actual corpus of utterances, and derive therefrom such regularities as would have generated these utterances—and would presumably generate other utterances of the language than the ones in our corpus. Statements about distribution are always made on the basis of a corpus of occurring utterances; one hopes that these statements will also apply to other utterances which may occur naturally. Thus when we say that the selectional difference in *oculist/lawyer* is greater than in *oculist/eye-doctor* (§2.3), or that the selection of nouns around the passive verb is the same as the selection around the active verb but with inverted order (§4.1) we mean that these relations will be approximated in any sufficiently large corpus (especially one built with the aid of eliciting), and that they will presumably apply to any sufficiently large additions to the corpus.

In much linguistic work we require for comparison various utterances which occur so infrequently that searching for them in an arbitrary corpus is prohibitively laborious. To get around this, we can use various techniques of eliciting, i.e. techniques which favor the appearance of utterances relevant to the feature we are investigating (without influencing the speaker in any manner that might bring out utterances which would not have sometimes occurred naturally). In particular, investigations of the selections of particular morphemes (§2.3, 4.1) can hardly be carried out without the aid of eliciting. Eliciting is a method of testing whether a certain utterance (which is relevant to our investigation) would occur naturally: in effect, we try to provide a speaker with an environment in which he could say that utterance—if he ever would naturally say it—without extracting it from him if he wouldn't. For example, if we are testing the active/passive relation we might offer a speaker noun $A_1$ + transitive verb $B_1$ and ask him to complete the sentence in many ways, obtaining a particular selection $C_1$, $C_2$, . . . after the verb. Then we can offer a speaker the passive verb $B_1$ + $A_1$ and ask him to begin the sentence in many ways, checking whether we get about the same selection $C_1$, $C_2$, . . . before the verb. We can repeat this for various $A_i$, and then for various $B_i$.

**4. Distributional relations.**

The methods of §3 yield first of all a representation of each utterance as a combination of elements. They also yield a set of statements about the utterances: what elements and regularities of combination suffice to represent the utterances. One can go beyond this and study the kinds of regularities, and the kinds of relations among elements. As was pointed out at the end of §2.3(b), certain correlations may be discovered even in those distributional facts which are too individual to be directly useful.

**4.1.** As an example of the latter we may consider selectional similarity. For

instance, it is impossible to list all the verbs that follow each particular noun, or all the verbs that follow *who*. But it is possible to state the following relation between the verb selection of nouns and the verb selection of *who:* Under an eliciting test as in §3.6, we will get after *The pianist*— much the same verbs as we will get after *The pianist who*—, and so for every noun. This means that the verb selection of *who* is the same as the verb selection of the noun preceding *who*. We have here a distributional characteristic that distinguishes such pronominal elements from ordinary nouns.

Or we may consider the active/passive relation mentioned in §3.6. If we take a large number of sentences containing a transitive verb in English, e.g. *The kids broke that window last week*, we can elicit sentences consisting of the same verb but with the passive morpheme, the same nouns before and after it but in reverse order, and the same remainder of the sentence, e.g. *That window was broken by the kids last week*. Some of these sentences may be stylistically clumsy, so that they would not occur unless some special circumlocution were involved; but they are obtainable by otherwise valid eliciting techniques.[21] In contrast, if we seek such inversion without the passive, we will fail to elicit many sentences: we can get *The kids saw Mary last week* and *Mary saw the kids last week;* but to *The kids saw the movie* we will never—or hardly ever—get *The movie saw the kids* (even though this sentence is grammatical). Or if we seek such selectional similarity (with or without inversion) for *broke/will break* or the like, we will find the same selection as to preceding and following nouns, but not always as to the rest of the sentence: *The kids broke that window* and *The kids will break that window*, but not *The kids will break that window last week* or *The kids broke that window if they don't watch out.* It thus appears that, using only distributional information about an ordinarily elicited corpus, we can find a relation between the active verb and the passive verb which is different from the relation between *-ed* and *will.*

**4.2.** The distributional regularities can themselves be a subject of study. One can consider recurrent types of dependence and substitutabilities that are found in a language (or in many languages), and find on one level such relations as "subject" and "object" (semantic names for distributional positions), and on a higher level of generality such relations as "constituent" and "head of a construction" (if A occurs in environment X, and AB does too, but B does not, then A is the head of AB). One can consider the parts of a grammar which permit alternative distributional analyses, and check their relation to language change and dialect or idiolect interrelations (since probably every linguistic structure has some points which are structurally in flux). One can investigate what are the structural characteristics of those parts of a language which are productive. Furthermore, one can survey what is similar and what is different in a great many language structures, and how linguistic systems in general differ from such partially similar systems as mathematics and logistic "languages," sign languages, gestures, codes, music.

[21] There will be a few exceptions where the passive is not obtainable. And if we try to elicit the active on the basis of the passive, we run into the difficulty of distinguishing between *by* of the passive (*The letter was finished by Carl*) and *by* as preposition (*The letter was finished by noon*).