DISCOURSE ANALYSIS

ZELLIG S. HARRIS
University of Pennsylvania

This paper presents a method for the analysis of connected speech (or writing).[1] The method is formal, depending only on the occurrence of morphemes as distinguishable elements; it does not depend upon the analyst's knowledge of the particular meaning of each morpheme. By the same token, the method does not give us any new information about the individual morphemic meanings that are being communicated in the discourse under investigation. But the fact that such new information is not obtained does not mean that we can discover nothing about the discourse but how the grammar of the language is exemplified within it. For even though we use formal procedures akin to those of descriptive linguistics, we can obtain new information about the particular text we are studying, information that goes beyond descriptive linguistics.

This additional information results from one basic fact: the analysis of the occurrence of elements in the text is applied only in respect to that text alone—that is, in respect to the other elements in the same text, and not in respect to anything else in the language. As a result of this, we discover the particular interrelations of the morphemes of the text as they occur in that one text; and in so doing we discover something of the structure of the text, of what is being done in it. We may not know just WHAT a text is saying, but we can discover HOW it is saying—what are the patterns of recurrence of its chief morphemes.

Definite patterns may be discovered for particular texts, or for particular persons, styles, or subject-matters. In some cases, formal conclusions can be drawn from the particular pattern of morpheme distribution in a text. And often it is possible to show consistent differences of structure between the discourses of different persons, or in different styles, or about different subject-matters.

1. PRELIMINARIES

1.1. The Problem. One can approach discourse analysis from two types of problem, which turn out to be related. The first is the problem of continuing descriptive linguistics beyond the limits of a single sentence at a time. The other is the question of correlating 'culture' and language (i.e. non-linguistic and linguistic behavior).

The first problem arises because descriptive linguistics generally stops at sentence boundaries. This is not due to any prior decision. The techniques of linguistics were constructed to study any stretch of speech, of whatever length. But in every language it turns out that almost all the results lie within a rela-

[1] It is a pleasure to acknowledge here the cooperation of three men who have collaborated with me in developing the method and in analyzing various texts: Fred Lukoff, Noam Chomsky, and A. F. Brown. Earlier investigations in the direction of this method have been presented by Lukoff, Preliminary analysis of the linguistic structure of extended discourse, University of Pennsylvania Library (1948). A detailed analysis of a sample text will appear in a future number of LANGUAGE.

tively short stretch, which we may call a sentence. That is, when we can state a restriction on the occurrence of element *A* in respect to the occurrence of element *B*, it will almost always be the case that *A* and *B* are regarded as occurring within the same sentence. Of English adjectives, for instance, we can say that they occur before a noun or after certain verbs (in the same sentence): *the dark clouds, the future seems bright*; only rarely can we state restrictions across sentence boundaries, e.g. that if the main verb of one sentence has a given tense-suffix, the main verb of the next sentence will have a particular other tense-suffix. We cannot say that if one sentence has the form *NV*, the next sentence will have the form *N*. We can only say that most sentences are *NV*, some are *N*, and so on; and that these structures occur in various sequences.

In this way descriptive linguistics, which sets out to describe the occurrence of elements in any stretch of speech, ends up by describing it primarily in respect to other elements of the same sentence. This limitation has not seemed too serious, because it has not precluded the writing of adequate grammars: the grammar states the sentence structure; the speaker makes up a particular sentence in keeping with this structure, and supplies the particular sequence of sentences.

The other problem, that of the connection between behavior (or social situation) and language, has always been considered beyond the scope of linguistics proper. Descriptive linguistics has not dealt with the meanings of morphemes; and though one might try to get around that by speaking not of meanings, but of the social and interpersonal situation in which speech occurs, descriptive linguistics has had no equipment for taking the social situation into account: it has only been able to state the occurrence of one linguistic element in respect to the occurrence of others. Culture-and-language studies have therefore been carried on without benefit of the recent distributional investigations of linguistics. For example, they list the meanings expressed in the language by surveying the vocabulary stock; or they draw conclusions from the fact that in a particular language a particular set of meanings is expressed by the same morpheme; or they discuss the nuances of meaning and usage of one word in comparison with others (e.g. in stylistics). Culture-and-language studies have also noted such points as that phrases are to be taken in their total meaning rather than as the sum of the meanings of their component morphemes, e.g. that *How are you* is a greeting rather than a question about health—an example that illustrates the correlation of speech with social situation. Similarly, personality characteristics in speech have been studied by correlating an individual's recurrent speech features with recurrent features of his behavior and feeling.[2]

1.2. Distribution within discourse. Distributional or combinatorial analysis within one discourse at a time turns out to be relevant to both of these problems.

On the one hand, it carries us past the sentence limitation of descriptive

[2] Correlations between personality and language are here taken to be not merely related to correlations between 'culture' and language, but actually a special case of these. The reason for this view is that most individual textual characteristics (as distinguished from phonetic characteristics) correlate with those personality features which arise out of the individual's experience with socially conditioned interpersonal situations.

linguistics. Although we cannot state the distribution of sentences (or, in general, any inter-sentence relation) when we are given an arbitrary conglomeration of sentences in a language, we can get quite definite results about certain relations across sentence boundaries when we consider just the sentences of a particular connected discourse—that is, the sentences spoken or written in succession by one or more persons in a single situation. This restriction to connected discourse does not detract from the usefulness of the analysis, since all language occurrences are internally connected. Language does not occur in stray words or sentences, but in connected discourse—from a one-word utterance to a ten-volume work, from a monolog to a Union Square argument. Arbitrary conglomerations of sentences are indeed of no interest except as a check on grammatical description; and it is not surprising that we cannot find interdependence among the sentences of such an aggregate. The successive sentences of a connected discourse, however, offer fertile soil for the methods of descriptive linguistics, since these methods study the relative distribution of elements within a connected stretch of speech.

On the other hand, distributional analysis within one discourse at a time yields information about certain correlations of language with other behavior. The reason is that each connected discourse occurs within a particular situation—whether of a person speaking, or of a conversation, or of someone sitting down occasionally over a period of months to write a particular kind of book in a particular literary or scientific tradition. To be sure, this concurrence between situation and discourse does not mean that discourses occurring in similar situations must necessarily have certain formal characteristics in common, while discourses occurring in different situations must have certain formal differences. The concurrence between situation and discourse only makes it understandable, or possible, that such formal correlations should exist.

It remains to be shown as a matter of empirical fact that such formal correlations do indeed exist, that the discourses of a particular person, social group, style, or subject-matter exhibit not only particular meanings (in their selection of morphemes) but also characteristic formal features. The particular selection of morphemes cannot be considered here. But the formal features of the discourses can be studied by distributional methods within the text; and the fact of their correlation with a particular type of situation gives a meaning-status to the occurrence of these formal features.

1.3. Conjunction with grammar. The method presented here is thus seen to grow out of an application of the distributional methods of linguistics to one discourse at a time. It can be applied directly to a text, without using any linguistic knowledge about the text except the morpheme boundaries. This is possible because distributional analysis is an elementary method, and involves merely the statement of the relative occurrence of elements, in this case morphemes. To establish the method for its own sake, or for possible application to non-linguistic material, no prior knowledge should be used except the boundaries of the elements.

However, when we are interested not in the method alone but in its results, when we want to use the method in order to find out all that we can about a

particular text, then it is useful to combine this method with descriptive linguistics. To this end we would use only those statements of the grammar of the language which are true for any sentence of a given form. For example, given any English sentence of the form N_1VN_2 (e.g. *The boss fired Jim*), we can get a sentence with the noun phrases in the reverse order N_2—N_1 (*Jim — the boss*) by changing the suffixes around the verb:[3] *Jim was fired by the boss*. The justification for using such grammatical information in the analysis of a text is that since it is applicable to any N_1VN_2 sentence in English it must also be applicable to any N_1VN_2 sentence in the particular text before us, provided only that this is written in English. The desirability of using such information is that in many cases it makes possible further applications of the discourse-analysis method.

How this happens will appear in §2.33; but it should be said here that such use of grammatical information does not replace work that could be done by the discourse-analysis method, nor does it alter the independence of that method. It merely transforms certain sentences of the text into grammatically equivalent sentences (as N_1VN_2 above was transformed into $N_2V^*N_1$), in such a way that the application of the discourse-analysis method becomes more convenient, or that it becomes possible in particular sections of the text where it was not possible to apply it before. And it will be seen that the decision where and how to apply these grammatical transformations need not be arbitrary but can be determined by the structure of the text itself.

The applicability of the discourse-analysis method in particular texts can be further increased if we not only use the ordinary results of grammar but also extend descriptive linguistics to deal with the special distributions of individual morphemes. There are cases, as will be seen in §2.33 below, when we would like to use information not about all the morphemes of some class (like the transformability of V into V^*) but about a particular member of the class, about a restriction of occurrence which is true for that one morpheme but not for the others of its class. Such information is not in general available today; but it can be obtained by methods which are basically those of descriptive linguistics.

Finally, the applicability of discourse analysis in particular texts can sometimes be increased if we draw our information not only from the grammar of the language but also from a descriptive analysis of the body of speech or writing of which our text is a part. This larger body of material may be looked upon as the dialect within which the text was spoken or written, and we can say as before that any distributional statement which is true for all sentences of a given form in that dialect will also hold for any sentence of that form in the text under consideration.

2. The Method

2.0. The nature of the method. We have raised two problems: that of the distributional relations among sentences, and that of the correlation between language and social situation. We have proposed that information relevant to

[3] When the verb is transformed to suit such an inversion of subject (N_1 above) and object (N_2), we may call the new verb form the conjugate of the original form, and write it V^*. Then an active verb has a passive verb as its conjugate, and a passive verb has an active verb as its conjugate.

both of these problems can be obtained by a formal analysis of one stretch of discourse at a time. What KIND of analysis would be applicable here? To decide this, we consider what is permitted by the material.

Since the material is simply a string of linguistic forms arranged in successive sentences, any formal analysis is limited to locating linguistic elements within these sentences—that is, to stating the occurrence of elements. We cannot set up any method for investigating the nature or composition of these elements, or their correlations with non-linguistic features, unless we bring in new information from outside.

Furthermore, there are no particular elements, say *but* or *I* or *communism*, which have a prior importance, such as would cause us to be interested in the mere fact of their presence or absence in our text. Any analysis which aimed to find out whether certain particular words, selected by the investigator, occur in the text or not, would be an investigation of the CONTENT of the text and would be ultimately based on the MEANINGS of the words selected. If we do not depend upon meaning in our investigation, then the only morphemes or classes which we can deal with separately are those which have grammatically stated peculiarities of distribution.

Since, then, we are not in general interested in any particular element selected in advance, our interest in those elements that do occur cannot be merely in the tautologic statement THAT they occur, but in the empirical statement of HOW they occur: which ones occur next to which others, or in the same environment as which others, and so on—that is, in the relative occurrence of these elements with respect to each other. In this sense, our method is comparable to that which is used, in the case of a whole language, in compiling a grammar (which states the distributional relations among elements), rather than in compiling a dictionary (which lists all the elements that are found in the language, no matter where).

Finally, since our material is a closed string of sentences, our statement about the distribution of each element can only be valid within the limits of this succession of sentences, whether it be a paragraph or a book. We will see in §2.33 that we can sometimes use information about the distribution of an element outside our material; but this can be only an external aid, brought in after the distribution of the element within the discourse has been completely stated.

2.1. General statement of the method. It follows from all this that our method will have to provide statements of the occurrence of elements, and in particular of the relative occurrence of all the elements of a discourse within the limits of that one discourse.

2.11. ELEMENTS IN IDENTICAL ENVIRONMENTS. We could satisfy this requirement by setting up detailed statements of the distribution of each element within the discourse, just as in descriptive linguistics we could set up individual statements summarizing all the environments (i.e. the distribution) of each element in various sentences of the language. However, such individual statements are unmanageably large for a whole language, and are unwieldy even for a single text. In both cases, moreover, the individual statements are an inconvenient basis for inspection and comparison, and for the deriving of general statements.

Therefore, in discourse analysis as in descriptive linguistics, we collect those elements which have like distributions into one class, and thereafter speak of the distribution of the class as a whole rather than of each element individually.

When two elements have identical distributions, this operation of collecting presents no problem. In descriptive linguistics, however, the opportunity rarely occurs, since few words have identical distributions throughout a language.[4] It may occur more frequently in a repetitive text, where two words may be always used in identical parallel sentences—e.g. in stylistically balanced myths, in proverbs, in sloganeering speeches, and in 'dry' but meticulous scientific reports.

2.12. ELEMENTS IN EQUIVALENT ENVIRONMENTS. In the much more frequent case where two elements occur in environments which are almost but not quite identical, we may be able to collect them into one distributional class by setting up a chain of equivalences connecting the two almost identical environments.[5] This is done in descriptive linguistics when we say that the class of adjectives A occurs before the class of nouns N, even though a particular A (say *voluntary*) may never occur before a particular N (say *subjugation*). It is done in discourse analysis when we say that two stretches which have the same environment in one place are equivalent even in some other place where their environment is not the same.

Suppose our text contains the following four sentences: *The trees turn here about the middle of autumn*; *The trees turn here about the end of October*; *The first frost comes after the middle of autumn*; *We start heating after the end of October*. Then we may say that *the middle of autumn* and *the end of October* are equivalent because they occur in the same environment (*The trees turn here about* —), and that this equivalence is carried over into the latter two sentences. On that basis, we may say further that *The first frost comes* and *We start heating* occur in equivalent environments. (The additional word *after* is identical in the two environments.) Such chains, which carry over the equivalence of two stretches from one pair of sentences where their environment is indeed identical to another pair of sentences where it is not, must of course be constructed with adequate safeguards, lest everything be made equivalent to everything else, and the analysis collapse. This problem appears also in setting up classes in descriptive linguistics; the kind of safeguards necessary in discourse analysis will be discussed in §2.21.

More generally, if we find the sequences AM and AN in our text, we say that M is equivalent to N or that M and N occur in the identical environment A, or that M and N both appear as the environment of the identical element (or sequence of elements) A; and we write $M = N$. Then if we find the sequence BM and CN (or MB and NC) in our text, we say that B is (secondarily) equivalent to C, since they occur in the two environments M and N which have been found to be equivalent; and we write $B = C$. If we further find BK and CL, we would write $K = L$ by virtue of their having occurred in the secondarily equivalent environments B and C; and so on. As an example, let us continue our text

[4] Two personal names may have identical distributions. Thus, for every sentence containing *Bill* we may find an otherwise identical sentence containing *Jim* instead.

[5] I owe a clarification of the use of such chains to the unpublished work of Noam Chomsky.

fragment with the following sentence: *We always have a lot of trouble when we start heating but you've got to be prepared when the first frost comes.* Then we would say that *We always have a lot of trouble* is equivalent (for this text) to *but you've got to be prepared.*

Saying that $B = C$ does not mean that they are IN GENERAL equal to each other, or that they MEAN the same thing. The equal-sign is used only because the relation between B and C satisfies the technical requirements of the relation which is generally marked by that sign. All we mean when we write $B = C$ is that this relation is a step in a chain of equivalences: on the one hand, B and C are found in equivalent environments (M and N); and on the other, any two environments in which B and C are found will be considered equivalent (K and L).

It is not relevant to ask, 'Is it TRUE that $B = C$?' or 'Have we the RIGHT to say that $K = L$ merely because $B = C$ and because BK and CL occur?' All that is proposed here is a method of analysis; the only relevant questions are whether the method is usable, and whether it leads to valid and interesting results. Whether the method is usable can be judged on the basis of its operations, without regard to its results, as yet unseen. Whether these results are of interest will be considered in Section 3 below, where we will see that the chains of equivalence reveal a structure for each text. There is no question whether we have the 'right' to put $K = L$, because all we indicate by $K = L$ is that BK and CL occur and that $B = C$. The justification will depend on the fact that when we put all the equivalences together we will obtain some information about the structure of the text.

2.13. EQUIVALENCE CLASSES. After discovering which sequences occur in equivalent environments, we can group all of them together into one equivalence class. In our formulaic statement we have $A = B$ (both occur before M), and $A = C$ (both before N), and $B = C$, so that we consider A, B, C all members of one equivalence class. Similarly, M, N, K, L are members of another single equivalence class. In our example, *The trees turn here in* (T_1) and *The first frost comes after* (T_2) and *We start heating after* (T_3) are all members of one equivalence class T, while *the middle of autumn* (E_1) and *the end of October* (E_2) are members of another equivalence class E. There is yet a third class E' consisting of *We always have a lot of trouble when* and *but you've got to be prepared when.* E' is obviously related to E, since both occur with the last two members of T. But E occurs AFTER T, whereas E' occurs BEFORE T.

In terms of these classes, the five sentences of our text fragment can be written as six formulas (since the last sentence was a double one): TE, TE, TE, TE, $E'T$, $E'T$. It is clear that we cannot make one class out of E and E'; but we can say that when the order of E and T is reversed (when E is 'reflected' in T), we get E' instead of E. If we change the members of E' to the form they would have if they came after T instead of before, then those changed members of E' become regular members of E. For example, we might say *We start heating at the cost of a lot of trouble always, but the first frost comes in a way you've got to be prepared for.* This sentence has the form $TE\ TE$. The new phrase *at the cost of a lot of trouble always* is a member of E by virtue of its occurrence after T; we can mark it E_3. Of course, we must show that it is equivalent to *We always have a*

lot of trouble, except for the reversed position in respect to T; to show this, we need techniques which will be discussed in §2.33. Similarly, we must show that the new E phrase *but ... in a way you've got to be prepared for* (E_4) is the T reflection of the E' phrase *but you've got to be prepared when*. If we can show these two reflection-equivalences, we can replace the two E' phrases by the changed phrases which we get when we put them in the E position. As a result we have two more members of E, and no peculiar E' class.

In such ways we can set up equivalence classes (like E) of all sequences which have equivalent environments, i.e. the same equivalence classes on the same side (before or after), within the text. The elements (or sequences of elements) which are included in the same equivalence class may be called equivalent to, or substituents of, each other. We will see later (§3.3) that in some respects (especially in extensions of the text) they may be considered substitutable or interchangeable for each other. In that case the equivalence class may also be called a substitution class.

Note especially that the operation of grouping non-identical forms into the same equivalence class does not depend upon disregarding small differences in meaning among them, but upon finding them in equivalent environments. This means either finding them in identical environments (*the middle of autumn* and *the end of October* both occur in the environment *The trees turn here in* —) or else finding them in environments which are at the ends of a safeguarded chain of equivalences (*The first frost comes* and *We start heating* occur in the equivalent environments *after the middle of autumn* and *after the end of October*). The method is thus fundamentally that of descriptive linguistics and not of semantics.

2.14. SENTENCE ORDER. At this point we come to an operation not used in descriptive linguistics: representing the order of successive occurrences of members of a class. In descriptive linguistics order comes into consideration only as the relative position of various sections of a sequence, as when the order of article and noun is described by saying that the first precedes the second along the line of a noun phrase. In discourse analysis we have this kind of order as among the sections of a sentence, e.g. the different orders of E and E' in respect to T.

The order of successive sentences, or of some particular word class in various sentences (say, the relation of successive subjects), is not generally relevant to descriptive linguistics, because its distributional statements are normally valid within only one sentence at a time. Here, however, where we are dealing with a whole discourse at once, this problem is a real one. If we were considering each sentence separately, and relating it to others only for purposes of structural comparison, we could say (as in descriptive linguistics) that each sentence in our text fragment consists of TE. But since we are speaking of the text as a whole, we cannot say that it consists merely of TE six times over. The particular members of E and of T are different in the various sentences; and these differences may be (for all we know) peculiar to this text, or to a group of similar texts.

Our text fragment can be structually represented by a double array, the horizontal axis indicating the material that occurs within a single sentence or sub-

sentence, and the vertical axis (here broken into two parts) indicating the successive sentences:

$$\begin{array}{ll} T_1\,E_1 & T_3\,E_2 \\ T_1\,E_2 & T_3\,E_3 \\ T_2\,E_1 & T_2\,E_4 \end{array}$$

In this double array, the various symbols in one horizontal row represent the various sections of a single sentence or subsentence of the text, in the order in which they occur in the sentence (except insofar as the order has been altered by explicit transformations in the course of reducing to symbols, as in the change from E' to E). The vertical columns indicate the various members of an equivalence class, in the order of the successive sentences in which they occur.

The reason why the order of symbols in a row may differ from the order of elements in a sentence, is that our linguistic knowledge of sentence structure enables us to deal with the elements separately from their order. We do this when we disregard in our symbols any order that is automatic and that would reappear as soon as our symbols are translated back into language, as when *but* ... is included in E_4 even though it is necessarily separated from E_4 in the actual sentence (since *but* generally occurs at the beginning of a sentence structure, no matter which section of the sentence it may be related to). We also perform this separation of elements from their order when we replace some non-automatic order which has morphemic value by the morphemes which are grammatically equivalent to it; for example, when we replace N_1VN_2 by $N_2V^*N_1$ (replacing *The boss fired Jim* by *Jim was fired by the boss*); or when, in our text fragment, E' before T is replaced by E after T.

In contrast with this cavalier treatment of horizontal order, we cannot alter anything about the order within a vertical column. Here we have no prior linguistic knowledge to tell us which orderings of sentences (if any) are automatic and therefore not to be represented, or which orderings can be replaced by different but equivalent orderings. A closer study of sentence sequences in the language may some day give us such information in the future; for instance, to take a very simple case, it might show that sentence sequences of the form P *because* Q are equivalent to sequences of the form Q *so* P, or that P *and* Q is interchangeable with Q *and* P (whereas P *but* Q may not be similarly interchangeable with Q *but* P).[6] Furthermore, a closer study of a particular text, or of texts of a particular type, may show that certain whole sequences of sentences are equivalent or interchangeable; and with this information we may be able to simplify the vertical axis of the double array, for example by finding periodically repeated vertical patterns. Pending such specific information, however, the vertical axis is an exact reproduction of the order of the sentences or subsentences in the text.

2.15. SUMMARY. We can now survey the whole method as follows. We call elements (sections of the text—morphemes or morpheme sequences) equivalent

[6] Mathematics, and to a greater extent logic, have already set up particular sentence orders which are equivalent to each other. This equivalence can be rediscovered linguistically by finding that the distribution of each sequence is equivalent to that of the others. Our interest here, however, is to discover other equivalences than those which we already know to have been explicitly built into a system.

to each other if they occur in the environment of (other) identical or equivalent elements. Each set of mutually equivalent elements is called an equivalence class. Each successive sentence of the text is then represented as a sequence of equivalence classes, namely those to which its various sections belong. We thus obtain for the whole text a double array, the horizontal axis representing the equivalence classes contained in one sentence, and the vertical axis representing successive sentences. This is a tabular arrangement not of sentence structures (subjects, verbs, and the like), but of the patterned occurrence of the equivalence classes through the text.

If the different sentences contain completely different classes, the tabular arrangement is of no interest; but this is generally not the case. In almost every text there are passages in which particular equivalence classes recur, in successive sentences, in some characteristic pattern. The tabular arrangement makes it possible to inspect this pattern; and we can derive from it various kinds of information about the text, certain structural analyses of the text, and certain critiques of the text. For the equivalence classes, which are set up distributionally, the tabular arrangement shows the distribution. For the text as a whole, the tabular arrangement shows certain features of structure.

2.2. Procedure. We will now illustrate the procedure in detail by applying it to a specific text, of a type as common today as any other that reaches print:[6a]

Millions Can't Be Wrong!

Millions of consumer bottles of X– have been sold since its introduction a few years ago. And four out of five people in a nationwide survey say they prefer X– to any hair tonic they've used. Four out of five people in a nationwide survey can't be wrong. You too and your whole family will prefer X– to any hair tonic you've used! Every year we sell more bottles of X– to satisfied customers. You too will be satisfied!

2.21. DETERMINING THE EQUIVALENCE CLASSES. The first step in discourse analysis is to decide which elements are to be taken as equivalent to each other, i.e. placed in the same column of the tabular arrangement. This is not always automatic—simply a matter of finding which elements have identical environments; for (1) there may be several ways of breaking a sentence down into equivalent parts, and (2) we must decide which way to look for the less obvious equivalence chains.

The simplest starting point is to consider the more frequently repeated words of the text. Almost every text has particular words which occur a great many times;[7] and these will often be key words of that text. The various occurrences

[6a] This is the actual text of an advertisement, found on a card which had presumably been attached to a bottle of hair tonic. A considerable number of advertisements have been analyzed, because they offer repetitive and transparent material which is relatively easy to handle at this stage of our experience with discourse analysis. Many other kinds of texts have been analyzed as well—sections of textbooks, conversations, essays, and so on; and a collection of these will be published soon.

[7] This will be true, though to a lesser extent, even in the writing of those who obey the school admonition to use synonyms instead of repeating a word. In such cases the synonyms

of such a word can certainly be put into one column, i.e. one equivalence class. And the neighboring words can be put into another single equivalence class because they occur in identical environments. In our text no key words are apparent; but we can start with the identical, and hence of course equivalent, repeated sequence *can't be wrong*. Then *Millions* is equivalent (for this text) to *Four out of five people in a nationwide survey*, since both occur before that sequence.

This first step might of course also be performed for such repeated words as *of*. But if we were to collect all the environments of the word *of*, we could not use the resulting equivalence class to build up a chain of further equivalences, because nothing else would be found in their environment. Whereas the class containing *Millions* and *Four out of five* ... , which we obtain from repetitions of *can't be wrong*, will be found, in the paragraphs below, to tie up with other sections of our text.

From this utilization of repetitions we go on to construct chains of equivalence—that is, we ask what other environments occur for *Millions* and for *Four out of five*... . For *Millions* we have one other environment, namely *of consumer bottles*, etc. It will turn out in our further work (§3.2) that this environment clashes with the environments of *Four out of five*... . Therefore we will tentatively set aside the sequence *of consumer bottles*, etc. As for *Four out of five people in a nationwide survey*, we find it in one other environment: before *say they prefer X– to any hair tonic they've used.*

We proceed along this equivalence chain by looking for some other environment in which *say they prefer X–* ... occurs. There is one such occurrence, but it differs by having *you* where the first occurrence has *they*. At first it seems that this difference makes it impossible for us to consider the two sequences equivalent, since our method provides for no approximation technique, no measurement of more and less difference, such as might permit us to say that these two sequences are similar enough to be considered equivalent. Indeed, since we do not operate with the meanings of the morphemes, the replacing of *they* by *you* might constitute a great difference (as it would if the whole text dealt with the distinction between 'you' and 'they'). As they stand, therefore, these two sequences would be left unrelated by our method; at most that method could separate out the identical and the different portions. It so happens, however, that a little consideration shows these two sequences to be contextually identical —that is, identical in respect to their relevant environment or context. This will be seen in §2.31.

In constructing chains of equivalence the first safeguard is adherence to the formal requirements of the method. If we never make any approximations, never overlook some 'small' difference in environment, we will be certain that any two members of one equivalence class have at least one environment in common. If we wish to put two elements into one class even though no environ-

---

will often be found in the same environments as the original not-to-be-repeated word. In contrast, when a writer has used a different word because he intends the particular difference in meaning expressed by it, the synonym will often occur in correspondingly different environments from the original word.

ment of one is identical with some environment of the other, it will have to be at the cost of some explicit assumption, added to the method, which equates the two environments or nullifies their difference.

The final factor in our decision to include or not to include two elements in one equivalence class is the way the resulting class will function in the analysis of the text, i.e. the kind of double array we get by using that class. This factor must play a part, since there are often various possible chains of equivalence that equally satisfy our method. The criterion is not some external consideration like getting the longest possible chain, but rather the intrinsic consideration of finding some patterned distribution of these classes, i.e. finding some structural fact about the text in terms of these classes. In other words, we try to set up such classes as will have an interesting distribution in our particular text. This may seem a rather circular safeguard for constructing equivalence chains. But it simply means that whenever we have to decide whether to carry an equivalence chain one step further, we exercise the foresight of considering how the new interval will fit into our analyzed text as it appears when represented in terms of the new class. This kind of consideration occurs in descriptive linguistics when we have to decide, for example, how far to subdivide a phonemic sequence into morphemes.[8]

One might ask what right we have to put two words into one equivalence class merely because they both occur in the same environment. The answer is that the equivalence class indicates no more than the distributional work which its members do in the text. If the two words occur only in identical or equivalent environments in this text, then in this text there is no difference in their distribution (aside from their order in the column, which is preserved). We are not denying any difference in meaning, or in distribution outside this text.

So far we have recognized two equivalence classes. One, which we will mark P, at present includes

Millions
Four out of five people in a nationwide survey

The other, which we will mark W, at present includes

can't be wrong
say they prefer X– to any hair tonic they've used

2.22. SEGMENTATION. Once we have a rough idea of what equivalence classes we wish to try out in our text, we segment the text into successive intervals in such a way as to get, in each interval, like occurrences of the same equivalence classes. If our classes so far are P and W, and if we have a few PW successions, we try to segment into intervals each containing precisely one P and one W. For example, the title of the advertisement is represented by PW. The first sentence after the title seems to contain a P (the word *Millions*), but the rest of the sentence neither equals nor contains W; hence the sentence is as yet unanalyzed, and even its P is in doubt.

[8] Cf. Harris, Methods in structural linguistics 160 (Chicago, 1951). It goes without saying that this vague use of foresight is a preliminary formulation. Detailed investigations will show what may be expected from different kinds of equivalence chains, and will thus make possible a more precise formulation of safeguards.

Assignment of an element to a particular class is always relative to the assignment of its environment. The elements are not defined except in relation to their environment. For all we know, *Millions* in this sentence might not even be the same word as *Millions* in the title. In descriptive linguistics two phonemically identical segments are the same morpheme only if they occur in the same morpheme class: *sun* and *son* would presumably have to be considered the 'same' morpheme, no less than *table* (of wood) and *table* (of statistical data). If they occur in different morpheme classes, e.g. *sea* and *see*, they certainly are not the same morpheme; and if we want to keep in view the connection between (*a*) *table* and (*to*) *table*, we have to speak of classed and unclassed morphemes, and say that the unclassed morpheme *table* appears both in the N class and in the V class. Similarly, if *Millions* occurs twice we try to consider it a repeated 'same' morpheme (hence in the same class), and so consider its two environments equivalent. But we may find later that a better text-analysis is obtained by not considering those two environments equivalent (because the first environment is equivalent to one sequence A in the text, while the second is equivalent to a different sequence B which is not equivalent to A). In that case we may have to consider the two occurrences of *Millions* as belonging to two different classes. In §3.2, we will find this to be the case here.

To return to our segmentation. The second sentence in our text is PW, and the third is PW. Hence we try to segment our text into successive stretches each of which will contain just PW and no more. These stretches will then be the successive rows of our double array. They will often be whole sentences, but not necessarily: they may also be the separate sections of a compound sentence, each of which has its own sentence structure (as in the two $E'T$ of §2.13). But they may also be any other stretches taken out of the sentence. For example, if we found in our advertisement the sentence *Millions of people—four out of five—can't be wrong when they say they prefer X–*, which as it stands seems to consist of $PPWW$, we would try to reduce it to two PW intervals. Such less obvious segmentations require care, since we want not only the P and the W occurrences to be the same in each interval, but also the relation between P and W to be the same. When each whole sentence in a string is reduced to PW, the relation between P and W in each interval is the same; from descriptive linguistics we know it is the relation of subject to predicate. We do not need to use this specific information in tabulating our text as a succession of PW, but we do assume that whatever the relation between P and W in one interval, it is the same in all the other intervals. Otherwise we would be wrong in saying, when we see such a double array as the successive TE of §2.14, that the successive intervals are identical in terms of T and E. Techniques for checking the sameness of the relation between the equivalence classes in each row will be discussed in §§2.32–3.

2.23. SETS OF LIKE SEGMENTS. The attempt to divide a text into intervals containing the same equivalence classes (in the same relation to each other) will not generally succeed throughout a whole text. There may be individual sentences here and there which simply do not contain these classes. These may turn out to be introductory sentences, or offshoots of some other set of equivalence classes. And there may be successive sections of the text, each of which contains

its own equivalence classes different from those of other sections. These may be paragraph-like or chapter-like sub-texts within the main text.

In the course of seeking intervals which contain the same classes, our procedures will discover the limits of this sameness, i.e. the points at which we get text-intervals containing different classes. In the general case, then, a text will be reduced not to a single set of identical rows (each row, like *TE*, representing an interval with the same equivalence classes), but to a succession of sets of identical rows, with occasional individually different rows occurring at one point or another.

Having obtained this result, we compare the various sets and individual rows to see what similarities and differences exist among them in the arrangement of their classes, whether the specific classes are different or not. We try to discover patterns in the occurrence of such similarities among the successive sets and individually different rows. For example, let a text come out to be *AB TE TE TE A'B' EP EP AB KD LM LM K'D' MS MS MS FBV MS*. Then, using [*TE*] to indicate a set of *TE* intervals, and temporarily disregarding the *FBV*, we can represent the text by *AB* [*TE*] *A'B'* [*EP*] *AB KD* [*LM*] *K'D'* [*MS*]. We note, further, that *AB* [*TE*] *A'B'* [*EP*] and *KD* [*LM*] *K'D'* [*MS*] are structurally identical: both have the form *w* [*xy*] *w'* [*yz*]. This form is a particular relation of *w*, *x*, *y*, and *z*. Our text consists of two occurrences of this structure, with the *w* of the first occurrence (that is, the *AB*) appearing again between the two structures (or before the second structure), and with a unique *FBV* before the end of the last structure.

2.3. Accessory techniques. The main procedure, as described in the foregoing section, must be refined and supplemented by a number of accessory techniques.

2.31. INDEPENDENT OCCURRENCE. The distribution of equivalence classes (their pattern of occurrence), and the segmentation of intervals containing them, depend on what we recognize as an occurrence of an element. At first sight, this would seem to be trivial: in the stretch *say they prefer X– to any hair tonic they've used* we obviously find *say* once, *they* twice, and so on. Closer consideration, however, will show that not all occurrences of elements are independent: there are some elements which occur, in a given environment, only when some other element is present. This situation is known from descriptive linguistics; for example, the *-s* of *he walks* is taken not as an independent element but as an automatic concomitant of *he*, by comparison with *I walk*, *you walk*;[9] and in forms like *both he and I* the *and* always occurs if *both* is present, so that *both ... and* can be taken as one element rather than two. In the same way, if in a particular text we find identical (repeated) or different elements, of which one occurs only if the other is present, we conclude that these occurrences are not independent of each other, and mark their joint occurrence as a single element in the representation of the text.

For *they prefer X– to any hair tonic they've used*, our only comparison is *You too and your whole family will prefer X– to any hair tonic you've used*. In each case,

[9] The *-s* is also a part of all singular nouns (*The child walk-s*, etc.). Or else *walks*, *goes*, and the like can be taken as alternants of *walk*, *go*, etc. after *he* and singular nouns.

the stretch before *prefer* contains the same word that we find before *'ve*. We can therefore say that the word before *'ve* is not independent; rather, the choice of one or the other member of the set *they/you* depends on which word of that set occurs before *prefer*. Writing *Q* as a sign to repeat that member of the set *they/you* which occurs in the stretch before *prefer*, we obtain:

they prefer X– to any hair tonic Q've used
You ... will prefer X– to any hair tonic Q've used

It now appears that by reducing these stretches to their independent elements, the latter sections have become identical. On this basis, the beginning sections of these two sentences are found to have identical environments, and hence to be equivalent. Since the first of these beginning sections was included in our class *P*, we can now include the section *You too ...* in *P* as well.[10]

This is only one kind of dependent occurrence. There are many others which have to be investigated; and the resulting information is of use both to discourse analysis and to a more detailed descriptive linguistics.

One major example is that of the pronouns. If the advertisement had read *You ... will prefer it* instead of *You ... will prefer X–*, we would at first regard *it* as a new element, to be placed in a new equivalence class. However, the occurrence of *it* is dependent on the occurrence of *X–*: if the preceding *X–* had contained the plural morpheme (*X–s*), the pronoun in this sentence would have been *them*. Other words of the *it* group, say *he* or *you*, will not occur here as long as *X–* occurs in the preceding sentence; but they could occur if certain other words were used in place of *X–*. The same is true of words like *this/these*, *who/which*, which also depend on particular words occurring somewhere else in the passage. Without using any information about the meaning of these pronouns, or about their 'referring' to preceding nouns, we can conclude from their distribution in the text that they are not independent elements: they contain a (discontinuous) portion of the occurrence of the morpheme with which they correlate.

Another type of dependent occurrence is found in such expressions of cross reference as *each other* and *together*, which carry out in language some of the functions filled in mathematical expressions by variables—but in the vaguer and more complex way that is characteristic of language. The sentence *Foster and Lorch saw each other at the same moment* is normal; but if we drop the words *and Lorch*, every native speaker of English will immediately replace *each other* by something else. To put it differently: we will not find any sentence that contains *each other* but does not contain either the expression *and Z* or a plural morpheme in the relevant noun. Furthermore, though we will find the sentence *Electrons and positrons attract each other*, we will not find—at least in a physics textbook—the same sentence with the words *and positrons* omitted, unless there are also other changes such as *repel* in place of *attract*.

[10] Before this can be done, some further operations must be carried out to reduce *Four out of five ... say they prefer ...* to two *PW* sequences: *Four ... say ...* and *They prefer ...*, with the sentence *You ... will prefer ...* as a third *PW* sequence. Otherwise, the words *say they* would be left hanging, since the *P* section (equivalent to *Millions*)' is only *Four out of five people in a nationwide survey*, and the corrected *W* section (identical with the *W* of *You ... will prefer ...*) is only *prefer X– to any hair tonic Q've used*. See §3.2 below.

It may be noted that dependent elements are especially prone to be assigned to different equivalence classes in their various occurrences, since each occurrence of them is assigned to the class of whatever element correlates with that particular occurrence. If the text contained *You will prefer X–, You will prefer it, The survey showed, It showed,* the first occurrence of *it* would be assigned to the class of *X–*, the second *it* to the class of *survey.*

In all such cases the special relations of dependent occurrence among particular elements can be eliminated by considering the dependent element to be simply a portion of that element with which it correlates (upon which its occurrence depends). It should be clear that when we speak of dependence, the term is only required to apply within a particular text. The dependence of pronouns or cross-reference words upon some neighboring noun may hold in every text in which these words occur; but the dependence between the two occurrences of *they* or of *you* in our text is peculiar to this text. Elsewhere we may find the sentence *they prefer X– to any hair tonic you've used*; but in this particular text such a sentence does not occur. It is for that reason that in this text we can tell what the second pronoun must be by looking at the first one.

2.32. SUBDIVISIONS OF SENTENCES. The recognition of dependent elements affects our decision concerning the number of intervals into which a particular sentence is to be subdivided.

Where an element has dependent portions spread over a domain, we generally have to consider the whole domain as entering into one interval with that element. For example, in *they prefer X– to any hair tonic they've used* we have established that the two occurrences of *they* are interdependent in this text. Hence we can analyze this section into *they* (occurring over both positions) plus ... *prefer X– to any hair tonic ... 've used*; and similarly for the sentence with *you* (also over both positions). This is a more general treatment that that of §2.31, which gave favored status to the first occurrence of *they* and of *you* by considering the second occurrence to be dependent on the first, and which made the identity of the two sentences in their latter portions depend on their both containing the same kind of dependence (*Q*). The present treatment eliminates dependence by viewing the single *they* or *you* as occurring over two positions, and makes the second parts of the sentences identical without qualification. The effect of this new treatment is that since the two-position *they* stretches over almost the whole length of the second part, the whole of that second part has to be kept in the same interval as *they.* The consolidation of the two occurrences of *they* thus precludes our setting up two intervals here; otherwise we might have set up two intervals: *they prefer ...* , and either *they've used* or *Q've used.*

On the other hand, there are cases where recognition of dependence leads us to distinguish more intervals than we might otherwise. Take the sentence *Casals, who is self-exiled from Spain, stopped performing after the fascist victory.* If we investigate the text in which this is imbedded we will find that the *who* is dependent upon *Casals,* much as the second *they* above is dependent upon the first: the text includes *And the same Casals who ...* , but later *The records which ...* . We may therefore say that the *who* 'contains' *Casals,* i.e. either continues it or repeats it. But which does it do? If *who* continues *Casals,* we have one interval,

the first section (C) being *Casals who*, while the second section (S) is *is self-exiled ... stopped ....* If *who* repeats *Casals* instead of continuing it, we have two intervals, one imbedded in the other: the first consists of *Casals* (again C) plus *stopped performing* (marked S_1), the second of *who* (taken as an equivalent of *Casals*) plus *is self-exiled* (S_2). We would be led to the second choice only if we could show in terms of the text that *is self-exiled ...* and *stopped performing ...* are two separate elements (not just two portions of one long element)—for example, if we found in the text two additional sentences: *The press failed to say why he stopped performing, etc. But he has stated publicly why he is self-exiled, etc.* In either case *who* contains *Casals*. But if the original sentence is *Casals who S*, we analyze it as CS, whereas if (on the basis of the later sentences) we view the original sentence as *Casals who* S_2S_1, we analyze it as CCS_2S_1, and divide it into two intervals CS_2 and CS_1, with the result that S_2 and S_1 are equivalent since they both occur after C. The only difference between taking a dependent element as a continuation and taking it as a repetition is in the number of intervals — one or two — into which we then analyze the total.

We have seen here that when a sentence contains an element A which is dependent upon B, we have the choice of taking the whole sentence as one interval, with A simply a continuation of B, or as two intervals—one containing B and the other containing A in the same class as B. The latter choice will generally be taken if the rest of the sentence can be divided into two comparable sections, one to go with A and the other with B.

Choices of this type can arise even where there are no dependent forms. For example, in our second text we have the further sentence *The self-exiled Casals is waiting across the Pyrenees for the fall of Franco.* We wish to put *self-exiled* in the same class as *is self-exiled ...*, since the same morphemes are involved (provided we can show from the text itself that *self-exiled* is equivalent to *self-exiled from Spain*). This gives us the peculiar sentence structure S_2CS_3, as compared with the previous CS sentences. Now if by good fortune the text also contained the sentence *Casals is waiting across the Pyrenees for the fall of Franco* (which is too much to ask in the way of repetition), we would be in position to make the following analysis. We have as sentences of the text CS_1, C *is* S_2, S_2CS_3, CS_3. The sequences S_1 and S_2 and S_3 are all members of one equivalence class S, since they all occur after C. Our problem lies with the maverick S_2CS_3. Let us now say that any sentence X_1AX_2 can be 'transformed' into A *is* X_1*:* AX_2.[10a] This means that if X_1AX_2 occurs in the text, then A *is* X_1*:* AX_2 also occurs in the text. In that case we will consider X_1AX_2 equivalent to A *is* X_1*:* AX_2; as a new structure our maverick has disappeared. We replace S_2CS_3 by the transformationally equivalent C *is* S_2 and CS_3, both of which occur elsewhere in the same text.

We may proceed on this basis even to transformations which are not already justified by the text, provided they do not conflict with the text. Thus, we find in the text the sentences *The memorable concerts were recorded in Prades ... The*

[10a] In such formulas as A *is* X_1*:* AX_2, the italic colon indicates the end of a sentence or interval. (It is used instead of a period because that might be mistaken for the period at the end of a sentence in the author's exposition.)

concerts were recorded first on tape. We can represent this as MNR_1: NR_2 (the equivalence of R_1 and R_2 being shown, let us suppose, elsewhere in the text), and we would transform the first sentence into *N is M:* NR_1. This does not mean that we claim that our transformation *N is M* (*The concerts were memorable*) actually occurs in the text, or that there is no stylistic or other difference between saying *The memorable concerts were recorded in Prades* and saying *The concerts were memorable: The concerts* (or *They*) *were recorded in Prades.* All that our transformation means is that MNR_1 is taken as equivalent to *N is M:* NR_1 because S_2CS_3 is actually found as an equivalent of *C is* S_2: CS_3, in the sense that both occur in the modified text.

On the one hand, we have eliminated from our tabular arrangement the peculiar interval structure MNR_1 or S_2CS_3—peculiar because the other intervals all have the form NR or CS. On the other hand, we have discovered that M (or rather *is* M) is a member of the R class. But our most important result is that a sentence may be represented as two intervals even when it does not contain two sets of the requisite equivalence classes. This happens when we can show that a single class in the sentence relates independently to two other classes or elements elsewhere. That class is then repeated, once in each interval; and each interval will indicate separately its relation to one of the other classes.[11]

These difficulties in dividing a sentence into intervals arise from questions about the manner in which the equivalence classes relate to each other. In a sentence, the various morphemes or sequences do not merely occur together; they usually have a specific relation to each other which can be expressed by one or more morphemes of order: *You wrote Paul* and *Paul wrote you* differ only in their morphemic order. If we find several CS intervals in our text, that means that C has a particular relation to S—that of occurring with it and before it. Since we are operating without meaning, we do not know what this relation is, but we are careful to represent the same morphemic order in the sentence by the same class order in the interval. Now when we find S_2CS_3, we do not know how this order relates to the order CS, and we can make no comparison of the two sentences. It is therefore desirable to rearrange the unknown S_2CS_3 so that it will contain the same classes in the same order as other intervals—and of course we must show that the rearrangement is equivalent, for this text, to the original. In most cases this can be done only if we break the unknown sentence, by means of such transformations as have been discussed above, into two or more intervals, in such a way that the smaller intervals have a form which occurs in this text.

In this way we get a great number of structurally similar intervals even in a text whose sentences are very different from each other.

2.33. GRAMMATICAL TRANSFORMATIONS. Up to this point we have seen how the structure of a text can be investigated without using any information from outside the text itself. The straightforward procedure is to set up equivalence classes, and to discover patterned (i.e. similar or partly similar) combinations

[11] The case which we have been considering here is the important one of the sequence adjective + noun + verb, in which the noun relates independently to the adjective and to the verb. The adjective can be represented as a predicate of the noun in the same way as the verb. This will be discussed in §2.33 below.

of these classes in successive intervals of the text. Often, however, we get many small classes and dissimilar intervals, because the sentences are so different from each other; when this happens, we find that by comparing the sentences of the text we can sometimes show that one section of one sentence is equivalent (for this text) to a different section of another sentence, and therefore contains the same classes. The extent to which we can do this depends upon the amount of repetition in the text.

We raise now the question of advancing further in the same direction by using information from outside the text. The information will be of the same kind as we have sought inside the text, namely whether one section of a sentence is equivalent to another (in the sense that MNR is equivalent to *N is M: NR*). It will go back to the same basic operation, that of comparing different sentences. And it will serve the same end: to show that two otherwise different sentences contain the same combination of equivalence classes, even though they may contain different combinations of morphemes. What is new is only that we base our equivalence not on a comparison of two sentences in the text, but on a comparison of a sentence in the text with sentences outside the text.

This may seem to be a major departure. One may ask how we know that any equivalence discovered in this way is applicable to our text. The justification was given in §1.3 above: if we can show that two sequences are equivalent in any English sentences in which they occur, then they are equivalent in any text written in English. If in any English sentence containing XAY, the XAY is equivalent to *A is X: AY*, then if we find S_2CS_3 in our English text we can say that it is equivalent to *C is* S_2*:* CS_3.

But what is 'equivalence'? Two ELEMENTS are equivalent if they occur in the same environment within the sentence. Two SENTENCES in a text are equivalent simply if they both occur in the text (unless we discover structural details fine enough to show that two sentences are equivalent only if they occur in similar structural positions in the text). Similarly, two sentences in a language are equivalent if they both occur in the language. In particular, we will say that sentences of the form A are equivalent to sentences of the form B, if for each sentence A we can find a sentence B containing the same morphemes except for differences due to the difference in form between A and B. For example, N_1VN_2 is equivalent to N_2 *is V-en by* N_1 because for any sentence like *Casals plays the cello* we can find a sentence *The cello is played by Casals.*

We do not claim that two equivalent sentences necessarily mean exactly the same thing, or that they are stylistically indifferent. But we do claim that not all sentences are equivalent in this sense: the relation of equivalence is not useless, as it would be if it were true for all sentences. For example, N_1VN_2 is not equivalent to N_1 *is V-en by* N_2, because the latter form will be found only for certain N_1 and N_2 forms (*I saw you* and *I was seen by you*) but not for all forms (we will not find *Casals is played by the cello*).[12] We claim further that the application of

[12] True, one might claim that this last sentence is still 'grammatical'. But present-day grammar does not distinguish among the various members of a morpheme class. Hence to require that sentence B must contain the same morphemes as sentence A is to go beyond grammar in the ordinary sense.

this grammatical equivalence from outside the text will enable us to discover additional similar intervals in our text, beyond what we could get merely from comparing the text sentences with each other. Thus, we can show that in various environments *who, he,* etc. are grammatically equivalent to the preceding noun, and that N_1 *who* V_1V_2 is equivalent to N_1V_2: N_1V_1. Then, in *Casals, who is self-exiled ... stopped performing ...*, we have two intervals CS_1: *C is* S_2. We would have this result (without having to worry whether *Casals who* is one continued occurrence of *C* or two repeated occurrences) even if there were no other occurrences of *who* within the text, i.e. when no analysis could be made of *who* on internal textual grounds. The usefulness of grammatical equivalence is especially great if, for example, we have a number of intervals all containing *Casals,* besides many others interlarded among the first but containing *he,* and if we can find no common textual environments to show that *Casals* and *he* are equivalent. As soon as we accept this equivalence grammatically, we can show that all the environments of *Casals* are equivalent to those of *he*; and this in turn can make other equivalences discoverable textually.

Grammatical equivalence can be investigated more systematically if we introduce a technique of experimental variation. Given a sentence in form *A* and a desired form *B*, we try to alter *A* by only the formal difference that exists between it and *B*, and see what happens then to our *A*. Given *The memorable concerts were recorded ...*, suppose that we want to make this *MNR* sentence comparable in form to previous intervals beginning with *N*. To this end, we seek a variation of the sentence beginning *The concerts*. We may do this by putting an informant into a genuine social speech situation (not a linguistic discussion about speech) in which he would utter a sentence beginning *The concerts* and containing the words *memorable* and *recorded.*[13] Or we may do it by the tedious job of observation, hunting for a sentence that begins with *The concerts* and contains *memorable* and *recorded.* By either method, we might get *The concerts were memorable and were recorded,* or something of the sort,[14] whence we learn that when *M* (or any adjective) is shifted to the other side of *N* (its following noun) one inserts *is*; *MN* is equivalent to *N is M*. In this way we discover that when *MNR* is shifted to a form beginning with *N*, an *is* appears between *N* and the following *M*.

This technique of varying the grammatical form of a sentence while keeping its morphemes constant cannot be used within a text; for there all we can do is to inspect the available material. But it can be used in the language outside the text, where we have the right, as speakers, to create any social situation which might favor another speaker's uttering one rather than another of the many

[13] To give a crude example, one can read the text sentence *The memorable concerts were recorded* in company with an informant, and then stop and say to him, in an expectant and hesitant way, 'That is to say, the concerts——', waiting for him to supply the continuation.

[14] We may find a great many sentences beginning with *The concerts* and containing the other two words, e.g. *The concerts were not memorable but were nevertheless recorded.* These sentences will contain various words in addition to those of the original sentence; but the only new word which will occur in ALL sentences of the desired form *NMR* (or rather in a subclass of the *NMR* sentences) will be a form of the verb *to be*. Hence this is the only new word that is essential when changing to that form.

sentences at his disposal. It is especially useful in a language like English, where so many morphemes occur in various grammatical classes.

The preceding paragraph indicates the basic safeguard in applying grammatical equivalence to extend our textual equivalence classes. We do not merely ask, What sentence-forms are equivalent to MNR? There may be many. We ask instead, Since $N...$ is a common form in this text, and since we find also MNR, can we replace this by an equivalent sentence of the form $N...$? The direction of change is not arbitrary, but comes entirely from the text. As before, it is a matter of dividing our sentences into the most similar intervals possible. All we ask is whether there is a grammatical equivalence which would connect MNR with the form $N...$; the answer is yes, provided an *is* appears in the form. This in turn yields *is* M as equivalent to R. As elsewhere in linguistics, the method does not collapse all sentences into any arbitrary form we choose; it simply enables us to describe the rarer forms of the text (MNR) in terms of the common ones ($N...$).

For analysis purely within the text, all we need to know are the morpheme boundaries. To utilize grammatical equivalences we need to know also the morpheme class to which each morpheme in our text belongs, since grammatical statements concern classes rather than individual morphemes. The grammatical statement in this instance is that adjective + noun is equivalent to noun + *is* + adjective; to apply it to our sequence MN, we must know that the M is an adjective and the N a noun.

It has been found empirically that a relatively small number of grammatical equivalences are called upon, time after time, in reducing the sentences of a text to similar intervals. Hence even a non-linguist can get considerable information about the text by using (in addition to the internal textual method) a prepared list of major grammatical equivalences for the language. Some frequently used equivalences are given here, without any evidence for their validity, and with only a very rough indication of the sentence-environments in which they hold:[15]

(1) If we find XCY, then $X = Y$ (X is equivalent to Y). The C is a conjunction like *and, but, or,* or else, under special circumstances, a phrase like *as well as, rather than, A-er than.* The X and Y must be in the same grammatical class. Thus, in *I phoned him but he was out*, X and Y are each NV; in *I saw it but went on*, the Y is only the verb phrase *went on*, and hence the X can include only the verb phrase *saw it* (not the whole sequence *I saw it*). It follows that $N_1V_1CN_2V_2$ is equivalent to two intervals N_1V_1: N_2V_2, and $NV_1CV_2 = NV_1$: NV_2.

(2) The sequence N_1 *is* N_2 indicates that $N_1 = N_2$. The class of *is* includes *remains* and other verbs.

(3) $\acute{N}_1\grave{N}_2$, with a primary stress on each N, indicates that $N_1 = N_2$; e.g. *The pressure P increases* is equivalent to *The pressure increases* and *P increases.*

(4) NV (*that*) $NV = NV$: NV; e.g. *I telegraphed that we'll arrive tomorrow* is equivalent to *I telegraphed: We'll arrive tomorrow.*

(5) $N_1VN_2 = N_2V^*N_1$, where V and V^* are respectively active and passive, or passive and active.

[15] A for adjective, N for noun, V for verb, P for preposition. Subscripts indicate particular morphemes, regardless of their class.

(6) $N_1PN_2 = N_2P^*N_1$; e.g. (*They seek*) *the goal of certainty* is equivalent to some such form as (*They seek*) *certainty as a goal.* The change in prepositions when two nouns are reversed is far greater than the corresponding change in verbs. In verbs the change is effected simply by adding or subtracting the passive morpheme and the word *by*; in prepositions it is effected by replacing one form by an entirely different form. The pairs of equivalent prepositions are not fixed: between certain nouns, the substitute for *of* may be *as*; between others, it may be *with.* Nevertheless, it is possible to find structures in which the nouns of the sequence N_1PN_2 are reversed.

(7) $N_1PN_2 = A_2N_1$, i.e. the morpheme of the second noun occurs in an adjectival form before the prior noun, as in *medical training* for *training in medicine.*

(8) Pronouns like *he*, and certain words with initial *wh-* and *th-*, repeat a preceding noun. Which noun they repeat (when there are several nouns preceding) depends on the details of the grammatical environment; usually it is the immediately preceding noun, or the last noun that occurs in a comparable grammatical environment. Thus, *who = the man* in *The man who phoned left no name* (N *who* $V_1V_2 = NV_2$: NV_1); *who = my roommate* in *The man spoke to my roommate, who told him to call again* ($N_1V_1N_2$ *who* $V_2 = N_1V_1N_2$: N_2V_2). There are many variant ways of determining which noun is repeated by a pronoun, and which verb belongs with each noun. In *the man who phoned,* no subject can be inerted before *phoned*, hence *who* must be taken as subject. In *The man I phoned was out,* we reduce first to *I phoned: The man was out*; then, since no object can be inserted after *phoned* in the original sentence, we set *the man* as the object[16] of *phoned* and obtain the equivalent *I phoned the man: The man was out* ($N_1N_2V_1V_2 = N_2V_1N_1$: N_1V_2).

(9) NV_1, V_2-*ing* = NV_1: NV_2; e.g. *They escaped, saving nothing* is equivalent to *They escaped: They saved nothing.*

(10) $N_1CN_2VX = N_1VN_2$: N_2VN_1. Here X represents a class of cross-reference expressions like *each other*; e.g. *The Giants and the Dodgers each beat the other twice* is equivalent to *The D beat the G twice: The G beat the D twice.* The equivalence differs somewhat for different groups of X forms.

(11) $ANV = N$ *is* A: NV, as in the example *The self-exiled Casals...* in §2.32. So also $NVAN_1 = NVN_1$ *who is* $A = NVN_1$: N_1 *is* A; e.g. *They read the interdicted books = They read the books which were interdicted = They read the books: The books were interdicted.*

(12) $N_1VN_2PN_3 = N_1VN_2$: N_1VPN_3. That is, a double object can be replaced by two separate objects in two intervals which repeat the subject and verb; e.g. *I bought it: I bought for you* for *I bought it for you.*

These grammatical equivalences preserve the morphemes and the grammatical relations among them, though in a changed grammatical form. We cannot get $N_1VN_2 = N_2VN_1$, because that would change the subject-object relation to the verb; but $N_2V^*N_1$ is obtainable as an equivalent of N_1VN_2 because the verb too is changed here, in a way that preserves its grammatical relation to the now

[16] The only way to express the exclusion of an object here purely in terms of occurrence of elements is to say that the object already occurs. This cannot be *I*, since that is the subject of *phoned*; hence it must be the other N, *the man.*

reversed nouns. Preservation of the grammatical relations is essential, because such relations are always to be found among the morphemes in a sentence. That is to say, there are restrictions of substitutability and order and intonation among the various morphemes (or morpheme classes) in a sentence; and when we move from one sentence to an equivalent sentence, we want upon moving back to the original sentence to get back the same restrictions—since the original, like all sentences, is defined by the restrictions among its parts. Therefore, when we break up a sentence into various intervals for a tabular arrangement, we do not want two combinations of the same equivalence classes (say our first and second *TE* combinations above) to represent different grammatical relations. Accordingly, when we transform a sentence containing certain equivalence classes, we are careful to preserve the original grammatical relations among them.

Sometimes, however, we find sections of a sentence which contain none of our equivalence classes; that is (in the simplest case), they contain no material which recurs elsewhere in the text. The grammatical relation of unique sections to the rest of the sentence must be preserved in our tabular arrangement no less than the relation of recurrent sections; but here we escape the problem of preserving their relation while changing their relative position, since we have no reason to change their position at all: it is only our equivalence classes that we wish to rearrange. All we want of this non-recurrent material is to know its relation to our equivalence classes, and to indicate this relation in our analysis. We may not be able to learn this from a study of our text alone; but we can learn it by bringing in grammatical information or experimental variation. For an example we return to the sequences *Casals, who is self-exiled from Spain* ... and *The self-exiled Casals* ... . If the latter is S_2C, the former is C, *C is* S_2 *from Spain*. Since *from Spain* does not recur, we want only to know where to keep it when we arrange our equivalence classes, i.e. what its relation is to these classes. From the grammar we know that in sentences in the form $NVAPN$ the smallest unit of which PN is an immediate constituent is APN, and that this APN is replaceable by A alone.[17] Therefore, if the A happens to be a member of one of our equivalence classes while the PN is not, we associate the PN with the A in its equivalence column by writing APN instead of A as the member of the class.

More generally, material that does not belong to any equivalence class, but is grammatically tied to a member of some class, is included with that member to form with it an expanded member of the class in question. Thus, *self-exiled from Spain* is now in the same class as *self-exiled*. The justification for this is that since the material does not occur again in this text (or occurs again only in the same grammatical relation to the same equivalence class), its only effect, when the text is represented in terms of particular equivalence classes, is precisely its relation to the particular member to which it is grammatically tied.

An interesting special case arises when two members of the same equivalence class constitute jointly the next larger grammatical unit of their sentence (i.e. are the immediate constituents of that unit), for example when the two are an adjective and a following noun, where $AN = N$. In such a case we may consider that the two together constitute just one member of their class, and fit together

[17] Semantically one would say that the PN 'modifies' the A.

into a single interval. If we took them as two occurrences of their class, we would have to put each occurrence into a separate interval.

Grammatical information is especially useful in the recognition of sentence connectives. These morphemes are easily identified from formal grammar, quite independently of their meaning, but may not be identifiable as such on purely textual evidence. Their importance lies in the fact that many sentences of a text may contain the same classes except for some unassigned words, often at the beginning, which are grammatically connecters or introducers of sentences; they stand outside the specific classes which comprise the sentence or interval. In our tabular arrangement these elements can be assigned, by their grammatical position, to a special front column. We can go beyond this and assign to this front column any material which is not assignable to any of the equivalence columns. Sometimes such connecting material is not immediately obvious; note that many sentences of the form *NV that* N_1V_1 can be analyzed as consisting of the equivalence classes N_1V_1, with the *NV that* relegated to the front column. Consider, for example, *We are proud that these concerts were recorded by our engineers.* Here the known members of equivalence classes are *concerts* and *recorded.* The preceding words do not recur in the text and are not grammatically tied to any particular class member. Quite the contrary, they can be grammatically replaced by introductory adverbs like *indeed,* even though in a purely grammatical sense they are the major subject and predicate of the sentence.

In addition to making use of the grammatical relations of whole grammatical classes, we can use information about the relation of particular morphemes or grammatical subclasses to grammatical classes. For instance, it is possible to establish that intransitive verbs (in some languages) form a subclass which never occurs with an object and which is equivalent to a transitive verb plus an object. In a given text, this may enable us to put a transitive verb with its object in the same column as a comparably placed intransitive verb.

Finally, there are a great many detailed equivalences which apply to particular morphemes. This information is not provided by descriptive linguistics, which deals generally with whole morpheme classes. But it can be obtained by linguistic methods, since it deals with matched occurrences and special restrictions, though in most cases it is necessary to study the restrictions over more than one sentence at a time. Suppose, for example, that we find the words *buy* and *sell* in a text. Their environments in that text may not be sufficiently similar to place them in the same equivalence class, even though it might promote the analysis of the text if we could do so. But if we investigate a good number of other short texts in which the two words occur, we will find that the two often appear in matched environments, and that in certain respects they are distributional inverses of each other; that is, we will find many sequences like N_1 *buys from* N_2*:* N_2 *sells to* N_1 (*I bought it from him at the best price I could get, but he still sold it to me at a profit*). If the environments of *buy* and *sell* in our text are similar to the matched environments of the other short texts, we may be able, by comparison with these wider results, to put the two into one equivalence class in our text after all, or even to analyze one as the inverse of the other.

In this way we can put more words into one textual class than would otherwise

be possible, and we can make use of what would seem to be special semantic connections between words (as between *buy* and *sell*, or even between a transitive verb and the presence of an object) without departing from a purely formal study of occurrences. The reason is that differences in meaning correlate highly with differences in linguistic distribution; and if we have two related words whose distributional similarities cannot be shown within the confines of our text, we will often be able to show them in a larger selection of texts, even of very short ones.

The kind of outside information which has been indicated here has been only sketched in scattered examples, both because the field is vast and because a great deal remains to be done. Further work in this direction will not only be useful to discourse analysis but will also have interest as an extension of descriptive linguistics.

3. Results

3.1. The double array. As a product of discourse analysis we obtain a succession of intervals, each containing certain equivalence classes. For a tabular arrangement we write each interval under the preceding one, with the successive members of each class forming a column, as in §2.14 above. The very brief text of §2.32 is arranged as follows:[17a]

C	S_1	
C	S_2	(S_2 after C is *is* S_2)
C	S_2	(= S_2C without the *is*)
C	S_3	
N	R_0	(= MN; R_0 = *is* M)
N	R_1	
N	R_2	

The horizontal rows show the equivalence classes present in each interval, arranged according to their order (or other relation) within the interval. The vertical columns indicate the particular members of each class which appear in the successive intervals. Material which is a member of no equivalence class, but is grammatically tied to a particular member of some class, is included with that member in its column; thus *in Spain* is included in the first S_2. Material which is a member of no equivalence class, and is not grammatically tied to a particular member of some class, is placed in a front column (not illustrated here), which will be found to include morphemes that relate the sentences or intervals to each other, or mark some change in several classes of a single interval. The tabular arrangement thus represents the original one-dimensional text in a two-dimensional array, where each element has two coordinates: one horizontal, in respect

[17a] The array given here represents the following sentences, taken from a review of some recent phonograph records: *Casals, who is self-exiled from Spain, stopped performing after the fascist victory ... The self-exiled Casals is waiting across the Pyrenees for the fall of Franco ... The memorable concerts were recorded in Prades ... The concerts were recorded first on tape.* (The other sentences analyzed in §2.32 were composed by me for comparison with these.) The sentences do not represent a continuous portion of the text. This fact limits very materially the relevance of the double array; but that does not concern us here, since the array is intended only as an example of how such arrangements are set up.

to the other elements of its interval, and one vertical, in respect to the other members of its class.

This double array can be viewed as representing the whole text, since every morpheme of the text is assigned to one class or another in the array, and since the array preserves the relations among the morphemes. Even when a large number of textual and grammatical transformations have been carried out, the classes and their members are defined at each step in such a way that the text can always be reproduced from the array plus the full definition of the classes in it. The individual intervals in the array may not be 'idiomatic'—that is, they may not naturally occur in speech. But the preservation of idiom is not one of the requirements of our method. All we ask is that the succession of intervals should be textually and grammatically equivalent to the original text. Although the array may suggest a critique or a possible improvement of the text, it is not meant to be used instead of the original.

The double array can also be viewed as indicating the purely distributional relations among the equivalence classes which figure in it. From this viewpoint we can operate upon the tabular arrangement and investigate its properties. We can develop ways of simplifying the array, for example by drawing out common elements, or by grouping together larger sets of equivalent sequences than we used in the formation of the array. We can learn how to accommodate various special cases, such as a mobile class which appears in close relation now with one class now with another, or which appears a different number of times in various intervals. We can try to regularize or 'normalize' the array by matching all the intervals, so as to establish a single 'normal' interval with which all the actual intervals can be compared: for instance, given an interval from which one of the classes is absent, we can try to transform it into one that includes all the classes, preserving equivalence during the transformation. We can attempt to formulate a general statement covering the changes in successive members of a class as we go down a column, in an effort to 'explain' or 'predict' the particular form taken by the classes of each interval—that is, to derive the successive intervals from the normal form.

All such operations with the array have the effect of isolating the most general independent elements in terms of which we can describe the text (ultimately the horizontal and vertical axes), and of bringing out their relations to each other in the text. In this sense all such operations are but further refinements of our initial procedures.

3.2. Findings. Various conclusions can be drawn about a particular text or type of text by studying the properties of its double array, either directly or in its most simplified forms. Many of these conclusions may well have been obtainable intuitively without such formal analysis; but intuition does not yield results that are either explicit or rigorous. In some respects, moreover, the complexity and size of the material make it impossible for us to draw all the relevant conclusions without painstaking formal analysis. The sample texts used in the present paper have been necessarily too short and too simple to show what kind of conclusions the analysis yields about a particular text or style; that must be left for a future presentation of a longer sample text, though the details of method and

the range of conclusions obtainable by means of it could be shown only through the analysis of a great many discourses. To give some slight idea of these conclusions, we will complete here the analysis of our first text (§2.2).

The analysis was left at the following point: P has as members *Millions*, *Four out of five people in a nationwide survey*, *You too will*, (*and*) *your whole family will*. W has as members *can't be wrong*, *prefer X– to any hair tonic ... 've used*. Four of the sentences (including the title) are represented by five PW intervals.

At this point it is difficult to proceed without recourse to grammatical equivalence (see fn. 10 above). In *Four out of five ... say they prefer...* we have P and W but with *say they* intervening. If our text happened to contain *they* and *four out of five ...* in equivalent environments, we could analyze this sentence directly. In the absence of this, we appeal to the grammatical equivalence of *they* with the preceding comparably-situated noun: *four out of five ...* as subject of *say*, parallel to *they* as subject of *prefer*. We therefore put *they* into the same class P as *four out of five*. Then the sentence becomes P *say* PW, which is analyzed as two intervals P *say:* PW, on the basis of the formula NV (*that*) $NV = NV$: NV; and on this basis *say* is a member of W, since it occurs after P to make a whole interval.

We now turn to the last sentence: *You too will be satisfied*. The first part is a known P; hence *be satisfied* is included in W. This gives us a start for working on the preceding sentence, *Every year we sell more bottles of X– to satisfied consumers*. Now *X– to satisfied consumers* is grammatically *X– to AN*, which is equivalent to *X– to N: N is A*. In this way we obtain an interval *consumers are satisfied*; and since the second part of this is W, we place *consumers* in P. The rest of the sentence contains new classes: Since *bottles* occurs elsewhere in the text, we regard it as representing a possible equivalence class and mark it B; with this occurrence of B we associate the word *more*, which does not occur elsewhere and which is grammatically tied to *bottles*. Since *sell* occurs elsewhere in *sold* (= *sell* + part of the passive morpheme), we mark it S; and we associate with it *every year*, which is grammatically tied to it. (*Every year* is similar in only one morpheme to *since ... years ago* in the first sentence; rather than try to get these phrases into new equivalence classes, we note that each is tied to the member of S that occurs near it, and we associate each phrase with its member of S). There remains *we*, which is not grammatically part of either the B phrase or the S phrase; even though it seems not to occur again, we place it tentatively in a new class I. (We will see below that a zero form of I may be said to occur in the first sentence.) Thus we get *ISB to P*. This in turn can be somewhat simplified, since it is grammatically equivalent to *ISB: IS to P*.

Finally there is the first sentence, *Millions of consumer bottles of X– have been sold since its introduction a few years ago*. If we start with *Millions* as a known P, we obtain an unanalyzable remainder beginning with *of*. Instead, we match *bottles of X– have been sold* with *we sell bottles of X–*. The first has the form N_1V; the second is N_2VN_1. Grammatically, *have been sold* is *sell* + past + passive; hence if we take *sell* as V, then *been sold* is V^*. Grammatically also, V + passive + *by* N is equivalent to V + passive alone (*is sold by us* = *is sold*). Hence the lack of any *by us* after *sold* does not prevent our matching the two clauses. To

we sell bottles as N_2VN_1 we match *bottles have been sold* as $N_1V^* = N_1V^*N_2$; we can even say that the passive morpheme, with or without the following 'agent' ($by + N$) is equivalent to the subject of the active verb (i.e. the verb without the passive morpheme). If *we sell bottles of X–* is *ISB*, then *bottles of X– have been sold* is the equivalent BS^*I with zero I. The section *since ... years ago* we associate with the preceding S^*, as also the past-tense morpheme, since neither of these figures elsewhere in our equivalence classes. *Millions* and *consumer* are both members of P,[18] but there is no way of making use of this fact. Grammatically, *consumer bottles* is $N_1N_2 = N_2$, and *millions of* N_2 is $N_3PN_2 = N_2$, so that the whole sequence is grammatically tied to *bottles* (as *more* was tied to *bottles* above), leaving the sentence as BS^*I. This means that there are two occurrences of P words which are lost by being included in an occurrence of B. There is no other distributional relation that this *Millions* and this *consumer* have to any other class occurrence in the text (except their analogy to *more*); hence there is no way of including them in the double array. The same morphemes indeed occur elsewhere as P, but in different relations to other classes.

This points up the confusing relation of the title to the first sentence. If we start with the title, we come upon *Millions* in the first sentence and assign it to P, on the basis of the title, only to find that there is no class P in the final analysis of the sentence. (The millions who can't be wrong turn out to be bottles.[19]) If we begin with the body of the advertisement, we have a class P (*four out of five*; *you*) which relates to W, and a class B (*bottles*, *millions of ... bottles*) which relates to S; and if we then proceed to the title, we find there the W preceded not by any known P word or by a new word which we can assign to P, but by a word which has elsewhere been associated with a member of B. (The bottles show up as people.) This is the formal finding which parallels what one might have said as a semantic critique—namely, that the text of the advertisement (millions of bottles sold; many people can't be wrong in preferring X–) fails to support the title (millions can't be wrong).

The double array for the advertisement is not interesting in itself:

PW	*Millions of People Can't Be Wrong!*
BS^*I (the B containing pseudo-P)	*Millions of consumer bottles ... have been sold ...*
CPW	*And four out of five people ... say*
PW	*they prefer X– ...*
PW	*Four out of five people ... can't be wrong.*
PW	*You too will prefer X– ...*
PW	*your whole family will prefer X– ...*

[18] We have *consumers* in P; and since the singular-plural distinction does not figure in our classes, we can associate the dropping of the *-s* with the occurrence of *consumers* in the first sentence. By dropping the *-s* from the P-element *consumers* we get a P-form *consumer* for the sentence.

[19] Since *millions of consumers* would be a natural English phrase (P_1 *of* $P_2 = P_2$), the effect of using the almost identical sequence *millions of consumer* in front of *bottles* is to give a preliminary impression that the sentence is talking about P; but when one reaches the word *bottles* one sees that the subject of the sentence is B, with the P words only adjectival to B.

*B S*I (= ISB)*	*Every year we sell more bottles of X–*
*S*I to P*	*we sell to consumers*
P W	*consumers are satisfied*
P W	*You too will be satisfied!*

3.3 Interpretations. The formal findings of this kind of analysis do more than state the distribution of classes, or the structure of intervals, or even the distribution of interval types. They can also reveal peculiarities within the structure, relative to the rest of the structure. They can show in what respects certain structures are similar or dissimilar to others. They can lead to a great many statements about the text.

All this, however, is still distinct from an INTERPRETATION of the findings, which must take the meanings of the morphemes into consideration, and ask what the author was about when he produced the text. Such interpretation is obviously quite separate from the formal findings, although it may follow closely in the directions which the formal findings indicate.

Even the formal findings can lead to results of broader interest than that of the text alone. The investigation of various types of textual structure can show correlations with the person or the situation of its origin, entirely without reference to the meanings of the morphemes. It can also show what are the inherent or the removable weaknesses (from some given point of view) of a particular type of structure. It can find the same kinds of structure present in different texts, and may even show how a particular type of structure can serve new texts or non-linguistic material.

Finally, such investigation performs the important task of indicating what additional intervals can be joined to the text without changing its structure. It is often possible to show that if, to the various combinations of classes that are found in the existing intervals of the text, we add intervals with certain new combinations of classes, the description of the textual structure becomes simpler, and exceptions are removed (provided we leave intact any intrinsic exceptions, such as boundary conditions). The adding of such intervals may regularize the text from the point of view of discourse analysis. If for example our text contains *AB: AC: ZB*, we may say that *Z* is secondarily equivalent to *A*, since both occur before *B*, but only *A* before *C*. If there are no textually intrinsic exceptions governing this restriction on *Z*, we can on this basis add the interval *ZC* to the text. In this extended text the equivalence $A = Z$ is now a matter of complete substitutability in an identical range of environments, rather than just the secondary result of a chain of equivalences. The addition of such intervals has a very different standing from the addition of arbitrary intervals to the text. If we want to know what is implied but not explicitly stated in a given text, or if we want to see what more can be derived from a given text than the author has already included, this search for adjoinable intervals becomes important.

4. SUMMARY

Discourse analysis performs the following operations upon any single connected text. It collects those elements (or sequences of elements) which have identical or equivalent environments of other elements within a sentence, and con-

siders these to be equivalent to each other (i.e. members of the same equivalence class). Material which does not belong to any equivalence class is associated with the class member to which it is grammatically most closely tied. The sentences of the text are divided into intervals, each a succession of equivalence classes, in such a way that each resulting interval is maximally similar in its class composition to other intervals of the text. The succession of intervals is then investigated for the distribution of classes which it exhibits, in particular for the patterning of class occurrence.

The operations make no use of any knowledge concerning the meaning of the morphemes or the intent or conditions of the author. They require only a knowledge of morpheme boundaries, including sentence junctures and other morphemic intonations (or punctuation). Application of these operations can be furthered by making use of grammatical equivalences (or individual morpheme occurrence relations) from the language as a whole, or from the linguistic body of which the given text is a part. In that case it is necessary to know the grammatical class of the various morphemes of the text.

Discourse analysis yields considerable information about the structure of a text or a type of text, and about the role that each element plays in such a structure. Descriptive linguistics, on the other hand, tells only the role that each element plays in the structure of its sentence. Discourse analysis tells, in addition, how a discourse can be constructed to meet various specifications, just as descriptive linguistics builds up sophistication about the ways in which linguistic systems can be constructed to meet various specifications. It also yields information about stretches of speech longer than one sentence; thus it turns out that while there are relations among successive sentences, these are not visible in sentence structure (in terms of what is subject and what is predicate, or the like), but in the pattern of occurrence of equivalence classes through successive sentences.