Before entering into the direct analysis of the level structure of language, we must study the crucial question of determining "possible" or "grammatical" sentences. This will occupy us in the next chapter.

Appendix to Chapter IV: Formal Construction of the Language of Grammar [omitted here].

GRAMMATICALNESS

The first problem that the linguist must face in constructing the grammar of a language is that of determining the subject matter of his description. Given a corpus of sentences, this problem breaks down into two parts, as we have already seen in §13. First, he must determine which of these utterances are phonemically distinct. Second, he must determine which utterances, whether in the corpus or not, are grammatical, hence to be described in the grammar to be generated in the manner discussed in Chapter IV. We have suggested that the first problem can be met by such devices as the paired utterance test (cf. §13.3); the second will concern us now.

It is clear that the set of grammatical sentences cannot be identified with the linguist's corpus of observed sentences. Not only are there many (in fact, infinitely many) nonobserved grammatical sentences, but, in addition, certain sentences of the corpus may be ruled out as ungrammatical, e.g., as slips of the tongue. Thus we must project the class of observed sentences to a larger, in fact, infinite class of grammatical sentences. And within linguistic theory we must define "grammatical sentence" in terms of "actual, observed sentence."

Investigating the conditions that we want this definition to meet, we find that a partition of utterances into just two classes, grammatical and nongrammatical, will not be sufficient to permit the construction of adequate grammars in terms of what we have broadly described as distributional analysis. If we wish to distinguish between two elements X

---

1 We will use the word "project" in a broad sense, it being understood that the projection of a class to a new class may drop some of the members of the original class. In the particular kind of projection that we are discussing in this chapter, the deletions will necessarily be of a minor and peripheral character. But the projection being discussed here is only a special case, a first stage in a series of projections. And in later stages (e.g., on the morphological level), the projection may be primarily deletion (cf. §47).
and Y in distributional terms, we must be able to exhibit a significant class K of contexts such that either X or Y (but not both) occurs in the contexts of K to form grammatical sentences. A difficulty arises when X occurs regularly in K, and occurrence of Y in K is somehow "near-grammatical." If Y is excluded from K, we lose the grounds for distinguishing these near-grammatical sequences from completely ungrammatical nonsense, while if occurrence of Y in K is ruled grammatical, we lose the distributional grounds for distinguishing X from Y.

This problem arises, for example, in the case of various subclasses of nouns in English. Clearly, any adequate grammar of English would have to distinguish proper nouns like "Jones," "Bill," etc., on the one hand, from "sincerity," "golf," etc. This requires a discriminatory class K of contexts, including, presumably such contexts as

1. (a) — admires Tom
   (b) — had lunch with Tom yesterday

in which "Jones" occurs to form grammatical sentences. But now consider the sentences formed by placing "sincerity" or "golf" in the blanks of 1. If these are considered grammatical, we lose the distributional grounds for distinguishing the class of proper nouns. But if they are ruled ungrammatical (as in some sense they surely are), we will have failed to indicate that these sentences are by no means as remote from English as, e.g., "the admires Tom," or "of had lunch with Tom."

To choose a somewhat different sort of example, an adequate grammar would certainly have to register the fact that the relation between "bring" and "brought" is the same as that between "like" and "liked." We might try to show this on the basis of such contexts as

2. they — it today
3. they — it yesterday

Clearly, "like" and "bring" occur in 2, and "liked" and "brought" in 3. But even if we agree that it is proper to exclude "like" and "bring" from 3, we face the further difficulty that "break" can occur in 2 and "brought" in 3. How, then, do we establish that the proper pairing is "bring"—"brought," "break"—"broke" rather than "bring"—"broke."

"break"—"brought." The most reasonable distributional grounds for the correct pairing would seem to be that "bring," "brought" can occur in such contexts as 4, and "break," "broke" in such contexts as 5:

4. wars — disease and famine
5. they — into the store

But now consider the class of sentences formed by placing "break," "broke" in the contexts 4, etc., and "bring," "brought" in the contexts 5, etc. Clearly these are somehow "more grammatical" than the sequences formed by placing, e.g., "Jones" in these contexts. On the other hand, if they are considered fully grammatical, we lose the basis for making a correct distributional analysis of the verbs in question (and, at the same time, we admit some rather odd sentences into the language).

This leads us to observe that an adequate linguistic theory will have to recognize degrees of grammaticality, so that substitution of "Jones" in 1, "bring" in 4, and "break" in 5 gives fully grammatical sentences; substitution of "sincerity" in 1, "break" in 4, and "bring" in 5 gives partially grammatical sentences; and substitution of "brought" in 1 and "Jones" in 4, 5 gives completely ungrammatical sequences.

We can approach this matter in somewhat different terms. A primary motivation for this study is the remarkable ability of any speaker of a language to produce utterances which are new both to him and to other speakers, but which are immediately recognizable as sentences of the language. We would like to reconstruct this ability within linguistic theory by developing a method of analysis that will enable us to abstract from a corpus of sentences a certain structural pattern, and to construct, from the old materials, new sentences conforming to this pattern, just as the speaker does.4

But we must be quite careful in determining just what we are to reconstruct. In the only presystematic terms we have (cf. §§13.5, 15), we can say that we are trying to explicate that intuitive concept of linguistic form that enables a speaker to distinguish such grammatical sentences as "colorless green ideas sleep furiously" from such non-grammatical sequences as "furiously sleep ideas green colorless."

The particular theory that we develop will determine what qualifies as a significant class.

---

8 We might try to argue this on the grounds that the correct pairing minimizes the number of consonant changes. But, on the other hand, we might argue that the incorrect pairing gives the parallelism break—brought, take—taught. √√√\downarrow\leftarrow\downarrow\downarrow\downarrow\downarrow\downarrow.4

4 See Chapter I. We will try to attach more content to such rather figurative modes of expression below.
But more generally, there is little doubt that speakers can fairly consistently order new utterances, never previously heard, with respect to their degree of "belongingness" to the language, as in the cases discussed above. Similarly, the following sequences may all be new in English:

6  
(a) look at the cross-eyed elephant  
(b) look at the cross-eyed kindness  
(c) look at the cross-eyed from

but I think it is clear that any one would arrange them in this order with respect to "belongingness" to English. In other words, we might say that the speaker projects his past linguistic experience to a set of more and more comprehensive extensions. Correspondingly, within linguistic theory we must develop a notion of "degree of conformity to the structural pattern," so that, given a reasonable sample of English not containing 6, we would be able to predict that 6a is perfectly grammatical, 6b partially so, and 6c not grammatical at all.

Thus from several points of view, we are led to consider as the goal of our reconstruction the notion of "degree of grammaticality."

33.1 The notion of "syntactic category" provides a first approach to the reconstruction of this process of projection. Let us assume that we have a finite corpus of sentences with word division marked. It might contain, for example,

7  
(a) John came  
(b) Bill ate  
(c) John saw Bill

We assign these words to categories. Let us call this assignment a syntactic analysis of the words of the language. We can now associate with each sequence of words a sequence of categories, replacing each word by the category to which it belongs. Thus if we assign "John," "Bill" to the category N, and "came," "ate," "saw" to the category V, we will have NV, NV, and NVN corresponding, respectively, to 7a, 7b, and 7c. There will in general be many fewer sequences of categories than sequences of words. Each sequence of categories may be called a sentence form, and we can construe the generated language of grammaticality (but perhaps nonoccurring) sentences as those which conform to one of these sentence forms. In this example, for instance, "John ate," "Bill saw John," etc., would be sentences, perhaps not in the corpus,

but conforming to the sentence forms NV, NVN constructed from the syntactic analysis.

The notion of degree of conformity to the structural pattern is easily derived. Instead of considering only one syntactic analysis of words into categories, we consider several analyses of various orders. That is, we have broad syntactic categories like Noun and Verb, subcategories of these, subclasses of categories, etc. Thus we might have a first order syntactic analysis into several hundred categories, a second-order analysis into perhaps fifty categories, etc., down to an analysis into a single category. For each order of syntactic analysis we have a set of sentence forms in terms of the categories of that order. For an order with many categories, hence small categories, we have sentence forms that are more selective and generate few sequences. For an order with fewer categories, hence large categories, we have sentence forms that generate a great many sequences. A nonoccurring sequence then has a higher degree of grammaticality if it conforms to the more selective sentence forms stated in terms of a many-category analysis.

Referring to the previous example, suppose that

6  
(d) look at the cross-eyed man

does occur in the corpus. We see that 6a has a high degree of grammaticality, since "man" and "elephant" are presumably co-members of the small category of Animate Common Noun. 6b is less grammatical, since "man" and "kindness" are co-members of no syntactic category smaller than Noun, and 6c is still less grammatical, since the only syntactic category containing both "man" and "from" is presumably the class of all words.

More precisely, we have a system \( \mathcal{C} \) of classes of words,

\[ \mathcal{C} = \{ C_n \} \]

where

(i) \( 1 \leq n \leq N; 1 \leq n_i \leq a_n \)  
(ii) \( a_1 > a_2 > \cdots > a_N \)  
(iii) \( w \in C_{n_i} = w \) is a word\(^5\)  
(iv) \( C_{n_i} \) is non-null; \( C_{n_i} \subseteq C_{n_j} \)  

There are many other conditions that we can put on \( \mathcal{C} \), but this will suffice for our immediate discussion.

\(^5\) When we fit this discussion directly into our system of levels, this stipulation will mean that \( w \) corresponds to a prime of a certain level \( W \).
For each $n$, the set $\mathcal{C}^n = \{C^a_1, ..., C^a_n\}$ will be called the syntactic analysis of order $n$, and the $C^a_n$ are called categories of order $n$.

We are given a corpus $K$ which we take to be a set of strings of words. A sequence $C^a_1, ..., C^a_n$ of categories of order $n$ is said to generate the word string $w_1 \sim ... \sim w_n$ if $w_i \in C^a_n, ..., w_n \in C^a_n$. Thus the set of word sequences generated by a category sequence is essentially, the Cartesian product of the categories.

A sequence of categories of order $n$ is said to be a grammatical sentence form of order $n$ if one of the word sequences generated by it is in $K$.

The sentences of order $n$ are those generated by the grammatical sentence forms of order $n$. Thus the highest-degree grammatical sentences are those of order 1, the order with the largest number of categories. If $a_n = 1$, then all sequences (of the length of members of $K$) are grammatical of order $N$.

34.1 Our immediate problem, then, is to construct a formal and abstract notion of syntactic category that will lead to an appropriate system of categories when we are presented with an adequate sample of linguistic material. Or, to put differently, we must develop a procedure such that given a corpus of sentences with word division marked (i.e., a set $K$, as in §33.2), we can, by means of this procedure, construct a system of classes $\mathcal{C}$ which will give the correct account of grammaticalness.

The most obvious approach to this problem seems to be some sort of substitution technique. The very nature of our goal dictates that these categories $\mathcal{C}_f$ be classes of elements that are, in some sense, mutually substitutable. But attempts to clarify this sense run up against serious difficulties. We will now briefly consider some of these difficulties and the possibilities of elaborating a direct substitution technique in such a way as to avoid them. Then, in §35, we will discuss in greater detail a somewhat different approach to the problem.

34.2 Consider first the problem of constructing the smallest categories, the categories $C^a_1, ..., C^a_n$ of order 1. This is the problem of defining highest-degree grammaticalness. The use of a substitution technique for this construction faces two immediate difficulties. In any sample of linguistic material, no two words can be expected to have exactly the same set of contexts. On the other hand, many words which should be in different categories will have some context in common. There are contexts like

9 it is ———

where elements of many different categories can appear (e.g., “John,” “red,” “mine,” “likely,” “up,” “here,” “not”). Thus substitution is either too narrow, if we require complete mutual substitutability for co-membership in a syntactic category $\mathcal{C}_f$, or too broad, if we require only that some context be shared. Thus neither of the simplest approaches to substitution is adequate without elaboration.

Though we can scarcely hope to find two words mutually substitutable everywhere in a linguistic corpus, we might find some behavioral test for mutual substitutability, so that we could run through our (finite) corpus testing every pair of words for this property in each context. Then the $\mathcal{C}_f$ could be defined as classes of words among whose members complete mutual substitutability holds.

The unacceptability of this approach lies in the fact that it simply avoids the very question with which we are concerned. An operational test for grammaticalness would simply record the speaker’s ability to project, but a definition of “grammatical sentence” in terms of rules of combination set up for observed utterances would give some insight into the nature and source of this ability, and as such, it would represent a fundamental contribution to our understanding of linguistic behavior. Replacing a systematic account of grammaticalness by a behavioral test would mean abandoning this goal, which is both important and, it seems, quite possible to attain.

To put it differently, this approach amounts to taking first-order grammaticalness as a primitive notion in linguistic theory. The purpose of a theory is to investigate to the fullest extent the relations between the notions that enter into the theory. Complete success in unearthing and expressing these interrelations is marked by the elimination of terms from the primitive basis. The discovery of an effective behavioral test for first-order grammaticalness, then, should be taken as posing (rather than solving) the problem of accounting for this fact (now presumably behaviorally marked) of first-order projection (cf. Note 6, Chapter II).

Even if such contexts as 9 can be excluded as illegitimate on some grounds, the difficulties faced by a substitution technique are not at an end. It is evident that substitutability is directly relevant only to the establishment of the smallest categories. For words which belongs to the same higher-order (i.e., larger and less selective) category, e.g., the nouns “horse” and “justice,” it would be difficult to find any natural sentence in which they can replace one another.

For this reason, it is necessary to develop the methods of substitution somewhat further. One way to do this is as follows. If we find two words in the same legitimate context, we assign them to the same category.
We now rewrite the corpus, replacing words assigned to the same category by a single symbol everywhere they occur. In other words, we disregard the distinction between words assigned to the same category. This reduces the number of distinct contexts, so that new pairs of words become mutually substitutable, and we can build larger categories. We can thus proceed to analyze the corpus in a constructive manner, using the information obtained at each step in determining substitutability relations for the succeeding step. Such a procedure will lead to a system of categories from which the system $\mathcal{G}$ may be selected. Although the problem of defining grammaticality is thus not completely solved, given such a technique, it is tremendously reduced in scope.

It seems that a substitution technique must be at least as powerful as this if it is to be effective. Extension to include substitutability of phrases (i.e., sequences of words) is possible, leading to a generalization of the notion of syntactic category to cover at least a partial analysis of phrase structure. One important result of this extension is that now the projection of the corpus can include sentences longer than those of the original corpus. We may be able to develop the infinite generative power required for a full account of grammaticality by continually repeating the analysis for the successive projections of the corpus.

The extension to include substitution of word sequences raises a further problem of segmentation. We cannot freely allow substitution of word sequences for one another. Thus we cannot set up a category containing "from New York was here" and "left," even though both occur in the context "my friend ——." Thus it may be necessary not only to exclude certain contexts as illegitimate, but also to rule out certain sequences as improper substituends.

34.4 A final difficulty for a substitution technique is the problem of homonyms. Such words as "will," /twu/, ("to," ("too," ("two"), /riw/ ("read," ("reed")), etc., are best understood as belonging simultaneously to several categories of the same order. Hence the technique of substitution must be designed in such a way as to lead to a system $\mathcal{G}$ with the categories of a given order overlapping in the syntactic homonyms.

The problem of homonymity is an important one. In preparation for later discussions of homonymity, we note now that only certain cases of homonymity are relevant to grammar, namely those that can be interpreted syntactically by assignment of the words in question to the overlap of syntactic categories of some order. Thus "will," /twu/, and /riw/ are instances of syntactic homonyms in this sense, while /sa/ ("son," "sun") or "will" in its various nominal uses are perhaps not. To a certain extent, this distinction is relative to the subtilty of our grammatical analysis. Thus a more refined account of highest-order grammaticality may convert certain cases of apparent purely semantic homonymy into cases with syntactic correlates.

Among the syntactic homonyms, there is one further distinction of importance. In the overlap of Noun and Verb, we find both "walk" and /riw/, but clearly these are very different kinds of homonyms. In the first case, we are inclined to say that we have to do with a single word which is in two categories, while in the second, we are more inclined to describe this as a case of two words, one a noun, one a verb, with the same phonemic shape. This distinction might be felt to be a wholly semantic one. We cannot deal with it at this stage in our investigations, but below, we will suggest syntactic grounds for it (§107. Chapter X).

It is apparent that substitution procedures require for their success that we make an intelligent choice of contexts and substituends, passing over such contexts as 9 and beginning our investigation of distribution with nonhominoms. There are many quite obvious suggestions for a precise and orderly specification of this element of "intelligent choice," and with the increasing availability of high-speed computers, it may be possible to put some of them to the test.

A somewhat less "constructive" (and perhaps more promising) way to approach this complex of problems is through the analysis of clustering. We define the distribution of a word as the set of contexts in which it occurs, and the distributional distance between two words as the number of contexts they share divided by the total number in which either can occur. More generally, we can define the cohesion of the class of words $w_1, \ldots, w_n,$ with distributions $D_1, \ldots, D_n,$ as the cardinal number of the logical product of $D_1, \ldots, D_n$ divided by that of the logical sum.

We are interested in choosing categories in such a way as to maximize the amount of mutual substitutability among the words belonging to the given category. Let us call the quantity that we would like to maximize the cluster value of the category. We can determine the cluster value of a category as a function of the cohesions of its subsets. There are various ways in which this can be done, each making precise a certain sense of the vague notion of mutual substitutability. Suppose we fix on one such

* For a detailed introductory study of the topics touched on in this section, and certain related problems (e.g., investigation of the equivalence of alternative formulations of such substitution techniques for various kinds of language), see Chomsky, "Systems of syntactic analysis."

* See Harris, Method, for discussion of this problem.
function. We might then define the \textit{joint cluster value} of a set of categories as a function of the cluster values of the categories in this set, e.g., their average. Given such a definition we would choose the set of categories as the set with the highest \textit{joint cluster value}.

This only gives us the best analysis into \( n \) categories for each \( n \). Thus we still have the problem of selecting a system \( C \) from this system of classes by choosing the \( n \)'s for which the \( n \)-class analysis is to be selected as a syntactic order. This might be done, for instance, by considering, for each \( n \), what the best joint cluster value of an \( n \)-class analysis would be if words were distributed randomly in sentences, and choosing those \( n \)'s for which the deviation of the joint cluster value from this value is above a certain amount, or is at a relative maximum, etc.

If we can develop an effective notion of clustering of sets, this can be used either to establish a system \( C \) directly, or to provide a basis for application of a system of the kind outlined in \( \S 34.3 \). We might use cluster analysis to exclude contexts like 9 (these being characterized by the fact that elements belonging to several clusters appear in them) and to determine homonyms (as elements belonging to several clusters).

This outline can be filled in with formal detail in several different ways. We would be led to favor one of these formulations if it were shown to have particularly interesting formal properties or (and this is the ultimate test) the correct empirical consequences. A certain amount of formal investigation has not led to any conclusive reason for choosing one of several formulations, and it would clearly require a tremendous amount of data to present even the most fragmentary empirical validation. For these reasons, I will not go on to present in detail any one of the various ways of realizing the program just outlined. Nevertheless, the study of this complex of notions seems to be of considerable importance and promise for distributional analysis, and despite its difficulties and uncertainties, it should certainly be pursued further.\footnote{A special case of this problem, the case of clustering of points on a line, has been investigated by A. W. Holt in an unpublished MIT master's thesis (mathematics).}

\[ 34.6 \]
The line of reasoning that underlies the discussion of \( \S 34.5 \) can be generalized beyond substitution techniques. Given any distributional property \( \varphi \) of words (or words and phrases), we can attempt to set up syntactic categories on the basis of \( \varphi \) by defining (i) the distance between words (or the cohesion of a set of words), (ii) the cluster value of a category of words, and (iii) the joint cluster value of a set of categories, all in terms of \( \varphi \). We have been considering the total distribution of a word in the corpus (unweighted by frequencies) to be the relevant distributional characteristic \( \varphi \). But there are many other distributional characteristics which are reasonable candidates for this type of category analysis. For example, we can define the \textit{gap} between two word tokens, \( x \) and \( y \), as the number of words occurring between them, and we can study the frequency distribution of gaps throughout the corpus for each ordered pair of words (here, word types). Fitting such a study into our framework, we note that for each word \( w \), there is a set of such frequency distributions, two for each word \( w_i \), ..., \( w_n \) in a corpus of \( n \) distinct word types. We can take this set of distributions (or certain of them) as the characteristic \( \varphi \), and can then attempt to define distance between two words in terms of the pairwise similarity between the associated gap distributions.

There are many other distributionally defined properties that deserve consideration here. Thus we might consider morphological criteria as being particularly crucial, or we might assign a special status to words of higher frequency like "the," "of," etc. There is no logical necessity to limit ourselves to one criterion. Some combination of these or other distributional properties may be selected as \( \varphi \).

The correct choice of a distributional characteristic \( \varphi \) as the basic datum for category analysis will be determined, ultimately, by the empirical consequences of that choice. But whatever choice is made, it seems that the line of reasoning sketched in (i)-(iii) above, or something quite similar, will have to be followed, in order to determine the empirical consequences. The investigation of \( \varphi \) alone is of limited interest, in this connection, until we state how a set of categories is to be derived from it.

The purpose of the rather unorganized remarks about substitution techniques and the like has been to indicate briefly the kind of problem that must be faced by such procedures, and to suggest that there is a large variety of ways of attacking these problems that have not been sufficiently studied. However, it is not surprising that such a technique should be so elusive and difficult to obtain.

If we could formulate a substitution procedure properly, we would have succeeded in meeting the strongest requirement for correspondence between the general theory and particular grammars (cf. \( \S 10.3 \)). The approaches we have considered so far aim to provide a practical and mechanical discovery procedure for syntactic categories. In other words, in each case the intention was to enable the syntactic categories to be constructed directly from the raw data by observation of some simple distributional property.

\footnote{Following a suggestion of V. Yngve, "Gap analysis and syntax."}
In view of the difficulties involved in carrying through such an ambitious program, I think it can prove interesting to lower our aims to the weaker correspondence between theory and particular grammars, and to try to construct a definition of syntactic category that begins not with an easily observable distributional characteristic of words, but with some measurable characteristic of completed syntactic solutions; that is, a definition that merely enables us to assign a value, say a number, to each proposed analysis, and thus to decide mechanically between two proposed analyses, with no concern as to how these analyses were arrived at. In accordance with this weaker aim, I would like to sketch a conception of syntactic category that seems to undercut many of the difficulties cited.

35.1 Suppose that we consider once again the method given in §33 for generating sentences once we have a set of categories of a certain order.¹⁰ We rewrite each sentence of the corpus as a sequence of categories, replacing each word by the category to which it belongs. This gives a set of sentence forms, and we may now generate all sentences of these forms. This gives a great many new sentences, since along with each original sentence we now have all sentences of the same form as this original sentence, whether or not they appeared in the corpus.

Let us suppose now, for the sake of simplicity of exposition, that all the sentences of the corpus are of the same length. We also assume, for the moment, that we are discussing only a given fixed order of syntactic categories, with a preassigned number of categories. We return to the latter assumption in §35.2 and to the former in §35.3.

Suppose that the categories of a proposed syntactic analysis are set up on the basis of complete mutual substitutability. That is, two words are members of the same category only if in the original sample, each word occurs in every position in which the other occurs. It is clear that for such a syntactic analysis, no sentence can be generated if it did not already appear in the original sample. Thus no new sentences are generated.

In general, it is clear that more new sentences will be generated to the extent that the distributions of the elements within a category differ. The number of sentences generated by an analysis thus gives some measure of the extent to which elements in the same category have similar distributions, and it seems reasonable to measure the value of

¹⁰ For many of the ideas of this section, I am indebted to Peter Elias. In the case of non-overlapping categories, the measure described for evaluating an analysis is the same as what Harwood calls “measure of negative fit” of an analysis to a corpus; “Axiomatic syntax; the construction and evaluation of a syntactic calculus.”

a syntactic analysis by the number of sentences that it generates, fewer sentences being generated by a better analysis. Thus we evaluate an analysis by seeing how good an approximation it gives to the original corpus—how few sentences it generates beyond those of the original corpus.

The foremost problem faced by a substitution technique was seen to be the difficulty of deciding how many contexts must be shared for two elements to be in the same category. Such questions are avoided here, since this technique does not build categories step by step, but rather provides a procedure for evaluating a completed solution. Elements may be in the same category in the highest-valued analysis even if they share no context. This property is important, since in actual linguistic material, the selectional restrictions on distribution are extremely heavy, and literal substitutivity is distinctly the exception rather than the rule. Nevertheless different nouns do substitute for each other in the sense that they all occur with some verb, though rarely with the same verb. Similarly, individual verbs are substitutable in contexts defined by the categories Noun, Adjective, etc., though rarely in the context of particular nouns and adjectives. This approach, as distinct from a substitution procedure, permits us to use this fact by, as it were, setting up these classes simultaneously.

Consider now the problem of homonyms. This is essentially the problem of when to put a word into two or more of the categories of the syntactic analysis. If a word is put into more than one category, there is always a loss in the value of the analysis in one respect. To see this, note that each time a certain category appears in a grammatical sentence form, a set of sentences is generated for each word in that category. Hence the more words in a category, the more sentences are generated, and the lower the value of the syntactic analysis. When a word is put into a second category, this second category now has an extra member, and it follows that more sentences are generated wherever this second category occurs, with a corresponding drop in the value of the analysis.

On the other hand, if the element is a real homonym, there may be a compensating saving in the following way. Consider an English homonym like /tʊ/ (“to,” “two,” “too”). If this word is put only into the category of prepositions, then, since /tʊ/ can appear in such sentences as: “there are two books on the table,” it follows that all prepositions will occur in the numeral position in the generated language. But if /tʊ/ is put in both the preposition and numeral categories, then a given occurrence of the word can be classed either as a numeral or as a preposition. Since the category of numerals will occur anyway in “there are — books on the table,” no new sentence forms are
generated if this occurrence of /tw/ is classed as a numeral, and there
is consequently a considerable saving in the number of generated
sentences. We see that assignment of a word to several categories may
increase the value of the analysis. This suggests a way of deciding when
to consider a word to be in fact a set of homonyms. We do this if the
loss incurred in assigning it to several categories is more than offset by
the gain; and there is always a numerical answer to this.

It must be shown, of course, that in terms of presystematic criteria,
the solution of the homonym problem given by this approach is the
correct one. Certain preliminary empirical investigations of this have
been hopeful, but the task of properly validating this (or any other)
conception of "syntactic category" is of course an immense one.

35.2 We can now determine, for each n, the best analysis in terms of n
categories. As n increases, the categories become smaller, and projection
is more limited and selective. Thus the degree of grammaticalness of
the projected sentences will be higher. Where n is the number of words
in the corpus, the set of sentence forms is exactly the corpus itself, and
no new sentences are generated. Where n = 1, there is only one sentence
form, but every possible sequence is generated in terms of it. The number
of sentences generated is thus a nonincreasing function of the number
of categories.

However, we still have the problem of selecting a system S from
this set of analyses. This is the problem of determining for which n
we actually set up the n-category analysis as an order of the system S,
that is, a set \( S^n \) (cf. §33.2). Our aim here is to select a certain n such that
the n-category analysis compares very favorably with the \((n - 1)\)-
category analysis, but is not much worse than the \((n + 1)\)-category
analysis. In other words, we are interested in minimizing a certain
function of n and the number of sentences generated by the n-category
analysis. At this point we can only speculate about which function
should be chosen for minimization. There are several possible can-
didates, and at this point there seems to be no compelling reason for
making a choice one way or another. This decision turns upon the
empirical consequences of the various choices, and we simply do not
have the requisite data at this stage of our knowledge.11 But it seems
reasonable to hope that this is no defect in principle, and that the proper

Perhaps it is worth emphasizing once again that the unavailability of such data is not
a reason for judging such theorizing as we have been engaged in to be pointless. We
cannot know what kind of data to collect until we have a theory that offers some hope
of solving the difficulties that we know in advance to exist. The purpose of these
remarks, of course, is to suggest such a theory (cf. Chapter I).

kind of empirical investigation may lead directly to a decision, thus
filling in the remaining gap.

Whatever function we do minimize, it must be remembered that we
are interested in relative minima, since we would like to construct
certain orders of analysis. We will see below (§39), however, that the
absolute minimum, and the dichotomous partition that it imposes on
sequences, may have a special significance.

We have seen how, given a set of sentences of fixed length \( \lambda \), we can
determine, for each n, the best analysis into n categories of the words of
which these sentences are composed. It would seem reasonable to assume
that for any n, \( \lambda_1, \lambda_2, ... \), the optimal n-category analysis for sentences
of length \( \lambda_1 \) and of length \( \lambda_2 \) will be identical, since we would hardly expect
the basic principles of sentence construction to vary markedly from one
sentence length to another. If this assumption is granted, then investiga-
tion of a single sentence length will be sufficient for determination of
the system \( S \).

Alternatively, we may consider the break between sentences to be
a "word" assigned to a special category \( \$ \), and we define a "discourse
form" as any sequence of sufficiently long12 sentence forms, with sentence
break marked.13 We can then measure the value of an analysis in terms of
the number of discourses it generates (generation of discourses from
discourse forms being exactly like generation of sentences from sentence
forms).

More precisely, we define an "initial discourse form" as the
beginning of any discourse form (with a discourse form as a special
case), and we let \( S_1 \) be the number of strings generated by a given
syntactic analysis from the set of all initial discourse forms of length \( \lambda \).14
The natural way to define the value of a given syntactic analysis \( A \) is as

\[
10 \quad \text{Val}(A) = \lim_{\lambda \to \infty} \frac{\log S_1}{\lambda}
\]

We choose as the best analysis into n categories that analysis \( A \) for which
\( \text{Val}(A) \) is minimal.

11 \( \lambda_1 \) and \( \lambda_2 \) being greater than some fixed \( \lambda_0 \). That is, we will not investigate the very
shortest sentences, where little grammatical structure appears, in determining \( S \).

12 Our theory has not yet been extended to account for prosodic features. A more complete
account would no doubt identify sentence breaks with intonation morphemes belonging
to a special category.

13 Since there is a finite upper bound to sentence length in our present discussion, it turns
35.4 This conception of syntactic analysis has an information-theoretic interpretation; in fact, it was initially suggested and motivated by this interpretation. We have in fact defined the best analysis as the one that minimizes the information per word in the generated language of grammatical discourses. We have to do here with a very special and elementary case of information, since the frequency of words and word sequences is nowhere considered.

An elaboration of this interpretation may prove illuminating. By the redundancy of language is meant, essentially, the restriction on the freedom of the choice of elements in discourse, and in our present context, it can be understood as a measure of restriction on the freedom of choice of words. We might picture this redundancy as being broken down into two factors, the first involving the restrictions provided by the grammatical structure of the language, and the second, those provided by all other factors, including the content of discourse and all its extra-grammatical concomitants. In other words, at every point in the stream of discourse the speaker must choose a particular single word, and it makes sense to ask to what extent his choice of a particular word was governed by the grammatical structure of the language, and to what extent it was governed by other factors. The more rigid the grammatical structure, the fewer discourses are permissible altogether (for each length), and the larger the share of the constraints contributed by the grammatical structure. Essentially, the conception of syntactic analysis given above has been designed in such a way as to minimize the number of possible discourses of each length, consistent in a special sense with the corpus, and thus to maximize the contribution of the formal grammatical structure to the total redundancy. As we move to lower, less selective degrees of grammaticalness, this contribution decreases. Even for highest-degree grammaticalness, we should expect it to be relatively slight.

36.1 This interpretation for the proposed constructions focuses attention on a characteristic feature of the linguist's ordinary conception of grammar. I have in mind the sharp distinction maintained between grammatical and statistical structure. In view of recent interest in statistical methods in linguistics, it seems important to give a somewhat more systematic statement of this distinction and its consequences, even at the cost of some repetition.

Customarily, the linguist carrying out grammatical analysis disregards all questions of frequency and simply notes the occurrence or nonoccurrence of each element in each context in his observed materials. A consequence of this approach is that the resulting grammar sets up a sharp division between a class $G$ of grammatical sentences and a class $G'$ of ungrammatical sequences. The formal properties of language might be studied in other ways. Instead of noting merely occurrence and nonoccurrence, we might present a statistical analysis of the corpus, tabulating the probability of occurrence of each element in each context or the conditional probability of occurrence of each element as the $n$th element of a sequence, given the first $n-1$ elements, etc.

The grammatical approach thus contrasts with a statistical approach that leads to an ordering of sequences from more to less probable, rather than a sharp division into two classes within which no such gradations are marked. This literally correct statement of two different approaches can be misleading. It would be easy to picture the grammatical approach as an attempt, motivated by the complexity of the statistical data, to impose a rough approximation to the full statistical variation, with all sequences of higher than a certain probability being assigned to $G$ and all others to $G'$. But this would be a gross misconception. We have already noted that if our theory is to begin to satisfy the demands that led to its construction, then $G$ will have to include such sentences as 11, while such sentences as 12 are assigned to $G'$.

11 colorless green ideas sleep furiously

12 furiously sleep ideas green colorless

But clearly these strings are not distinguished by their assigned probabilities. If probability is to be based on an estimate of frequency in some English corpus, then this probability will be zero in both cases. Nor can they be distinguished, in some more sophisticated way, in terms of the probability of their parts. The full statistical picture is not a direct generalization of the grammatical analysis with its simple yes-no system of constraints. There is no obvious tie-up between the two approaches. If we somehow rank sequences of English words in terms of their probability, we will find grammatical sentences scattered freely throughout the list. The grammatical approach cannot be interpreted

---

13 For simplicity of exposition, we will temporarily disregard the notion of degree of grammaticalness.

14 For example, in terms of order of approximation (cf. Note 36, Chapter II).
as giving a schematized and simplified description of the full variety of the "actual" language. Nor can the generalization to degrees of grammaticalness be understood as simply a closer approximation to this variety.

This is a simple but important point, and failure to appreciate it has occasionally led to serious misunderstanding of the nature of grammar. The linguist uses such words as "pattern" and "structure" quite freely in describing his own activities. He says that he is interested in describing the structure of the language, the pattern to which its utterances conform. The distinction between two kinds of nonsense, grammatical nonsense like II and ungrammatical nonsense like 12, can serve as a simple illustration of the significance of this reference to pattern and structure. Here we have two sequences of words, no part of which may ever have occurred in connected discourse. Yet any speaker of English will recognize at once that 11 is an absurd English sentence, while 12 is no English sentence at all, and he will consequently give the normal intonation pattern of an English sentence to 11 but not to 12 (cf. above, §13.5). Such examples as this give empirical content to the linguist's search for pattern and structure. The distinction between grammatical and ungrammatical nonsense cannot be explained by simply giving a more and more detailed description of observed linguistic behavior, ultimately, let us say, a tabulation of the probability of occurrence of each item in each context. In terms of such a description alone, both II and 12 will be excluded as equally remote from observed English. This distinction can be made (in this case, but not in many others that will concern us) by demonstrating that II is an instance of the sentence form Adjective-Adjective-Noun-Verb-Adverb, which is grammatical by virtue of such sentences as

13 revolutionary new ideas appear infrequently

that might well occur in normal English.

36.2 The custom of calling \( G \) the class of "possible" sentences, or those that "can occur," is no doubt responsible for much confusion here. It is natural to understand "possible" as "highly probable," and "impossible" as "highly improbable." When this interpretation is rejected, as it obviously must be, it becomes equally natural to take the next step of rejecting the notion "possible sentence" as mere mysticism.

17 See, for example, Hockett, A Manual of Phonology, pp. 3–17.

18 More properly, an absurd semi-English sentence, when we have set up degrees of grammaticalness.

Actually, although the notion of grammaticalness is undoubtedly complex and difficult to reconstruct, it is by no means mystical, and we have a good idea as to how to go about reconstructing it. Given a corpus of sentences, we define the set \( G \) to be the set of sentences conforming to the rules established for describing this corpus, whether or not these sentences happen to occur in the corpus. The problem of constructing \( G \), then, is the problem of determining how to provide a proper description for a fixed linguistic corpus—it is the problem of constructing a linguistic theory as we have several times described this project above. Linguistic theory must provide us with the system of formal structures that can be realized in language and with a procedure for evaluating any proposed realization of this system based on a given corpus. To construct such a theory is no mean task, but it is important to recognize that there is no difficulty in principle.

The system \( G \) is one such structure that can be given an explicit interpretation, given an adequate corpus, and in §35 we have suggested one way in which any interpretation might be evaluated. Describing a corpus in terms of \( G \) automatically produces a certain projection of the corpus. Further projection will be discussed below in terms of other structures. Whether or not any of our explicit proposals prove ultimately to be adequate, they do indicate that there is nothing mysterious about the project.

We have frequently noted that the problems of projection and phonemic distinctness are twin aspects of the problem of determining the subject matter of grammatical description. Such goals as that of distinguishing between grammatical and ungrammatical nonsense serve as a principle of relevance for linguistic description in that they determine the degree of detail to which it is necessary to analyze the corpus in the study of grammatical structure. Similarly, the paired utterance test (cf. §13.3, above) offers a principle of relevance on the phonemic level. There is no limit to the detail in which it is possible to characterize the phonetic shape of sounds, and such study may be perfectly proper. But it is also perfectly in order to draw the line just at the point where differences fail to be significant in the sense provided by the paired utterance test. Phonemic theory is developed by drawing the line just at that point.

Though we have strong reasons for a nonstatistical conception of the form of grammar, it might turn out to be the case that statistical considerations are relevant to establishing, e.g., the absolute, nonstatistical distinction between \( G \) and \( G \) (cf. §36.1). As mentioned in §34.6, the relevant distributional criterion \( g \) may turn out to be statistical in nature. There is no a priori way to determine whether the extradistributional
information utilized by a statistical approach to grammaticalness will prove essential, or whether it simply blurs important distinctions with irrelevant detail. At the present stage of our knowledge we must surely keep an open mind on this matter.\(^9\)

37 The notion of level of grammaticalness has some further implications that might be explored with profit. If we drop a certain sentence from the corpus, and apply the analysis to the corpus minus this sentence, we would ordinarily expect that this sentence will be generated at the highest degree of grammaticalness (i.e., by generation in terms of first-order categories). But for certain sequences, this will not be the case. Suppose, for instance, that a certain sequence of the corpus is a slip of the tongue, or is an interrupted sentence, or the like. Then if it is struck out of the corpus, it will not be reintroduced by the process of generation at any level of grammaticalness at all, above the lowest. Or consider a sentence like

14 misery loves company

This may be the only sentence of the form Abstract Noun-Verb1-Abstract Noun, where Verb1 is a certain class of verbs that occur otherwise only in such contexts as Proper Noun—Abstract Noun, etc. If 14 is dropped out of the corpus, then it will not be reintroduced at the highest level of grammaticalness, but only at some lower level, i.e., at the level at which "misery" and "John" are in the same category, since "John loves company" will surely be generated at the highest level. This suggests that we need not consider all occurring sentences as of the highest degree of grammaticalness just because they occur. Above, we

\(^9\) Note the similarity between this discussion of statistical approaches to grammaticalness and the discussion of semantic approaches in §13. In both cases we have to deal with positions that are often ardently maintained, though never carefully formulated. In both cases, our attempt to formulate them seems to show that they are quite beside the point. We must, of course, remain open to the possibility that there is some more significant formulation.

Note that there is no question being raised here as to the legitimacy of a probabilistic study of language, just as the legitimacy of the study of meaning was in no way brought into question when we pointed out (§13.7) that projection cannot be defined in semantic terms. Whether or not the statistical study of language can contribute to grammar, it surely can be justified on quite independent grounds. These three approaches to language (grammatical, semantic, statistical) are independently important. In particular, none of them requires for its justification that it lead to solutions for problems which arise from pursuing one of the other approaches. Nevertheless, these three approaches are in some way related. The object of investigation is ultimately the same, and ultimately, we might expect them to fall into place in some larger semantic theory.

were concerned with assigning highest-degree grammaticalness to certain nonoccurring sequences; now we have a way to assign some lower degree of grammaticalness to certain occurring sequences. The method is to strike them out of the corpus, redo the analysis on the reduced corpus, and see at what point the eliminated sentences are reintroduced. More generally, if a certain sentence form is inadequately represented, in some sense that must be defined precisely, we can drop it and investigate the level at which its instances are regenerated. Though this account is oversimplified, it points out the possibility that certain idioms or metaphors\(^10\) might be characterizable as sentences which occur, but are not of the highest degree of grammaticalness, and that mistakes might be characterizable as occurring sentences of the lowest degree of grammaticalness. In this way we may be able to develop a method of projection of the kind originally discussed in §31.

We see that in terms of the system \(\mathcal{S}\), such sentences as 14 have a special and exceptional status. They belong to sentence forms that are quite inadequately represented. \(\mathcal{S}\) is just one of the systems in terms of which we describe linguistic structure. We will see below (§117) that 14 has an exceptional status in the light of higher-level structures as well. We will also find other sources, on higher levels, for cases of semi-grammaticalness of a different sort.\(^11\)

\(^10\) And, for that matter, many other sentences. Partially grammatical sentences play a role in discourse and often have an important literary effect. For example, consider Veblen's phrase "perform leisure" or "conformable individuals." Such locations are not infrequent in the writings of certain authors. A recent tendency within philosophy has been to seek the source of philosophical perplexity and error in bad grammatical analogies. Here too, the statements criticized often appear to be semigrammatical.

\(^11\) Note that "conformable individuals," in the preceding footnote, is of a different type. Note that when we call a sentence "partially grammatical" we are not excluding it from consideration or declaring it meaningless. We will consider the grammar of a language \(L\) to be a device that generates the highest-degree grammatical sentences of \(L\), but if we have a system \(\mathcal{S}\) as a linguistic level, it will be possible to recover semigrammatical sentences from the grammar.

A familiar problem in linguistics, similar in many ways to that posed by semigrammatical but occurring sentences, is the problem of determining "analytic norms" (cf. Hockett, Manual, and my review of this book). An attempt to construct discovery procedures for grammar is faced by the difficulty that it must deal in a neutral manner with the total linguistic behavior of the informant, including slips, slurred speech, interrupted utterances, etc. A more limited approach will be satisfied with a grammatical description of a partially hypothetical language underlying actual speech in the sense that actual linguistic behavior can easily be characterized as a special deviation from underlying norms. In general, phonemic analysis is the study of fairly slow speech. It is possible to characterize rapid speech as a "blurred" variant of this, though the opposite procedure is out of the question. Similarly, interrupted fragments, semigrammatical statements,
The proposed analysis of the notion of syntactic category was introduced in §34.7 as having a more modest aim than what we have called a "procedural" formulation, since it provides a method for evaluating a proposed analysis, but not for arriving at the correct analysis directly. But this is not literally correct. Given a finite corpus, there is a finite (but astronomical) number of ways in which the words can be arranged into n classes, and the procedure of systematically running through these, evaluating each one, and choosing the best is of course a terminating procedure. But it is clear that this is not the sense in which we speak of a procedural definition or technique (cf. §§10.3, 25). However, the difference is not easy to characterize. It is not the difference between finite and infinite, but the more elusive difference between too large and not too large.

Clearly the definition we have constructed can never in fact provide a procedure for discovering the correct grammar, though it can provide a practical procedure for evaluating a given proposal. In this respect, it is much like the measure of simplicity discussed in Chapter IV. These constructions meet the requirements for linguistic theory laid down in §10. Nevertheless, it is interesting to investigate the possibility of actually constructing a stronger theory, i.e., of converting this evaluation procedure into a practical discovery procedure. A combination of the ideas of §§34, 35 might be useful to this end. The procedures discussed in §§34.3, 34.4 can be applied directly to data to provide a complex system of classes from which a system of can be selected in various ways. Applying the evaluation procedure to these various proposals, we can select the best of them. Even though the substitution procedure will not lead directly to the system of, it may reduce significantly the number of alternative analyses that have to be evaluated. Hence if we do have an effective evaluation procedure, it becomes quite important to develop substitution procedures (or other procedures of the general type outlined in §34.6) even if these prove to be only partially effective in themselves.

Of course the method of §35 cannot guarantee uniqueness of the best analysis in terms of n categories, unless we lay down certain formal restrictions on the sets of symbol sequences that are eligible for consideration as a linguistic corpus. The problem of stating the formal requirements that symbolic systems must meet to qualify as "language" is an important one. In a sense, it is the goal of linguistic theory to solve exactly this problem.

etc., can be explained in terms of the underlying constructed norms. This more hopeful approach is open to us once we have lowered our aims from the construction of discovery procedures to the development of procedures of evaluation in terms of simplicity, etc.

Once we have constructed a system of for a given language, we can give a relative sense to the expression "and Y have the same grammatical form," where X and Y are sequences of words of the same length. In this case, we can say that X and Y have the same order grammatical form if the lowest order of categories in terms of which they are instances of the same sentence form is the order n. Thus any two sentences of the same length have the same grammatical form on some order, since at least they are instances of the same sentence form in terms of the one-category analysis of (cf. §33.2). It will also be necessary to give an absolute sense to the expression "same grammatical form." The reasons for this are already evident from our discussion of grammatical and ungrammatical nonsense in §35, and further reasons will appear below, in succeeding chapters.

In §35.2 we suggested that the n's for which an n-category analysis is to be constructed as an order of can be selected by determining the relative minima of a certain function (which we left unspecified, pending further empirical investigation) of the number of categories and the number of sentences generated. We might then take the absolute minimum of the function as defining the absolute order of grammaticalness and the absolute categories, as we will refer to them below.

Speculating, it would seem reasonable to suppose that a proper choice of will give as the absolute order a set of fairly large classes, so that the absolute order will not correspond to one of the higher degrees of grammaticalness. The absolute analysis embodies the major grammatical restrictions. Presumably these will be stated in terms of such classes as Noun, Verb, Preposition, etc. There will be many further grammatical restrictions that have to do with limited and special contexts, and that will presumably, be reflected in superior degrees of grammaticalness (i.e., smaller, lower-order categories). These further restrictions correspond in part to what Harris has called selection. Thus selectional restrictions can be defined as those which refer to an account of grammaticalness which is more detailed and specific than that provided by the absolute analysis. Although Preposition may well turn out to be a class of the absolute analysis for English, there will be subclass of prepositions that occur with different nouns and verbs, etc., and at the first order, we may even find that although many of the categories are still quite large, the categories of prepositions may be extremely small, even unit classes.

Below we will find it necessary to make assumptions about absolute

---

8 Further specification is necessary if we cannot assume uniqueness for this.
9 See Harris, "Co-occurrence and transformation in linguistic structure," for discussion of many points related to the notion "degree of grammaticalness."
categories that are not warranted by any empirical evidence. This is 
unfortunate, and naturally we will try to reduce them to a minimum. But 
we cannot drop all further investigation into linguistic structure until 
all problems connected with grammaticalness are completely resolved. 
One reason why this is impossible is that problems of grammaticalness 
are not independent of later considerations. The analysis of grammaticalness 
must be carried out in such a way as to relate correctly to more 
advanced constructions, and we cannot know how extensive the description 
of this first step in determining grammaticalness must be until we 
have some idea about what can be accomplished on higher levels of 
analysis (cf. Preface). Whenever we make assumptions about the absolute 
analysis, these can be understood as conditions on the function \( f \). 
That is, not only will \( f \) have to be designed so as to give the correct 
relative minima, but also so as to give the correct absolute minimum. 
As is likely with any assumptions, investigation may reveal that these 
are not valid, i.e., that these conditions on \( f \) should be removed. This 
must be kept in mind when we suggest reliance on absolute categories 
below. Linguistic theory is a complex system with many interconnections 
between its parts. It is apparently necessary at this early stage in its 
development to let speculation outrun available evidence at many points 
in the theory, so that the data obtainable at other points may fit into some 
reasonable conceptual framework—these data then supplying conditions 
of adequacy which the underlying conceptual framework must meet.

Throughout this discussion we have been assuming that projection is 
determined by rules of combination framed in terms of words. But 
this is certainly a debatable assumption. There is nothing in our 
formulation of syntactic analysis that explicitly requires that words 
and not morphemes be considered as the fundamental units for this 
first, basic step in projection. However, the choice of words seems to 
me to be the correct one.

We need not require that the projection given by the system \( \mathcal{S} \) 
be final and immutable. If later considerations in terms of higher levels 
show that adjustment of the boundaries of the set of grammatical 
sentences eliminates complexities and irregularities, then such adjustment 
may be permissible. But it is easy to fall into circularity if we build 
into the theory a possibility for very radical readjustment at later levels. 
To reduce the danger of emptiness and circularity, we certainly try 
to choose as the elements categorized in forming the system \( \mathcal{S} \) those 
elements that are least likely to be revised and altered in the light of 
later syntactic considerations, and that occur as independently as possible 
in the stream of discourse. It seems apparent that these considerations

rule in favor or words rather than morphemes. Morphemes may be 
continuous or discontinuous, they may undergo complex contextually 
determined alternation, they are often subject to rather special restrictions 
on occurrence,\(^{24}\) and, as we will see below, choice of a morphemic 
analysis is heavily dependent on considerations of much higher levels. 
Words are relatively free from these difficulties.

The method of §35 cannot furnish a complete answer to the problem of 
projecting the corpus to a set of grammatical sentences. For one thing, 
this method does not generate sentences longer than those of the original 
corpus. But the class of grammatical sentences is infinite—there is no 
longest sentence. There are many other grammatical possibilities and 
restrictions that clearly cannot be adequately characterized in terms of 
syntactic categories. At best, then, the proposed account of grammaticalness 
represents only the first stage of projection. But an analysis of 
this account brings out the general character of each stage in the 
construction of the set of grammatical sentences.

At each stage of this construction we are presented with a set of 
sentences, and we are required to project this set to a new set, where 
"projection" is to be understood in the broad sense of Note 1. The 
process of projection is intimately bound up with the notion of linguistic 
level. We can picture each new level as being constructed in order to 
simplify the description of the sentences already generated in terms of 
the preceding levels (cf. §14.4). At each stage of linguistic analysis, we 
find that the description of the presented grammatical sentences can 
be simplified if we project this set to a new set, adding many new 
sentences and perhaps dropping certain sentences. Thus the generation 
of new sentences becomes an automatic consequence of the process 
of describing already given sentences in terms of the descriptive 
machinery available in the new level. Putting this a little differently, 
we might say that each level provides a certain point of view from which 
to investigate the structure of the presented set \( S \) of sentences already 
generated. Investigating \( S \) from the new vantage point offered by this 
higher level, we discover that the structure underlying the set \( S \) is only 
partially realized in this set—there are many gaps, and certain exceptions. 
By filling in the gaps with new sentences, and dropping the exceptions,

\(^{24}\) Thus investigation of Harris's results ("From phoneme to morpheme") on isolation of 
tentative segments in terms of independence seems to show that word boundaries are 
more much more clearly indicated than morpheme boundaries, and that it might be possible 
to determine words directly from the phonemic record. The prospects for morphemes 
may much more doubtful. See Hockett, "Problems of morphemic analysis," and 
Harris, Methods, Chapters 12, 13, for discussion of these problems.
we project $S$ to a new set $S^*$, which serves as the presented set for description in terms of the next level.

In the foregoing discussion of the first stage of projection, the set $S$ of presented sentences was the corpus itself. Introducing the level of syntactic categories (the system $\mathcal{C}$), we find that the corpus can be studied as a set of instances of a relatively small number of sentence forms. But the set of sentence forms is imperfectly realized in the corpus. We discard certain inadequately represented forms (in the manner of §37), and we form new sentences conforming to the adequately represented forms, thus projecting the corpus to a set $S^*$ of sentences of the highest degree (first order) of grammaticality (and incidentally, we presumably unearth a good deal of further information about lower-degree, partial grammaticality and absolute grammaticality). This set $S^*$ serves as the basis for study in terms of higher levels. We need not be concerned about the fact that partially grammatical sentences are not discussed on other levels. They have not been totally excluded from the grammar, and once we have the system $\mathcal{C}$, they can be derived from the set of highest-degree grammatical sentences.

In subsequent chapters, we will investigate this set $S^*$ of highest-degree grammatical sentences from the point of view of phrase structure and transformational structure. In each case, we see that the process of projection outlined above is repeated in terms of the descriptive potentialities of these higher linguistic levels.

41.2 This account of projection emphasizes the implications for the order of descriptive operations of the point of view we have adopted. It suggests that after phonemic analysis, the first step in grammar construction is the placing of word boundaries, and the second step is the limited projection provided by the method of §35 or something similar, i.e., by description in terms of the system $\mathcal{C}$ of syntactic categories of words. We then find that the relation between words and phonemes, as well as between sequences of words and phrase sequences, is simplified tremendously by morphological analysis of words. In this view, then, morphology appears as a higher level of analysis. Morphological representations are provided for grammatical sequences of words in such a way as to simplify the description of these sequences and the statement of their relations to phoneme and phrase sequences.

This outline of an order of descriptive operations should not, however, be taken literally. We are not attempting to present procedures for the grammatical analysis of languages, but only to discuss characteristics of a completed description. See §10.3 for further discussion of this point.

In carrying further the study of grammaticality and grammatical structure, it is necessary to make certain assumptions about this first stage of grammaticality. Some of these have such strong intuitive support that we can take it to be a criterion of adequacy for any proposed analysis of grammaticality that these assumptions be verified. As far as possible, we will try to develop our analysis of English in subsequent chapters on the basis of such assumptions.