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 undertstood 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.

out that as λ increases, the contribution of sentence fragments to S_λ diminishes, so that in the limiting case we are still measuring only the contribution of complete sentences, as is required.

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 $\bar{G}$ of ungrammatical sequences.[15] 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 nth 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 $\bar{G}$. 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 $\bar{G}$

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,[16] we will find grammatical sentences scattered freely throughout the list. The grammatical approach cannot be interpreted

[15] For simplicity of exposition, we will temporarily disregard the notion of degree of grammaticalness.

[16] 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.[17] 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 **11** 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,[18] 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 **11** 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 **11** 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 difficulty 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 $\mathscr{C}$ 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 $\mathscr{C}$ 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 at just that point. 3

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 $\bar{G}$ (cf. §36.1). As mentioned in §34.6, the relevant distributional criterion φ 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.[19]

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-*$Verb_k$*-Abstract Noun*, where $Verb_k$ 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 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[20] 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 $\mathscr{C}$, such sentences as **14** have a special and exceptional status. They belong to sentence forms that are quite inadequately represented. $\mathscr{C}$ 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.[21]

[19] 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 semiotic theory.

[20] 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 locutions 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.

[21] 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 $\mathscr{C}$ 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,