INTRODUCTION

Syntax is the study of the principles and processes by which sentences are constructed in particular languages. Syntactic investigation of a given language has as its goal the construction of a grammar that can be viewed as a device of some sort for producing the sentences of the language under analysis. More generally, linguists must be concerned with the problem of determining the fundamental underlying properties of successful grammars. The ultimate outcome of these investigations should be a theory of linguistic structure in which the descriptive devices utilized in particular grammars are presented and studied abstractly, with no specific reference to particular languages. One function of this theory is to provide a general method for selecting a grammar for each language, given a corpus of sentences of this language.

The central notion in linguistic theory is that of "linguistic level." A linguistic level, such as phonemics, morphology, phrase structure, is essentially a set of descriptive devices that are made available for the construction of grammars; it constitutes a certain method for representing utterances. We can determine the adequacy of a linguistic theory by developing rigorously and precisely the form of grammar corresponding to the set of levels contained within this theory, and then investigating the possibility of constructing simple and revealing grammars of this form for natural languages. We shall study several different conceptions of linguistic structure in this manner, considering a succession of linguistic levels of increasing complexity which correspond to more and more powerful modes of grammatical description; and we shall attempt to show that linguistic theory must contain at least these levels if it is to

provide, in particular, a satisfactory grammar of English. Finally, we shall suggest that this purely formal investigation of the structure of language has certain interesting implications for semantic studies.[1]

---

[1] The motivation for the particular orientation of the research reported here is discussed below in § 6.

THE INDEPENDENCE OF GRAMMAR

2.1   From now on I will consider a *language* to be a set (finite or infinite) of sentences, each finite in length and constructed out of a finite set of elements. All natural languages in their spoken or written form are languages in this sense, since each natural language has a finite number of phonemes (or letters in its alphabet) and each sentence is representable as a finite sequence of these phonemes (or letters), though there are infinitely many sentences. Similarly, the set of 'sentences' of some formalized system of mathematics can be considered a language. The fundamental aim in the linguistic analysis of a language L is to separate the *grammatical* sequences which are the sentences of L from the *ungrammatical* sequences which are not sentences of L and to study the structure of the grammatical sequences. The grammar of L will thus be a device that generates all of the grammatical sequences of L and none of the ungrammatical ones. One way to test the adequacy of a grammar proposed for L is to determine whether or not the sequences that it generates are actually grammatical, i.e., acceptable to a native speaker, etc. We can take certain steps towards providing a behavioral criterion for grammaticalness so that this test of adequacy can be carried out. For the purposes of this discussion, however, suppose that we assume intuitive knowledge of the grammatical sentences of English and ask what sort of grammar will be able to do the job of producing these in some effective and illuminating way. We thus face a familiar task of explication of some intuitive concept — in this case, the concept "grammatical in English," and more generally, the concept "grammatical."

Notice that in order to set the aims of grammar significantly it is sufficient to assume a partial knowledge of sentences and non-

sentences. That is, we may assume for this discussion that certain
sequences of phonemes are definitely sentences, and that certain
other sequences are definitely non-sentences. In many intermediate
cases we shall be prepared to let the grammar itself decide, when the
grammar is set up in the simplest way so that it includes the clear
sentences and excludes the clear non-sentences. This is a familiar
feature of explication.[1] A certain number of clear cases, then, will
provide us with a criterion of adequacy for any particular grammar.
For a single language, taken in isolation, this provides only a weak
test of adequacy, since many different grammars may handle the
clear cases properly. This can be generalized to a very strong con-
dition, however, if we insist that the clear cases be handled properly
for *each* language by grammars all of which are constructed by the
same method. That is, each grammar is related to the corpus of
sentences in the language it describes in a way fixed in advance for
all grammars by a given linguistic theory. We then have a very
strong test of adequacy for a linguistic theory that attemps to give a
general explanation for the notion "grammatical sentence" in terms
of "observed sentence," and for the set of grammars constructed in
accordance with such a theory. It is furthermore a reasonable
requirement, since we are interested not only in particular languages,
but also in the general nature of Language. There is a great deal
more that can be said about this crucial topic, but this would take
us too far afield. Cf. § 6.

2.2   On what basis do we actually go about separating grammatical
sequences from ungrammatical sequences? I shall not attempt to

---

[1] Cf., for example, N. Goodman, *The structure of appearance* (Cambridge,
1951), pp. 5-6. Notice that to meet the aims of grammar, given a linguistic
theory, it is sufficient to have a partial knowledge of the sentences (i.e., a
corpus) of the language, since a linguistic theory will state the relation
between the set of observed sentences and the set of grammatical sentences;
i.e., it will define "grammatical sentence" in terms of "observed sentence,"
certain properties of the observed sentences, and certain properties of grammars.
To use Quine's formulation, a linguistic theory will give a general explanation
for what 'could' be in language on the basis of "what *is* plus *simplicity* of the
laws whereby we describe and extrapolate what is". (W. V. Quine, *From a
logical point of view* [Cambridge, 1953], p. 54). Cf. § 6.1.

give a complete answer to this question here (cf. §§ 6.7), but I would like to point out that several answers that immediately suggest themselves could not be correct. First, it is obvious that the set of grammatical sentences cannot be identified with any particular corpus of utterances obtained by the linguist in his field work. Any grammar of a language will *project* the finite and somewhat accidental corpus of observed utterances to a set (presumably infinite) of grammatical utterances. In this respect, a grammar mirrors the behavior of the speaker who, on the basis of a finite and accidental experience with language, can produce or understand an indefinite number of new sentences. Indeed, any explication of the notion "grammatical in L" (i.e., any characterization of "grammatical in L" in terms of "observed utterance of L") can be thought of as offering an explanation for this fundamental aspect of linguistic behavior.

2.3 Second, the notion "grammatical" cannot be identified with "meaningful" or "significant" in any semantic sense. Sentences (1) and (2) are equally nonsensical, but any speaker of English will recognize that only the former is grammatical.

(1) Colorless green ideas sleep furiously.
(2) Furiously sleep ideas green colorless.

Similarly, there is no semantic reason to prefer (3) to (5) or (4) to (6), but only (3) and (4) are grammatical sentences of English.

(3) have you a book on modern music?
(4) the book seems interesting.
(5) read you a book on modern music?
(6) the child seems sleeping.

Such examples suggest that any search for a semantically based definition of "grammaticalness" will be futile. We shall see, in fact, in § 7, that there are deep structural reasons for distinguishing (3) and (4) from (5) and (6); but before we are able to find an explanation for such facts as these we shall have to carry the theory of syntactic structure a good deal beyond its familiar limits.

2.4 Third, the notion "grammatical in English" cannot be identi-

fied in any way with the notion "high order of statistical approximation to English." It is fair to assume that neither sentence (1) nor (2) (nor indeed any part of these sentences) has ever occurred in an English discourse. Hence, in any statistical model for grammaticalness, these sentences will be ruled out on identical grounds as equally 'remote' from English. Yet (1), though nonsensical, is grammatical, while (2) is not. Presented with these sentences, a speaker of English will read (1) with a normal sentence intonation, but he will read (2) with a falling intonation on each word; in fact, with just the intonation pattern given to any sequence of unrelated words. He treats each word in (2) as a separate phrase. Similarly, he will be able to recall (1) much more easily than (2), to learn it much more quickly, etc. Yet he may never have heard or seen any pair of words from these sentences joined in actual discourse. To choose another example, in the context "I saw a fragile—," the words "whale" and "of" may have equal (i.e., zero) frequency in the past linguistic experience of a speaker who will immediately recognize that one of these substitutions, but not the other, gives a grammatical sentence. We cannot, of course, appeal to the fact that sentences such as (1) 'might' be uttered in some sufficiently far-fetched context, while (2) would never be, since the basis for this differentiation between (1) and (2) is precisely what we are interested in determining.

Evidently, one's ability to produce and recognize grammatical utterances is not based on notions of statistical approximation and the like. The custom of calling grammatical sentences those that "can occur", or those that are "possible", has been responsible for some confusion here. It is natural to understand "possible" as meaning "highly probable" and to assume that the linguist's sharp distinction between grammatical and ungrammatical[2] is motivated by a feeling that since the 'reality' of language is too complex to be described completely, he must content himself with a schematized

---

[2]   Below we shall suggest that this sharp distinction may be modified in favor of a notion of levels of grammaticalness. But this has no bearing on the point at issue here. Thus (1) and (2) will be at different levels of grammaticalness even if (1) is assigned a lower degree of grammaticalness than, say, (3) and (4); but they will be at the same level of statistical remoteness from English. The same is true of an indefinite number of similar pairs.

version replacing "zero probability, and all extremely low probabilities, by *impossible*, and all higher probabilities by *possible*."[3] We see, however, that this idea is quite incorrect, and that a structural analysis cannot be understood as a schematic summary developed by sharpening the blurred edges in the full statistical picture. If we rank the sequences of a given length in order of statistical approximation to English, we will find both grammatical and ungrammatical sequences scattered throughout the list; there appears to be no particular relation between order of approximation and grammaticalness. Despite the undeniable interest and importance of semantic and statistical studies of language, they appear to have no direct relevance to the problem of determining or characterizing the set of grammatical utterances. I think that we are forced to conclude that grammar is autonomous and independent of meaning, and that probabilistic models give no particular insight into some of the basic problems of syntactic structure.[4]

---

[3] C. F. Hockett, *A manual of phonology* (Baltimore, 1955), p. 10.

[4] We return to the question of the relation between semantics and syntax in §§ 8, 9, where we argue that this relation can only be studied after the syntactic structure has been determined on independent grounds. I think that much the same thing is true of the relation between syntactic and statistical studies of language. Given the grammar of a language, one can study the use of the language statistically in various ways; and the development of probabilistic models for the use of language (as distinct from the syntactic structure of language) can be quite rewarding. Cf. B. Mandelbrot, "Structure formelle des textes et communication: deux études," *Word* 10.1–27 (1954); H. A. Simon, "On a class of skew distribution functions," *Biometrika* 42.425–40 (1955).

One might seek to develop a more elaborate relation between statistical and syntactic structure than the simple order of approximation model we have rejected. I would certainly not care to argue that any such relation is unthinkable, but I know of no suggestion to this effect that does not have obvious flaws. Notice, in particular, that for any n, we can find a string whose first n words may occur as the beginning of a grammatical sentence S_1 and whose last n words may occur as the ending of some grammatical sentence S_2, but where S_1 must be distinct from S_2. For example, consider the sequences of the form "the man who ... are here," where ... may be a verb phrase of arbitrary length. Notice also that we can have new but perfectly grammatical sequences of word classes, e.g., a sequence of adjectives longer than any ever before produced in the context "I saw a — house." Various attempts to explain the grammatical-ungrammatical distinction, as in the case of (1), (2), on the basis of frequency of sentence type, order of approximation of word class sequences, etc., will run afoul of numerous facts like these.

AN ELEMENTARY LINGUISTIC THEORY

3.1 Assuming the set of grammatical sentences of English to be given, we now ask what sort of device can produce this set (equivalently, what sort of theory gives an adequate account of the structure of this set of utterances). We can think of each sentence of this set as a sequence of phonemes of finite length. A language is an enormously involved system, and it is quite obvious that any attempt to present directly the set of grammatical phoneme sequences would lead to a grammar so complex that it would be practically useless. For this reason (among others), linguistic description proceeds in terms of a system of "levels of representations." Instead of stating the phonemic structure of sentences directly, the linguist sets up such 'higher level' elements as morphemes, and states separately the morphemic structure of sentences and the phonemic structure of morphemes. It can easily be seen that the joint description of these two levels will be much simpler than a direct description of the phonemic structure of sentences.

Let us now consider various ways of describing the morphemic structure of sentences. We ask what sort of grammar is necessary to generate all the sequences of morphemes (or words) that constitute grammatical English sentences, and only these.

One requirement that a grammar must certainly meet is that it be finite. Hence the grammar cannot simply be a list of all morpheme (or word) sequences, since there are infinitely many of these. A familiar communication theoretic model for language suggests a way out of this difficulty. Suppose that we have a machine that can be in any one of a finite number of different internal states, and suppose that this machine switches from one state to another by

producing a certain symbol (let us say, an English word). One of these states is an *initial state*; another is a *final state*. Suppose that the machine begins in the initial state, runs through a sequence of states (producing a word with each transition), and ends in the final state. Then we call the sequence of words that has been produced a "sentence". Each such machine thus defines a certain language; namely, the set of sentences that can be produced in this way. Any language that can be produced by a machine of this sort we call a *finite state language*; and we can call the machine itself a *finite state grammar*. A finite state grammar can be represented graphically in the form of a "state diagram".[1] For example, the grammar that produces just the two sentences "the man comes" and "the men come" can be represented by the following state diagram:

(7)
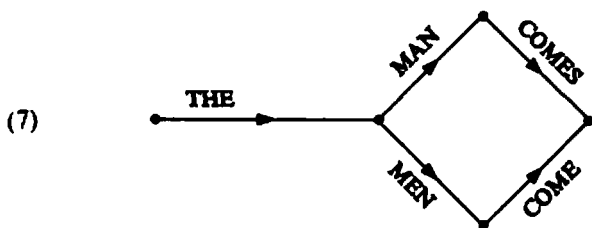
We can extend this grammar to produce an infinite number of sentences by adding closed loops. Thus the finite grammar of the subpart of English containing the above sentences in addition to "the old man comes", "the old old man comes", ..., "the old men come", "the old old men come", ..., can be represented by the following state diagram:

(8)
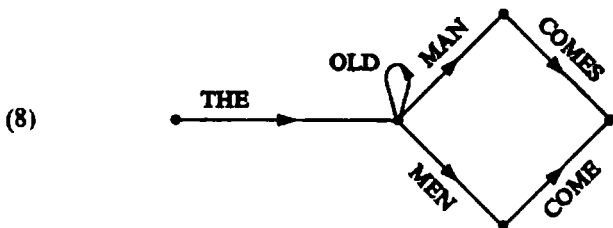
[1] C. E. Shannon and W. Weaver, *The mathematical theory of communication* (Urbana, 1949), pp. 15f.

Given a state diagram, we produce a sentence by tracing a path from the initial point on the left to the final point on the right, always proceeding in the direction of the arrows. Having reached a certain point in the diagram, we can proceed along any path leading from this point, whether or not this path has been traversed before in constructing the sentence in question. Each node in such a diagram thus corresponds to a state of the machine. We can allow transition from one state to another in several ways, and we can have any number of closed loops of any length. The machines that produce languages in this manner are known mathematically as "finite state Markov processes." To complete this elementary communication theoretic model for language, we assign a probability to each transition from state to state. We can then calculate the "uncertainty" associated with each state and we can define the "information content" of the language as the average uncertainty, weighted by the probability of being in the associated states. Since we are studying grammatical, not statistical structure of language here, this generalization does not concern us.

This conception of language is an extremely powerful and general one. If we can adopt it, we can view the speaker as being essentially a machine of the type considered. In producing a sentence, the speaker begins in the initial state, produces the first word of the sentence, thereby switching into a second state which limits the choice of the second word, etc. Each state through which he passes represents the grammatical restrictions that limit the choice of the next word at this point in the utterance.[2]

In view of the generality of this conception of language, and its utility in such related disciplines as communication theory, it is important to inquire into the consequences of adopting this point of view in the syntactic study of some language such as English or a formalized system of mathematics. Any attempt to construct a finite state grammar for English runs into serious difficulties and complications at the very outset, as the reader can easily convince himself. However, it is unnecessary to attempt to show this by

[2] This is essentially the model of language that Hockett develops in *A manual of phonology* (Baltimore, 1955), 02.

example, in view of the following more general remark about English:

(9)   English is not a finite state language.

That is, it is *impossible*, not just difficult, to construct a device of the type described above (a diagram such as (7) or (8)) which will produce all and only the grammatical sentences of English. To demonstrate (9) it is necessary to define the syntactic properties of English more precisely. We shall proceed to describe certain syntactic properties of English which indicate that, under any reasonable delimitation of the set of sentences of the language, (9) can be regarded as a theorem concerning English. To go back to the question asked in the second paragraph of § 3, (9) asserts that it is not possible to state the morphemic structure of sentences directly by means of some such device as a state diagram, and that the Markov process conception of language outlined above cannot be accepted, at least for the purposes of grammar.

3.2 A language is defined by giving its 'alphabet' (i.e., the finite set of symbols out of which its sentences are constructed) and its grammatical sentences. Before investigating English directly, let us consider several languages whose alphabets contain just the letters *a*, *b*, and whose sentences are as defined in (10i–iii):

(10) (i)   *ab, aabb, auabbb*, ..., and in general, all sentences consisting of *n* occurrences of *a* followed by *n* occurrences of *b* and only these;

   (ii)   *aa, bb, abba, baab, aaaa, bbbb, aabbaa, abbbba*, ..., and in general, all sentences consisting of a string *X* followed by the 'mirror image' of *X* (i.e., *X* in reverse), and only these;

   (iii)   *aa, bb, abab, baba, aaaa, bbbb, aabaab, abbabb*, ..., and in general, all sentences consisting of a string *X* of *a*'s and *b*'s followed by the identical string *X*, and only these.

We can easily show that each of these three languages is not a finite state language. Similarly, languages such as (10) where the *a*'s and *b*'s in question are not consecutive, but are embedded in other

strings, will fail to be finite state languages under quite general conditions.[3]

But it is clear that there are subparts of English with the basic form of (10i) and (10ii). Let $S_1, S_2, S_3, \ldots$ be declarative sentences in English. Then we can have such English sentences as:

(11) (i)   If S_1, then S_2.
      (ii) Either S_3, or S_4.
      (iii) The man who said that S_5, is arriving today.

In (11i), we cannot have "or" in place of "then"; in (11ii), we cannot have "then" in place of "or"; in (11iii), we cannot have "are" instead of "is". In each of these cases there is a dependency between words on opposite sides of the comma (i.e., "if"–"then", "either"–"or", "man"–"is"). But between the interdependent words, in each case, we can insert a declarative sentence S_1, S_3, S_5, and this declarative sentence may in fact be one of (11 i-iii). Thus if in (11i) we take S_1 as (11ii) and S_3 as (11iii), we will have the sentence:

(12) if, either (11iii), or S_4, then S_2,

and S_5 in (11iii) may again be one of the sentences of (11). It is clear, then, that in English we can find a sequence $a + S_1 + b$, where there is a dependency between a and b, and we can select as S_1 another sequence containing $c + S_2 + d$, where there is a dependency between c and d, then select as S_2 another sequence of this form, etc. A set of sentences that is constructed in this way (and we see from (11) that there are several possibilities available for such construction— (11) comes nowhere near exhausting these possibilities) will have all of the mirror image properties of (10ii) which exclude (10ii) from the set of finite state languages. Thus we can find various kinds of non-

---

[3] See my "Three models for the description of language," *I.R.E. Transactions on Information Theory*, vol. IT-2, Proceedings of the symposium on information theory, Sept., 1956, for a statement of such conditions and a proof of (9). Notice in particular that the set of well-formed formulas of any formalized system of mathematics or logic will fail to constitute a finite state language, because of paired parentheses or equivalent restrictions.

finite state models within English. This is a rough indication of the lines along which a rigorous proof of (9) can be given, on the assumption that such sentences as (11) and (12) belong to English, while sentences that contradict the cited dependencies of (11) (e.g., "either S_1, then S_2," etc.) do not belong to English. Note that many of the sentences of the form (12), etc., will be quite strange and unusual (they can often be made less strange by replacing "if" by "whenever", "on the assumption that", "if it is the case that", etc., without changing the substance of our remarks). But they are all grammatical sentences, formed by processes of sentence construction so simple and elementary that even the most rudimentary English grammar would contain them. They can be understood, and we can even state quite simply the conditions under which they can be true. It is difficult to conceive of any possible motivation for excluding them from the set of grammatical English sentences. Hence it seems quite clear that no theory of linguistic structure based exclusively on Markov process models and the like, will be able to explain or account for the ability of a speaker of English to produce and understand new utterances, while he rejects other new sequences as not belonging to the language.

3.3 We might arbitrarily decree that such processes of sentence formation in English as those we are discussing cannot be carried out more than n times, for some fixed n. This would of course make English a finite state language, as, for example, would a limitation of English sentences to length of less than a million words. Such arbitrary limitations serve no useful purpose, however. The point is that there are processes of sentence formation that finite state grammars are intrinsically not equipped to handle. If these processes have no finite limit, we can prove the literal inapplicability of this elementary theory. If the processes have a limit, then the construction of a finite state grammar will not be literally out of the question, since it will be possible to list the sentences, and a list is essentially a trivial finite state grammar. But this grammar will be so complex that it will be of little use or interest. In general, the assumption that languages are infinite is made in order to simplify

the description of these languages. If a grammar does not have recursive devices (closed loops, as in (8), in the finite state grammar) it will be prohibitively complex. If it does have recursive devices of some sort, it will produce infinitely many sentences.

In short, the approach to the analysis of grammaticalness suggested here in terms of a finite state Markov process that produces sentences from left to right, appears to lead to a dead end just as surely as the proposals rejected in § 2. If a grammar of this type produces all English sentences, it will produce many non-sentences as well. If it produces only English sentences, we can be sure that there will be an infinite number of true sentences, false sentences, reasonable questions, etc., which it simply will not produce.

The conception of grammar which has just been rejected represents in a way the minimal linguistic theory that merits serious consideration. A finite state grammar is the simplest type of grammar which, with a finite amount of apparatus, can generate an infinite number of sentences. We have seen that such a limited linguistic theory is not adequate; we are forced to search for some more powerful type of grammar and some more 'abstract' form of linguistic theory. The notion of "linguistic level of representation" put forth at the outset of this section must be modified and elaborated. At least one linguistic level *cannot* have this simple structure. That is, on some level, it will not be the case that each sentence is represented simply as a finite sequence of elements of some sort, generated from left to right by some simple device. Alternatively, we must give up the hope of finding a *finite* set of levels, ordered from high to low, so constructed that we can generate all utterances by stating the permitted sequences of highest level elements, the constituency of each highest level element in terms of elements of the second level, etc., finally stating the phonemic constituency of elements of the next-to-lowest level.[4] At the outset of § 3, we

---

[4] A third alternative would be to retain the notion of a linguistic level as a simple linear method of representation, but to generate at least one such level from left to right by a device with more capacity than a finite state Markov process. There are so many difficulties with the notion of linguistic level based on left to right generation, both in terms of complexity of description and lack

proposed that levels be established in this way in order to *simplify* the description of the set of grammatical phoneme sequences. If a language can be described in an elementary, left-to-right manner in terms of a single level (i.e., if it is a finite state language) then this description may indeed be simplified by construction of such higher levels; but to generate non-finite state languages such as English we need fundamentally different methods, and a more general concept of "linguistic level".

---

of explanatory power (cf. § 8), that it seems pointless to pursue this approach any further. The grammars that we discuss below that do not generate from left to right also correspond to processes less elementary than finite state Markov processes. But they are perhaps less powerful than the kind of device that would be required for direct left-to-right generation of English. Cf. my "Three models for the description of language" for some futher discussion.