Chapter 11

Discourse and Sublanguage

Zellig Harris

Discourse and sublanguage are two different objects for science, but they have similar points of departure out of the grammar of sentences, and even some similar structural properties. The major difference between them is that discourses are the directly observable events which constitute the occurrence of language, whereas a sublanguage is a construct – a structure that characterizes certain discourses, or certain parts of discourses, which occur in particular situations, e. g. in the discussions of scientists who work on a particular problem. The major similarity is that discourse and sublanguage can each be described as patterns of recurrence for the individual words in the various word-class positions in sentence structure.

1.

First, as to discourse. Its importance as the domain of some structural relation is overshadowed by the importance of the sentence. The sentence has its most obvious status, informally, as the "minimum free form" of Bloomfield. Whereas almost any word or construction (phrase) of words can be said by itself in answer to a question, or contrastively, or the like, a sentence can be said by itself independently of any particularly structured linguistic environment. However, the more important though less immediately observable property of a sentence is that any effective stochastic process which sets out to describe the word sequence of a discourse will be found to have recurrent points at which the stochastic process begins afresh: these are sentence boundaries.

In a stochastic description of the word-sequence, we say, very roughly, that the first word of a discourse can be (1) *the* or a quantifier, or an adverb, or an adjective, or a noun, or (2) a pronoun, or (3) a preposition, or (4) a verb, etc. If it is *the*, the next word can be any other word of (1). If we reach a quantifier, either as first word or as the follower of *the*, the next word can be any further word of (1) or else a conjunction (possibly plus *the*) plus quantifier. If we reach an adverb in this sequence, the next word is either an adjective or else a conjunction leading to an adverb. If we reach an adjective in this sequence, the next word is either a noun or else a conjunction leading to an adjective. If we reach a noun, the next word is either a preposition or a *wh*-word (zeroable in statable cases), or certain adjectives

(chiefly the participle), or a verb, or a conjunction leading to a noun. If we reach a preposition, the next word may in certain cases be another preposition, and in any case then a word of (1) or (2) above. And if we reach a verb, the next word is a word of (1) or (2), or else a preposition, depending on the particular verb. In such a statement of how the successive word-classes of a discourse depend upon the preceding sequence we come recurrently to points, which we call sentence boundaries, at which there is no regular dependence on the preceding (in terms of this stochastic process), and the possibilities are the same as for the beginning of the discourse.

The sentence is thus a sub-sequence of words which has a certain structure (given by the stochastic process on the word-classes), and independent in respect to it from the neighboring sentences. The independent sayability of a sentence is due to its structural independence.

The stochastic process could be formulated effectively only in terms of the dependence among word-classes, not among individual words. Both the sentence and the word-classes are necessary constructs in the course of formulating the structure of language.

The listing above is of course very sketchy. There may be more than one apparent sentence type (i. e. sentence-making sequence of word-classes), and each sequence may be interruptable at various points by insertable sequences. The whole description is simplified by transformations, which show that all the sentence types are transformed from a basic one, and that the insertions are themselves transforms of a whole sentence which is conjoined to the host sentence. But the recognition of sentences, and the ability to formulate their structure up to some reasonable level of detail, is common to all methods of grammatical description. The discourse remains as the actual datum of language, but it can be described at this stage by no more than an unrestricted or little-restricted succession of sentences.

The sequence of sentence-structures in a discourse does not specify all that there is to say about the structure of the discourse. That further structure, which characterizes a discourse, is not a matter of detailed restrictions on the sentence-structure sequence. (An example of such a restriction would be to reject the conjoining of a question with an assertion; but even this exists to some extent, as in *I'm going, and are you?* and in *Are you going?, because I am.*) Rather, the further structure turns out to be the fact that words recur in particular positions relative to other recurring words, within the word-class sequences which constitute the sentences of the discourse. This recurrence is visible already in the S_1CS_2 structure, i. e. in sentences composed of two or more sentences with a conjunction between every two component sentences. Here it is found that in many cases the two component sentences connected by a conjunction have a word in common. If they do not, the S_1CS_2 seems reasonable chiefly when there is a known semantic connection between some word of S_2 and some word of S_1. But this semantic connection can be stated in an additional S_a which can be

conjoined to S_1CS_2 and which can then be zeroed precisely because it is known. Then the S_1CS_2 without word-recurrence would be derived from a reconstructed $S_1CS_aCS_2$ (or $S_1CS_2CS_a$) where word-recurrence is satisfied in the pair S_1, S_a and in the pair S_a, S_2: e. g. *They put off the camping trip because rain was forecast* from something like *They put off the camping trip because rain, which would wash out the camping, was forecast.*

The constraints on word-recurrence are not simply a more detailed stage of the sentence-constraints on word-class sequences. For one thing, they consist in repetition, not in sequential dependence. For another, they do not specify the occurrence of particular subclasses of words, or of individual words. They consist of not much more than the requirement that there be some recurrence, and preferably in patterned ways which repeat throughout the given discourse. Indeed, were it possible to formulate the requirements for word-recurrence precisely, and so to state a grammar of discourse in terms of the individual words which may be combined in the successive sentences, we would have achieved the instructions for the fabled monkeys to type out Shakespeare. For if we consider the interpretation of these constraints, we find that the word-class constraints that comprise the grammar of sentences carry the dependencies that constitute information-giving per se – predicate, subject and object, modifier, secondary clause – while the way words recur in respect to each other in the positions defined by these constraints carries the specific information of the discourse.

Although the specific word-recurrences in the successive sentences of a discourse are unique to that discourse, various types of recurrence patterns seem to characterize various types of discourses. Sentences which have been strung together without having originated as a discourse (e. g. the first sentences of every hundredth page in an encyclopedia) do not in general show word-recurrence. Colloquial recountings of events have different patterns of word-recurrence than have disciplined reports of scientific observations (e. g. by the naturalists) or of instrument construction. Articles in the so-called "soft sciences" (e. g. sociology) or mixed fields (e. g. environmental studies) have different patterns than those in the "hard sciences". And above all, discourses, or sections of articles, which present an argument have different patterns than discourses or sections which present experimental results.

From all this we can gather that while the relative positioning of word-classes, i. e. their relative dependencies of occurrence in sentences, constitutes giving information, the fact of word-recurrence within these positions constitutes discoursing (i. e. saying something beyond one free-standing statement), and certain patterns of word-recurrence are necessary to the structure of colloquial narration, systematic report, conclusion-drawing, etc. The various types of word-recurrence are worth studying as the inherent carriers of various organizations of informations. And the particular pattern of word-recurrence in a given discourse or section is

useful as a framework of the particular information and information processing in that discourse.

2.

We now consider sublanguages. There is, aside from discourse structure, another constraint on word-occurrence which also goes beyond the grammar of sentences and which is partly related to the constraint in discourses. If we take a body of sentences, whether separate ones or covering whole discourses, which occur within a sufficiently systematic subject matter such as scientific articles in a single area, we find that in addition to the constraints on word-occurrence which are embodied in the word-class structure of sentences and discourses, there are explicit constraints on occurrence for the words in each class. For example, for particular verbs, some nouns may appear as objects but not as subjects. Thus *hydrochloric acid* can occur in an adverbial PN on *wash* but not as direct object of that verb: we can find *We washed the polypeptides in hydrochloric acid* but not **We washed the hydrochloric acid in polypeptides.* In the grammar of English as a whole the latter sentence cannot be excluded as being ungrammatical; it might even be said, with the aid of some metaphoric extension of the meaning of 'washing' in a flow of polypeptides. However, in the corpus of biochemical writings or conversations, that sentence will not appear. This fact alone would be of little moment in any structural description of biochemical language. However, when we survey a large number of such exclusions and inclusions of particular words in particular positions relative to other words, we find that word subclasses can be defined such that members of one subclass but not of another occur in particular positions relative to some yet other subclass. In the grammar of a whole language, nouns are distinguished from prepositions and verbs by such facts as that any noun, but no verb or preposition, can appear as the subject of some verb. Comparably here, a subclass which we may call molecule-nouns (including *polypeptides*) can be distinguished from other nouns and from non-nouns by the fact that they, but not the others, can appear as objects of a subclass of verbs including *wash,* or especially of *wash in hydrochloric acid*. In the grammar of the whole language such subclasses based on co-occurrence are not found, because the exclusion of co-occurrence is not sufficiently fixed. If subclasses of noun, etc. are recognized, it is usually on the basis of some grammatical property, of morphology or of the ability to occur with a major class, e. g. prepositions, rather than on the basis of their co-occurring with one subclass rather than another. In contrast, in the set of sentences for which the subject-matter subclasses hold, we have a number of noun subclasses, verb subclasses, etc., and each sentence structure consists of particular combinations of these: a particular

noun subclass for the subject, a particular verb subclass, and a particular noun subclass, for a family of $N_iV_jN_k$ sentence-structures, where the subscripts indicate particular subclasses. This differs from the grammar of the language as a whole, where all NVN sentences would be cases of a single structure, because there, as noted above, we cannot fully exclude co-occurrences that cut across the word subclasses. It also differs from mere co-occurrence preferences because the latter are variable, and not sufficiently sharp to permit subclasses in respect to co-occurrence.

From these considerations, we see that if we take as our raw data the speech and writing in a disciplined subject-matter, we obtain a distinct grammar for this material. The grammar is obtained by following the same procedures as yield the grammars of whole languages, but it is not identical with the grammar of the whole language. The sublanguage grammar has the same gross structure of word classes combining into sentence structures, but it has above all the novel feature of having families of sentence structures with the same gross form (e. g. NVN) but different subclasses. This conforms to the fact that the sublanguage deals with an organized, if not closed, part of the real world, whereas the whole language imposes only the broadest structuring upon our perception of the world.

If we work out the special grammar of a particular sublanguage in detail, and if we compare it with that of other subject-matters, we can see that the special grammar is not merely a linguistic exercise, but a classification of the relevant terms and relations of the given subject matter and a representation of its main fact-structures. It thus approaches being a grammar of the given science.

Whereas most sublanguages have to be analyzed on the basis of specially selected material, in a particular body of texts, there is one which does not need special sources. This is the grammar of the language, and the grammar of the grammar (which we will call "grammar^{2}"). A grammar of a language is itself a subset of that language, and its sentences can be described by a special grammar. The predicates in this sublanguage are *is a word, is a sentence, is next after, co-occurs with* and the like; the subjects and objects are (mentions of) the phonemes, words, word-sequences, etc. of the language. When the subject-matter is the grammar of the language, then its grammar in turn ("grammar^{2}") is a further sublanguage, in which *being a word, being next after*, etc. are among the terms, and various relational words such as *is similar to* are among the predicates. These grammars are one kind of metalanguage, each one a metalanguage of the one below it; and although there is an infinite regress of such metalanguages, it is found that after three stages their grammars become identical: the structure of "grammar^{4}" and all higher grammars is the same as the structure of "grammar^{3}". This situation is mentioned here not only as a special case of sublanguages, but also to show how constructing grammars of sublanguages can yield results about the sublanguages in question.

It remains to specify the relation of the sublanguages to the whole language. The sentences of each sublanguage are sentences of the whole language, since the words (even the new technical terms) are in the whole language, and the gross grammar of the whole language is satisfied by the sentences of the sublanguage. If $N_iV_jN_k$ (for particular subclasses) is a sentence-structure of the sublanguage, then any sentence having that form has meaning in the sublanguage and can be included in it. If we reinterpret the gross sentence-structures, such as NVN, to be not only the combining of any N and V words but also an envelope of all combinings of N subclasses with V subclasses, we can say that the $N_iV_jN_k$ structures of a particular sublanguage are included in the NVN structures of the whole language. In that case, the sublanguage, as a set of sentences formed by the $N_iV_jN_k$ and other structures is closed with respect to certain structures of the language – namely these subscripted structures. Hence this set of sentences can indeed be called a sublanguage of the whole language.

In the case of a whole language, a grammar is initially constructed to characterize the discourses of the language by their regularities; but by organizing the regularities it becomes able to indicate if some parts of the discourses are not in the language – for example, a French sentence in an English text. Similarly, the grammar of a sublanguage, although drawn from texts in a given subject-matter, can enable us to say that certain sentences in the text are in the ordinary language rather than in the sublanguage. The subgrammar thus becomes not simply a description of certain texts, but a specification of the relations among the relevant terms of a disciplined subject-matter.