# ON GRAMMARS OF SCIENCE

Zellig Harris
University of Pennsylvania

The languages of sciences can be studied as sublanguages of natural language, in a way that shows what can be done with the distributional methods of linguistics, and that also throws light on the structure of science.

Whatever else may be said about language, about its meaning and use, there is a directly observable structural property in that, first, every occurrence of language is a linear combination of phonemes, and of words, with intonational or other suprasegmental features, and, second, that certain combinations of these are found in occurrences of language and certain others are not. Grammatical properties which do not appear to be immediately combinatorial--above all, transformations--are obtained as secondary properties of combinatorial structure, specifically as equivalence relations on them. (The term "combinatorial" is used without its specific meaning in mathematics, in place of the somewhat aberrant linguistic use of the term "distributional".)

A language can be characterized by (a) which phoneme combinations, and especially what word combinations, are in it as against (b) those which are not. Between (a) and (b) lie combinations which are marginal, i.e. about which no decision or agreement can be reached. It is not practicable to make this characterization by listing all the (a), i.e. grammatical combinations; therefore we have recourse to statements of regularities as to which kinds of combination are admitted and which are excluded. It is essential that these statements be as unredundant as possible, e.g. that they not state certain exclusions twice, as cases of different regularities. The reason for this stems from the fact

that language has no outside metalanguage in which its structure can be described. Any statements which characterize the words and sentence-structures of a language have to be given in the same language (or in some other one) using already the same kinds of words and sentence-structures which have to be defined. Hence the ultimate entities and operations cannot be defined externally but must be distinguished by their combinations in respect to each other, by the constraints which characterize their departures from randomness. These constraints on combination observably exist in language, and are certainly related to the information which language expresses. The redundancies inherent in these constraints thus characterize (or "predict") the ultimate entities and operations. Hence any further redundancies which are due to the way the grammar states the constraints will muddy the characterization: the grammar must predict the existing combinations on the basis of the fewest constraints possible.

When this is done, it is found that the sentences of a language can be predicted by a partial ordering (rather than some linear combination) on words, the partial order being determined by the standing of the words in a single hierarchy of dependence. In addition, a discourse is characterized by there being some recurrence of families of these partial orderings. These structural (combinatorial) properties of language are accompanied by informational properties. The partially-ordered dependence is called an "operator-argument" relation and has the semantic effect that the operator states a property (event, act, etc.) of its arguments. The recurrence of sentences of a particular family has the semantic effect of discussing (as against merely announcing) the event or situation common to the sentences of that family.

When the combinatorial investigations are made in discourses that arise around a particular, relatively narrow, subject matter, the grammatical description which is restricted to just these discourses differs in important respects from the grammar of the whole language. In particular, if the subject matter is in a well-organized science, the special grammar describes a sublanguage which is closed under the operations of the language; and the grammar is seen to reflect the objects and relations of the science. We

find several classes of elementary arguments (roughly, nouns) each of which occurs only under a particular class of elementary operators (roughly, verbs or adjectives). This is a situation which does not occur in a whole language, for there every elementary operator (i.e. one whose arguments are elementary, e.g. *fall*) can in principle occur on any elementary argument. The likelihoods for these operators (in the whole language) are different: under *falls* we can readily find *stone* and also *word*, and *night*, quite unlikely *day* (except perhaps to parallel *night*), and hardly at all *vacuum*. But one cannot exclude even *Vacuum fell* from the language, and one cannot establish subclasses of nouns and of verbs such that only nouns of a given subclass can occur, in sentences of the language, under a verb of a stated subclass. In contrast, the exclusion of particular subclasses from occurring with certain words is common in each particular science. In immunological articles we can find *Lymphocytes secrete antibody*, *Lymphocytes produce antibody*, *Plasma cells produce antibody*, *Plasma cells produce agglutinin*, *Plasma cells contain antibody*, with *C (lymphocytes, plasma cells)* being first argument and *A (antibody, agglutinin)* second argument of *V (contain, produce, secrete)*, but never the opposite order of arguments. In English as a whole, if someone said the opposite order, e.g. *Antibody secretes lymphocytes* we would say he is innocent of biochemistry but we could not say he is innocent of English grammar, or he is not speaking English. In immunological articles we find *The tissue was inflamed*, while *The antibody was inflamed* is excluded; but again the latter cannot be excluded from English grammar.

Thus each science language has elementary operator classes which are restricted to occurring only on particular elementary argument classes, whereas the whole language does not have such definite restrictions, only great differences of likelihood.

In the material in a science, certain words which are placed in the same class because they occur under the same operator have to be assigned to different subclasses because there are other operators under which one of the words occurs while another is excluded. For example, some immunological articles have (in their analyzed form) *Lymphocytes contain antibody but they do not produce antibody.* This would be represented by *CVA but CV˜A* (using ˜ to indicate negation on the preceding symbol). Since this looks like a

that language has no outside metalanguage in which its structure can be described. Any statements which characterize the words and sentence-structures of a language have to be given in the same language (or in some other one) using already the same kinds of words and sentence-structures which have to be defined. Hence the ultimate entities and operations cannot be defined externally but must be distinguished by their combinations in respect to each other, by the constraints which characterize their departures from randomness. These constraints on combination observably exist in language, and are certainly related to the information which language expresses. The redundancies inherent in these constraints thus characterize (or "predict") the ultimate entities and operations. Hence any further redundancies which are due to the way the grammar states the constraints will muddy the characterization: the grammar must predict the existing combinations on the basis of the fewest constraints possible.

When this is done, it is found that the sentences of a language can be predicted by a partial ordering (rather than some linear combination) on words, the partial order being determined by the standing of the words in a single hierarchy of dependence. In addition, a discourse is characterized by there being some recurrence of families of these partial orderings. These structural (combinatorial) properties of language are accompanied by informational properties. The partially-ordered dependence is called an "operator-argument" relation and has the semantic effect that the operator states a property (event, act, etc.) of its arguments. The recurrence of sentences of a particular family has the semantic effect of discussing (as against merely announcing) the event or situation common to the sentences of that family.

When the combinatorial investigations are made in discourses that arise around a particular, relatively narrow, subject matter, the grammatical description which is restricted to just these discourses differs in important respects from the grammar of the whole language. In particular, if the subject matter is in a well-organized science, the special grammar describes a sublanguage which is closed under the operations of the language; and the grammar is seen to reflect the objects and relations of the science. We

find several classes of elementary arguments (roughly, nouns) each of which occurs only under a particular class of elementary operators (roughly, verbs or adjectives). This is a situation which does not occur in a whole language, for there every elementary operator (i.e. one whose arguments are elementary, e.g. *fall)* can in principle occur on any elementary argument. The likelihoods for these operators (in the whole language) are different: under *falls* we can readily find *stone* and also *word*, and *night*, quite unlikely *day* (except perhaps to parallel *night)*, and hardly at all *vacuum.* But one cannot exclude even *Vacuum fell* from the language, and one cannot establish subclasses of nouns and of verbs such that only nouns of a given subclass can occur, in sentences of the language, under a verb of a stated subclass. In contrast, the exclusion of particular subclasses from occurring with certain words is common in each particular science. In immunological articles we can find *Lymphocytes secrete antibody, Lymphocytes produce antibody, Plasma cells produce antibody, Plasma cells produce agglutinin, Plasma cells contain antibody,* with *C (lymphocytes, plasma cells)* being first argument and *A (antibody, agglutinin)* second argument of *V (contain, produce, secrete),* but never the opposite order of arguments. In English as a whole, if someone said the opposite order, e.g. *Antibody secretes lymphocytes* we would say he is innocent of biochemistry but we could not say he is innocent of English grammar, or he is not speaking English. In immunological articles we find *The tissue was inflamed,* while *The antibody was inflamed* is excluded; but again the latter cannot be excluded from English grammar.

Thus each science language has elementary operator classes which are restricted to occurring only on particular elementary argument classes, whereas the whole language does not have such definite restrictions, only great differences of likelihood.

In the material in a science, certain words which are placed in the same class because they occur under the same operator have to be assigned to different subclasses because there are other operators under which one of the words occurs while another is excluded. For example, some immunological articles have (in their analyzed form) *Lymphocytes contain antibody but they do not produce antibody.* This would be represented by *CVA but CV˜A* (using ˜ to indicate negation on the preceding symbol). Since this looks like a

contradiction, and so is not a satisfactory representation, we set up two subclasses, $V_i$ *(contain,* and in the reverse direction *is found in)* and $V_p$ *(produce, form, synthesize)* yielding $CV_iA$ *but* $CV_P^{\sim}A$.

In the elementary arguments, many subclasses (which are highly technical terms) are found to contain only one meaning. Some of the words in the subclass have different meanings in the science, but the difference is irrelevant to the particular article or to the particular research problem whose grammar is being investigated. An example is seen in *antibody* and *agglutinin* above. A more important case of synonymy relative to the immediate subject-matter is seen in many operator-classes, such as $V_p$ above, where it seems as if many different words are members of a single subclass. In sciences such as mathematics, physics, chemistry, and much of biology, it is found that the different operators which appear in the same class are synonymous in respect to their arguments in the science. That is to say, their meaning differences in the whole language are not used in the science material. For example, the differences between *synthesize* and *produce*, which are reflected in their different arguments and farther environments in English, do not apply in the immunology articles, where both verbs have the same arguments. The importance of this synonymity lies in the fact that the open-endedness of the English vocabulary in science is only apparent and not real: an author can draw for the $V^p$ position upon any word that even remotely means "to make"; but in so doing he is not using the particular meaning of the word, but merely using different phoneme sequences for the one entity $V_p$. Hence the science is operating not with an open vocabulary, but with a small explicit set of word-classes and subclasses, many of which have only one member, i.e. do not have different members with different meanings and details of environment.

The effect of this is that we have for each small subscience a well-defined set of word-classes and subclasses, which constitute the vocabulary sufficient for stating the facts of that subscience. These classes have been established in respect to their operator-argument combinations (and secondarily in respect to their farther environment); therefore their occurrences necessarily constitute sentences. These are the elementary sentences of the science, which in

addition to the operators and their arguments may have modifiers or local additions (all of which are ultimately derived from secondary operators) attached to one or another of the main words of the sentence. Each elementary sentence in the subscience can therefore be written as a formula in the fixed word classes and subclasses and modifiers of the science. For the moment we consider the formulas only as a normal form for the sentences, enabling us to know where each item of information is to be found, and enabling us to compare sentences in a regular way, and the like. Ultimately one can consider the formulas to be the sentences of the science.

The conjunctions which operate on pairs of elementary sentences (or formulas) have not as yet been found to fall into a fixed vocabulary of subclasses with fixed constraints on their combinability with the sentences or with each other. So far, we can only say that the conjunctions are used as in the language as a whole. If a special structure is found in the use of conjunctions in science, it would presumably be not unique to each science (and so part of the particular grammar of that science), but rather common to a certain set of sciences and so belonging to the grammar of scientific discussion in general. There is, however, one field where the conjunctions are organized into specified subclasses with specified constraints, so that they are as much part of the grammar of the science as are the sentences on which the conjunctions operate. This case is mathematics, where therefore not only the possible sentence structures but also the possible sequences of them (determined by the conjunctions on them) are well-defined, i.e. have well-formedness conditions. This last is seen in proof structure.

In contrast, there is many a subject-matter looser than the above-mentioned sciences which has a largely unstructured open vocabulary of operators, drawn from the whole language and used in the subject-matter in the same meanings they have in the whole language. Some of these fields, e.g. history, may lack a more-or-less closed vocabulary even in their nouns. Such openness of vocabulary may hold even for fields such as law which have a great amount of technical terminology, if it is found that the full breadth of natural language can be combined with the special terminology. In all these open-vocabulary cases, it is not

possible to use a priori combinatorial grounds so as to reduce all sentences of the field to formulas of fixed structure.

To all this we have to add that any sentence of the sublanguage, or conjunctional combination of sentences, can be an argument of a metalinguistic operator: e.g. *It has been shown that lymphocytes contain antibody.* Other metalinguistic operators have nouns (primitive arguments) of the sublanguage as their arguments: *They studied lymphocytes.* The metalinguistic operators and certain material attached to them are constructed according to the vocabulary and grammar of the language as a whole and do not accord with the sublanguage grammar, even though they are part of the language of science.

As an example of the language of a particular subscience, consider here the grammar of a set of articles, published in the course of some 35 years, on the question of which cell was the producer of antibody. On combinatorial grounds, it was found that the articles contained the following word-classes: *G (antigens), J (is injected), B* (animal or body-part), *F (infection), U (move), T (tissue), C (cell), S* (structures within cell), *W (respond,* with *T, C,* or *S* as subject), *A* (antibody), *V* (verbs with *A* subject and *C* or *T* indirect object), *Y* (verbs with *C* subject and object, e.g. *is called, develops into).* These classes combined into just a few sentence-structures, namely *GJB* (as in *Antigen was injected subcutaneously), GUT (The antigen travelled to the lymph node)* and *GUC (The antigen was taken up by lymphocytes), TW (The lymph nodes were inflamed), CWT (Lymphoid cells leave the lymph nodes), CW (The lymphocytes disintegrate), SCW (The cytoplasm in the lymphocytes was broadened), AVC (Antibody is present in plasma cells), CYC (The plasma cells were derived from blast cells).* Disregarding synonyms (e.g. *nodes* and *glands,* or *is found in* and *is contained in),* and words whose meaning differences in the science are not relevant to the combinatorial possibilities (or results) in those articles (e.g. as among the various antigens here discussed), only a few of these classes contained subclasses whose combinations differed, in a detail, from each other. There is $U^{i}$ *(is arrested in),* $U^{d}$ *(perishes in),* in addition to *U.* There are almost 20 subclasses of *T:* $T^{b}$ *(blood),* $T^{n}$, *(lymph nodes),* etc.; and 7 of *C:* $C^{y}$ *(lymphocytes),* $C^{z}$ *(plasma cells),* etc.; and some 10 of

*S:* $S^c$ *(cytoplasm),* $S^n$ *(nucleus),* etc. Of *W* there are some 15 subclasses, such as $W^a$ *(react),* $W^i$ *(is in),* $W^p$ *(multiply),* $W^c$ *(change),* $W^m$ *(mature),* $W^u$ *(move),* $W^d$ *(disintegrate),* and several specific to particular *S*, such as $W^e$ *(eccentric)* with $S^n$ subject. Of *A* the main subclasses are $A^a$ *(substance)* and $A^p$ *(protein);* otherwise it is *antibody* or the equivalent. In *V* there is $V^i$ *(appears in),* $V^p$ *(is produced in),* $V^u$ *(passes through),* $V^s$ *(is secreted from),* $V^t$ *(is stored in). Y is, is same as,* but $Y^c$ is *develops (from, into).*

There are no relevant combinatorial differences among the words within a subclass (written as symbol with subscript), or a class without subscript. Hence each subclass or subscriptless class symbol is just one word in the vocabulary of this subscience. To these words, in the above sentence-structures, there are occasionally added modifiers (derived from relative clauses, hence from subordinate sentences) and local operators: e.g. aspectual local operators such as *begin to,* and negative ones such as *fail to;* negation and quantity modifiers such as *not, few, increased;* prepositional operators such as *from.* There is a local operator which is important in this subscience *have a role in, be capable of;* and time modifiers *(until the 7th day);* and a few modifiers relevant for this field such as *in vitro, mature, disintegrating, family.*

The sentences of the articles can be transformed, using a priori precisely stated transformations, into a sequence of formulas each composed of some of the stated subclasses or classes, possibly carrying stated modifiers, organized into one of the sentence types listed above. These formulas are in many cases joined together by means of conjunctions. One class of (largely synonymous) conjunctions is so restricted in its environment that sentence pairs connected by it may be considered a single "macro-sentence" of the science. This is *is followed by, thereafter,* etc., symbolized here by colon which occurs between *GJB* or *GUT* (or *GUC)* and a following *W* or *V* or $Y^c$ sentence. The *J* and *U* sentences can be considered to report the stimulus, and the *W, V,* or $Y^c$ sentences the response. The macro-sentence formula is thus

*GJB:* ... *U* ... : ... $W/V/Y^c$ ...

(where ...*X*... indicates a sentence whose operator--e.g. verb--is *X),* with ...*U*... being most often unmentioned and *GJB:* often zeroed.

Finally, these sentences, and less frequently individual words, of the science-language, often appear as arguments of operators which may loosely be called metalinguistic. There are several classes of these: e.g. *Investigators* (or a name) *have suggested that* followed by a science sentence; or *We excised* followed by a science word (e.g. *the nodes*).

The validity of the analysis is shown not only by its being acceptable as a reasonable organization of the content of the articles, but also by the fact that in those places where it is known that the articles disagreed, or that new information or conclusions were reached, the above structure of the formulas shows appropriate differences. This in itself is evidence for the adequacy of the distributional method in linguistics in achieving a structure that conforms to the meanings of what is being talked about. One could of course argue that the reality with which a science (or any other language use) is dealing is in general, or grossly, a continuous and continuously varying object which cannot be fully described by any system of discrete phonemes and words. But that is a question for the science itself, rather than for its language. Each science that succeeds in describing or predicting something about the world does it with a use of language and mathematics, which consist essentially of discrete symbols. The methods mentioned above suffice to organize the words and sentences of scientific reports into a vocabulary and a formulaic sentence structure which are at each point in correspondence with the information of the science and with changes in that information.

The grammar sketched above has three structurally separated components. One is the sentence (or macrosentence) types of the particular subscience, which may contain sentences of a priori science as arguments or subordinate clauses. A second is the conjunctions, or hierarchies of conjunctions, on sequences of science-language sentences (excluding, in the case above, the colon which is part of the science-language macrosentence), which may differ appreciably as among different types of sciences. Last is the "metalinguistic" material operating on sentences or words of the science-language; this material may differ only in secondary respects (e.g. in its verbs) as among different types of science.

The conjunctions are particularly important because they may help in characterizing constructions which carry out for science some of the functions of proofs in mathematics. In the discussion section of articles it is seen that statements of conclusions are in general preceded by a sequence of sentences of the science, which provide the grounds for the conclusion. The preceding sentences and the conclusion are of the same sentence types, and the sequence apparently has to satisfy certain special conditions as to which words may occur in a given word-class position in successive sentences (e.g. as to having matched modifiers, or being classifier-words). Given such a sequence, certain hierarchies of conjunctions may suffice (together with certain metalinguistic operators) to assure that the last sentence of the sequence follows from the preceding ones.

The whole subject presented here can be looked at somewhat differently. In the case of the immunology articles discussed above, it was found that when French articles in the field were analyzed, the same word-classes and sentence types appeared as in the English articles. The language of each set of articles can be considered a sublanguage of its particular natural language. But the language common to them all, consisting of the word subclass symbols (which suffice as a vocabulary) and their sentence types, is not a sublanguage of either English or French. Instead, it can be looked upon as an independent linguistic system sufficient for articles in a particular research area. As such, it has certain statable structural relations to the grammars of its prior sciences and of its immediately neighboring research areas (e.g., in this case, the transfer of donor cells from sensitized animals to others, in order to investigate immune response in the recipients); and it has various similarities to the grammars of other sciences.

If we consider the grammars of various sciences in comparison to those of natural languages and of mathematics, we see certain common properties to all, and certain major and understandable differences among the three types of structures. The greatest differences between the science languages and natural language are that the metalanguage (in the technical sense) of a science language is outside the science language, whereas the metalanguage of a natural language is necessarily in it. Because of this, all

the word-classes of a natural language can be defined only in respect to a common co-occurrence requirement (argument-operator), while each word-class of a science-language can be defined (in English, etc., as metalanguage) as co-occurring with arbitrary other classes. Within this over-all difference, the differences between grammars of sciences, or of different periods and problems within a single science, fit the differences in content--that is, the differences in the constraints of reality--in the various subject-matters. More generally, the differences between science languages, natural language, and mathematics fit the different constraints of reality that are involved in the subject-matters of each of these. Thus the structure of grammars is seen to be related to the constraints of dealing with the real world, in ways reminiscent of the views of the Vienna positivists and even of some of the American pragmatists.