ANNALS, *AAPSS*, **495**, January 1988

Science Sublanguages and the Prospects for a Global Language of Science

By ZELLIG HARRIS and PAUL MATTICK, Jr.

ABSTRACT: Scientists have limited access to results published in languages in which they are not fluent. One solution to the problem is suggested by some results of investigation into the nature of language generally and the language of various sciences in particular. The information provided in language is given not only by the meanings of individual words but also by the relations among words, especially by the regularities of their co-occurrence. Particular sciences, furthermore, are characterized by particular sets of such relations among words. These relational structures are shared by discourses within the same scientific field in different languages; these structures can thus be seen as expressing the information carried by language in the field irrespective of national language. Because the informational structures are discoverable in a computable way, the solution suggested here to the problem of international communication in science would at the same time provide facilities for the computer processing and retrieval of scientific information on a large—potentially a global—scale.

Zellig Harris is Benjamin Franklin and University Professor of Linguistics at the University of Pennsylvania. He is the author of numerous books, including Methods in Structural Linguistics, Mathematical Structures of Language, *and* A Grammar of English on Mathematical Principles.

Paul Mattick, Jr., teaches philosophy at Bennington College and is a fellow of the Institute for the Humanities at New York University. He is the author of Social Knowledge.

THE development of science in the modern period coincided with the growth in importance of national linguistic boundaries in cultural life. As signaled by Descartes's presentation of his *Meditations* and certain of his scientific works in French as well as in Latin, this reflected, inter alia, a new secular order of ideas and institutions in opposition to the traditional order represented by Latin as the language at once of classical philosophy and Christian theology. The new cultural order took form in the context of the formation of modern nation-states in Europe, in which societies and academies of the sciences and arts displaced the earlier international system of universities, which were themselves soon to be transformed.

The general issue of European political-cultural disunity was very much the inspiration for the first great project of an international language of science, Leibniz's plan for a *Characteristica Universalis*, a symbolic representation of conceptual elements calculational operations on which would resolve all disputed questions. Leibniz himself wrote in French and German as well as Latin, which remained a basic language of science until the nineteenth century, when the ever more rapid pace of scientific development within national university systems, often in close connection with industrial development, led to its abandonment. On the other hand, certain areas of research became identified with particular languages, so that, for example, students of organic chemistry were obliged to learn German in order to read important research results.

The idea that international understanding would be fostered by a universal language lay behind a number of attempts at inventing such a language, of which Esperanto has been the most significant. Interlingua was invented in 1951 for use at scientific and medical meetings, but it has had little impact, partly as a result of being based on English and Romance languages only. At the present time, English is the closest to an international language of science, due largely to the economic and political dominance of the United States. But the bulk of scientific work is published in many national languages. This limits the access of scientists to results published in languages in which they are not fluent. At the same time science remains by nature an international and transcultural enterprise. The continuing explosion of scientific research around the world makes the question of a global language of science an important one. Considering this explosion of scientific research and the facility that advanced communications technology imparts to scientific interchange, the possibility of a global language of science becomes a reasonable one to examine.

LANGUAGE AND INFORMATION

In each area of science, and more generally in many specific subject matters, the use of language is limited in particular ways—and limited in the same ways no matter what language is being used. This is why it is easier to translate scientific texts than literary ones. These limitations of use, and the interlanguage similarity of the limitations, are due to an essential property of language.

This property is that the information provided in language is given not only by the dictionary meanings of individual words but also by the relations among words, especially by the regularities of their co-occurrence, or combination in sentences. When the grammar of a language is described in its most compact

and essential form, it is found that every contribution to the structure of a sentence—which words combine in what grammatical relation—makes a fixed contribution to the meaning of the sentence. This is an underlying form-content relation not altered by grammatical transformations, which change the form of a sentence but not its information—for example, the reduction of "I prefer for me to leave last" to "I prefer to leave last" does not change the information imparted.

LANGUAGE AND STRUCTURE

An important specialization of the form-content property is the sublanguage structure. It has been found that the use of a language in the texts or talk of a reasonably well-structured subject matter, especially a science, is limited in ways that go far beyond the limitations of ordinary grammar. In ordinary language, sentences consist of verbs with nouns—or whole sentences—as subject and in many cases also as object, with very few hard and fast restrictions as to which nouns can be subjects or objects of which verbs. Thus "child" may be a much more frequent subject of "sleep"—as in "The child slept"—than is "chair" or "universe"; but the latter cannot be excluded from the grammar—as in "That chair slept for years in the attic" and "Until the Big Bang, the universe slept."

In scientific writing, in contrast, we find sharp restrictions on word co-occurrence. In biochemistry, for example, one can say, "The polypeptides were washed in hydrochloric acid," but "Hydrochloric acid was washed in polypeptides," while a grammatical English sentence, cannot appear in a biochemistry article. For each science we find particular sets of nouns that can occur as subject, or object, of particular sets of verbs, to make not just a general noun-verb-noun sentence type as in English or French but a family of distinct sentence types, each with its particular subsets of verbs and of nouns.

LANGUAGE AND SUBLANGUAGE

In any system of a mathematical type, if there is a subset of the system that is closed under operations of the system then that subset is called a subsystem of the whole. "Closed" here means that an operation on any member, or pair of members, of the subset yields another member of the subset.

The subset of English sentences found in texts of a science has this character; grammatical operations on a sentence of the science will produce another sentence that can occur in texts of that science. For example, the active form of "The polypeptides were washed in hydrochloric acid" is "We washed the polypeptides in hydrochloric acid," which is also a sentence of biochemistry. Similarly, the active of "Hydrochloric acid was washed in polypeptides," which is not a biochemistry sentence, would be "We washed hydrochloric acid in polypeptides," which would also not be found in a biochemistry article. The set of English sentences in biochemistry, or in some subfield thereof, constitutes a sublanguage of English.

SUBLANGUAGE FORMULAS

A further linguistic property makes those previously mentioned relevant to the problem of international scientific communication. In every language in which there are texts and conversations in biochemistry, there is a biochemistry sublanguage, and so for every such field.

If we examine the structure of, for instance, the biochemistry sublanguage of French and the biochemistry sublanguage of English—that is, the subsets of nouns, verbs, and other elements and the various sentence types made of them—we find that they are in all essentials identical. If we mark the various word subsets in the English biochemistry sublanguage by letters—for example, by using *P* for "polypeptides" and other molecules that might be treated by washing, *W* for certain laboratory operations, and *S* for certain solutions—we could represent the sentence types by sequences of these word-class symbols. Such a sequence would be "PWS" here. We can show that the same symbol classes and sentential symbol-sequences suffice to characterize the word classes and sentence types of the French biochemistry sublanguage. This means that articles in whatever language in the given biochemical field could be represented by sequences of the same types of formulas. Starting with a science sublanguage, expressed in the words of one language or another, we have reached a science language expressed in symbols.

WHAT THE FORMULAS REPRESENT

The importance of the formulas is not that they are reminiscent of mathematics or chemistry. Indeed, a universe of interformula relations defined a priori, which is at the heart of mathematical equations and chemical-reaction formulas—and such as Leibniz dreamed of for his *Characteristica Universalis*—does not exist for science languages. The science-language formulas are more like the formulation of numbers in a particular notation such as the customary decimal expansion, or like the formulas for each individual chemical compound.

The science-language formulas have two major properties, however. One is that, like any fixed representation, they locate each item under discussion in preset positions relative to other items. If we want to know about any particular object or interobject relation studied in a field, we know where to look for it in the formulaic representation of any document or sentence. The other property of the formulas is that they allow us to free the representation of information from the noninformational features of language. Many languages have grammatical requirements that can go beyond what is needed to express information. For example, English requires that each verb carry a tense—say, present, past, or future—even in cases where tense is irrelevant to the information carried, as in the case of general statements or universal laws such as "Two plus two equals four." The formulas dispense with everything except what is relevant to the information that is distinguishable in the given field. It is therefore not surprising that the same formulas represent the same information irrespective of the language used.[1]

A NATURAL SCIENCE EXAMPLE: IMMUNOLOGY

To show what a science sublanguage is like, we present a very brief sketch of the language of immunology research papers circa 1935-66. This was a period when this field was far smaller and more

1. For a detailed examination of a science sublanguage—that of immunology—and a study of the essential identity of sublanguage symbols and symbol sequences in English and French, see Zellig Harris et al., *The Form of Information in Science: Analysis of an Information Sublanguage*, Boston Studies in the Philosophy of Science 104 (Dordrecht: Reidel, forthcoming).

inspectable than it is now and when it had a central research problem of determining which cell was the producer of antibodies. There was a controversy as to whether it was the lymphocyte or the plasma cell, both of the lymphatic system. After it was shown, by electron microscopy and other methods, that both cell types produced antibodies, the controversy was resolved by the understanding that the two cell names pertained to different stages of development of the same cell line. The purpose of the analysis that will be summarized here was to see if one could give a formal representation, in an orderly and usable way, of all the information contained in articles written in this area, if one could locate in the sentence structures—and characterize structurally—the changes in information over the years, and if one could locate and characterize the disagreements between the scientific workers involved.

Analysis of the literature

The study began with a detailed grammatical analysis of each sentence in each of 14 research papers studied, utilizing linguistic methods to recast sentences, where necessary, into forms facilitating the search for patterns of word repetition. For example, passive constructions were transformed into active ones. Words, or groups of words, were considered discourse equivalent when they appeared in the identical linguistic environment; thus nouns found in the context "_____ was injected" were classed together as antigen words. Because the word classes are defined by their occurrence in particular syntactic relations to other words, which thereby fall into other word classes, the procedure yielding these classes simultaneously yields the sequences of them that constitute the sublanguage sentence types.

The immunology sublanguage

In its barest outline, the sublanguage discovered by this process of analysis contained some 15 word classes. The chief ones, each followed here by the capital letter used to represent it in sublanguage formulas, are those for "antigen" (G), "antibody" (A), "inject" (J), "tissue" (T), "cell" (C), "body part" (B); then for verbs occurring between A and C (V; for example, "appear in," "produced by," "secreted by"), verbs occurring between two cell names (Y; for example, "is similar to," "develops into"), and verbs appearing with T or C words (W; for example, "T inflames," "C proliferates"). Words of these classes appeared, combined, in fewer than 10 major sentence types, chiefly those exemplified by "Antigen is injected into body"; "Antigen moves to tissue"; "Cells or tissues change or have some property"; "Antibody appears in cell"; "Cell is the same as or develops into another cell."

Formulaic representation of sentences

The many sets of synonymous words, especially verbs, are considered to be just variant forms of a single word, and the variants are not indicated. Writing each class with its letter symbol, we can represent the information in each sentence by a formula constructed of letters; thus "Antibody appears in lymphocytes" is AVC. The nonsynonymous words within a class are marked by subscripts, as in V_i for "appears in" and, synonymously, "present in," "contained in"; V_p

for "produced by"; and V_s for "secreted by." There are modifiers on certain verbs—such as "not," "increase," the pair "from" and "to," "begin to," and "have a role in"—and on certain nouns—such as "much," "immature," and "family of [cells]." These are marked by superscripts on the word-class letter.

We thus have these major sentence types, illustrated here by generic sentences rather than actual examples from particular papers:

—GJB, for "Antigen is injected into a body part or an animal."
—$GU^{ft}TT$, for "Antigen moves from tissue to some tissue"; the *ft* superscript indicates "from" and "to."
—TW and CW, for "A tissue [or cell] has some property or undergoes a change."
—AVC, for "Antibody appears in, is produced by, or is secreted from a cell."
—CYC, for "Some cell is similar to or is called some cell."
—CY_cC, for "A cell develops into another cell."

In donor research, in which antigen is injected into one animal, and then lymphocytes are injected, or transferred, from that animal to another, with antibodies then being sought in the second animal, an additional sentence type is found: $CI^{ft}BB$, for "Cells are injected from an animal into another animal."

There is a special conjunction, internal to a particular sentence-type sequence, that appears or is implicit in almost all occurrences of the pair GJB and AVC. This is "thereafter" and its synonyms, marked in our formulaic representation by a colon (:). It often carries a time modifier. An example is $GJB:^{t}AVC$, for "Antigen is injected into a body part; some time later antibody appears in cells." In inverse order the sentence would read, "Antibody appeared some time after ingestion of antigen." This conjunction takes different grammatical forms, for example, "to" in "The cell contained antibody to the antigen." All these forms synonymously connect AVC—or CW or TW—to GJB.

To recapitulate the analytic procedure

This sketch of the immunology sublanguage is sufficient to indicate the advances in the analysis of science information obtainable from the codification of the sublanguage structure. To begin with, metascience material, which states scientists' relations to the information of the science, can be distinguished from the latter, which appears in the form of nominalized sentences embedded in a recognizable set of contexts, such as "Researchers have shown that _____," "_____ as was expected," or "It was found that _____," as in "It was found that antibody is in lymphocytes," or the equivalent, "Antibody was found in lymphocytes."

We obtain a gross framework for representing the information in the field: the word-class sequence formulas, such as AVC. We also obtain a representation for the specific information in each sentence: the individual formulas with subscripts for different class members and superscripts for modifiers, as in $AV_p^rC_y$, for "Lymphocytes have a role in the production of antibody." The superscript *r* indicates participating in production as against actually producing.

We find, in this particular sublanguage, tightly knit sentence sequences marked by a colon, as in GJB:AVC, for "Antigen injection is followed by antibody appearing in cells." Insertions are possible, as in GJB:GUT:AVC, for "Anti-

gen injection is followed by antigen moving to a particular tissue after which antibody appears in cells." Alternative paths are also possible, as in GJB:TW, for "Antigen injection is followed by a particular tissue being altered." We see how related research lines differ. In the donor research mentioned earlier we have $GJB_1 \therefore CI^{ft}B_1B_2{:}AVCB_2$, for "Antigen is injected into animal one; thereafter lymphocytes are injected from animal one to animal two; thereafter antibody appears in lymphocytes in animal two." Subscripts here distinguish the two animals.

Some analytical results

Within most papers we find differences in sentence types between the Procedures, Results, and Discussion sections, allowing discrimination between different kinds of science information. Across papers, we can locate change over time. First AVT is replaced by AVC, as attention shifts from whole tissues to cells. Later, a new sentence type, CYC, enters, when more cell types are distinguished and their similarities noted, and when the proliferation of cell names is controlled by saying that some different names are for the same cell. In this connection, we can locate unclarity, as when the proliferation of cell names is not supported by different properties—in the W class—reported for the differently named cells, with the unclarity being finally recognized by CYC sentences stating that these are names for the same cell.

We can locate the disagreements between papers and see their structural status. The disagreements appear as symbol differences at specific points in the formulas. The chief case here is that one set of papers has AV_pC_y, for "Antibody is produced by lymphocytes" or "Lymphocytes produce antibody," while another set has AV_pC_z, for "Antibody is produced by plasma cells," and $AV_p^rC_z$, for "Lymphocytes have only a role in antibody production," and even $AV_p^{\sim}C_y$, for "Lymphocytes do not produce antibody," but does not have $AV_p\ C_y$. The contradiction between $AV_p^{\sim}C_y$ and AV_pC_y is overt.

We can also locate the resolution of this disagreement, when $C_yY_cC_z$, for "Lymphocytes develop into plasma cells," appears in the final papers. Sentences of the form CY_cC, stating that one cell is a later stage of a previous cell, were becoming frequent in the later papers as many cell names and cell-stage names appeared in the course of various experiments. But the two contenders for antibody production, C_y and C_z, had never appeared in the context ——— Y_c———; that is, the development was not recognized as reaching from one antibody-producing cell to the other. When both cells were shown to be producing antibody, the explanation—that they were in the same cell line—was expressed by extending Y_c to the pair of C_y and C_z: $C_yY_cC_z$.

A SOCIAL SCIENCE EXAMPLE: SURVEY INSTRUMENTS

The social sciences are in general not immediately amenable to sublanguage analysis, largely because they are wide-ranging in their topics, and their discussions readily extend into related fields or into examples from daily life. The language of some types of social science survey instruments, however, is restricted in the necessary ways. Within each instrument and within different instruments in one area—for example, that of income and wealth—only a relatively small number of words are used, and they are

used in very few combinations with other sets of words. With the word classes represented by symbols, the questions constructed from them can be mapped to symbol sequences, sublanguage formulas.

Application of the method

The analytical procedures applied to immunology texts have in fact been applied to small samples of the instrumentation used in three major national survey series: the Survey of Income and Program Participation, the Panel Study of Income Dynamics, and the National Longitudinal Surveys of Labor Market Experience. Patterns of word-co-occurrence were studied, in order to discover the classes of words, and sequences of these, appearing regularly in questionnaires employed in these series. The main word classes found are for the subject of the question, generally the respondent in the survey; for verbs indicating relation to employment or other income sources; for words for work or other income sources; and for the other categories of information sought by such surveys: duration of employment, amount of pay, and so forth. Three main question types, constructed from these word classes, were found; they appear in the survey questions in various combinations, joined by linguistic connectives to form more complex questions. As in the immunology material, instances of the sentence types may carry modifiers, also of specified types, that qualify the information.

Some analytical examples

The three main question types ask about employment, about welfare program participation and nonemployment income, and about conditions relevant to qualification for program participation. Table 1 shows two examples, representing the information requested by typical survey questions. We give the question formula, a description of each of the constituent word classes, and the question words, put into a standard order for intersentence comparison. The first example, "What was the main reason R could not take a job during those weeks?" typifies questions asking about employment. The second example, "Have you ever received Social Security disability benefits?" exemplifies questions asking about welfare programs and nonemployment income.

In these examples, the word classes are represented by words and mnemonics rather than by single letters, but the principle is exactly the same as in the immunology case. The method of analysis is based only on the occurrences of the words, and not on conceptions of their meanings or any other considerations contributed by the analyst. Nonetheless, this method makes it possible to code, store, and compare the information in sentences and whole documents. For example, once the questions in a group of instruments have been mapped to their formulaic representations, one can easily locate all questions that utilize a particular type of information and also those that utilize these types in particular combinations.

COMPUTABILITY OF THE METHOD

To generalize, an important property of sublanguage structures and science-language formulas is that they are discoverable by the application of fixed procedures of finding the regularities of word combination in a field and not on the basis of subjective judgments or of

TABLE 1
SUBLANGUAGE ANALYSIS OF SOCIAL SCIENCE SURVEY QUESTIONS

Question 1	What was the main reason R could not take a job during those weeks?					
Question formula	SUB	VJOB	JOB	DUR	WHY	
Word-class description	respondent	verb for job	job	duration of job	reason	
Question words in standard order	R	could not take	job	during those weeks	for what main reason	
Question 2	Have you ever received Social Security disability benefits?					
Question formula	SUB	VREC	INC	TYPE	SOURCE	DUR
Word-class description	respondent	verb for recipience	income	type of income	source of income	duration
Question words in standard order	you	have received	benefits	disability	Social Security	ever

semantic properties that lie beyond the capacity of computers. Because of this, computer programs can be developed to represent the sentences of documents in the field by the appropriate formulas and to subject the information represented in the formulas to various sorts of processing. Writing such programs is a daunting task and means adding a major capability to computers. But it has been done, in early stages, for some fields of science and medicine,[2] and also for the social science survey instruments just described.

SUBLANGUAGE COMMUNICATION

We can now consider what this new informational representation means for science communication and for international science cooperation. First, it means that methodologically unified "grammars" of science are possible. That is, there can be languagelike systems in which anything that cannot be said in the science, as irrelevant, meaningless, or grossly nonsensical, is ruled out as "ungrammatical." Second, it means that a computable representation of the specific information in scientific documents—including, in principle, conversations—is a possibility. Third, it means that in each field, scientists of whatever national language are even today speaking a global language, although it is expressed in the sounds and grammatical requirements of their particular languages. That is, while each science language can be viewed as a sublanguage within the spoken national language of each scientist, these languages as used in science communication can also be viewed as just particular pronunciations of the global science language.

It may be possible to overcome some of the communicational differences between the global science-language of a field and its divisive national-language pronunciations. It is unrealistic to expect scientists of whatever national linguistic backgrounds to begin to think, speak, write, hear, or read the statements of their science in formulas. Even in mathematics this is not quite what is done. But there are many ways in which the formulaic representation can be a communicational aid. It would be no great task for scientists to become acquainted with the formulas of their field, once these have been obtained by analysis from texts in that field. Scientists' articles and their papers for international conferences could be accompanied by abstracts or subtitles written in formulas or by formulas for the lead sentences of the main paragraphs. The international nature of the science formula language might also serve to limit the dominance exercised within the world of science by speakers of the leading national languages.

Implications for telescience

Aside from facilitating information processing for the needs of individual scientists and facilitating international cooperation among scientific workers, the fact that the science language is computable offers enhanced capabilities for science institutions. It makes possible computer processing of language data from articles or research reports, including language material added to standard data forms. It also opens a way for the construction and maintenance of data bases and other accessible and processible archives of information, even in real time, beyond what has hitherto been thought possible for computers. It offers certain safeguards on the privacy

2. See Naomi Sager et al., *Medical Language Processing: Computer Management of Narrative Data* (New York: Addison-Wesley, 1987).

and confidentiality of data in processing, as when indications of sources in documents are unseen in computer processing of information in the documents and not in human processing. The fact that differences between languages of origin are irrelevant to the sameness of science formulas means that remote-access multinational archives and data bases can be maintained in real time with little more difficulty than single-language ones, once the translation to the formulas has been worked out for each participating language. In such ways the solution promised by sublanguage analysis to the problem of international communication in science would at the same time provide facilities for the computer processing and retrieval of science information on a large—potentially a global—scale.

This solution would, of course, not suffice to overcome the basic contradiction between the rational and universal character of science, with its implication of the need of all interested human groups for free and equal access to scientific information, and the actual control of science as a political and economic resource by the nationally and socially distinct possessors of social power. The problem that Leibniz experienced is one not of difficulties in communication but of differences in interests. Any development in the direction of freer communication, however, at least points in the direction of a more egalitarian mode of creating and utilizing human knowledge.