ACADEMIC PRESS

Available online at www.sciencedirect.com

SCIENCE @ DIRECT®

Journal of Biomedical Informatics 35 (2002) 215–221

Journal of Biomedical Informatics

www.elsevier.com/locate/yjbin

The structure of science information

Zellig S. Harris[†]

Received 1 March 2002

Abstract

The organization of information within science can be investigated in a principled way through analysis of science language. The restricted use of language in science enables description of the informational structure of science and of particular subfields, with strong similarities to structures in mathematics and programming languages. This result rests on decades of research into the relation between form and content in language, based on an information-theoretic approach to the structure of information. Examples are provided from immunology and the social sciences. Practical applications include storage of science information in databases, indexing the literature, and identification and resolution of controversy.

1. Form and content

The search for a language of science is not new, and is known especially from the work of Rudolf Carnap [1]. Recent work has shown that there is a particular structure to science information in general, and to the information of each subscience in particular. Such structures are exhibited in the restricted use of language that carries the information. This restriction of language use in science is a special case of the form-content correspondence that characterizes all information.

Consider the differences between the ordinary use of language (which we will call here, colloquial), and science writing, programming languages, and mathematics. In colloquial language we have word-classes such as nouns and verbs and adjectives, with sentences being formed by particular sequences of these words. These sentences present statements about the world, something which, for example, musical notation could not do no matter what values were given to its symbols. The statements may be true or false or non-sensical, and may be about anything. In the case of science writing we find the same basic structure, except that for each subscience there are particular subsets of nouns that occur with particular subsets of verbs or other words: in a biochemical field we may find nouns for molecules and their parts, and nouns for cells and their parts, appearing in specified grammatical relations to each other. If we now turn to programming languages, we find that each one has particular sequences of symbols or words that are defined as making a statement, and specified kinds of sequences of statement-types that make a program. The result of these restrictions on symbol-occurrence is a computation device. Finally in mathematics there are wellformedness conditions on symbols to make what is in effect a sentence, and then complex conditions on sentence sequences, called proof, which supply the “not less true than” meaning that connects the output of each such sequence to its inputs (A implies B means that B is not less true than A). In all of these cases, if the structural conditions are altered, the system will no longer do its work or carry its information.

For language-like systems except the colloquial, we can define the symbols or words, and the operations on them in a metalanguage of that system: the items of mathematical notation, or of programming languages, are defined in, say, English. For colloquial language, however, no external metalanguage is available. Any language in which we could describe and define the words and word-classes of English, and state what sequences of these constitute sentences, would itself have to consist of words and sentential sequences of words, in order for it to be able to speak about English. That language would thus have to consist of the very structures that it is being used to define. This circularity can

[†] Deceased.

1532-0464/02/$ - see front matter © 2003 Published by Elsevier Science (USA).
doi:10.1016/S1532-0464(03)00011-X

be circumvented. The entities and operations of a language can be directly exhibited within its utterances, and thus recognized without a metalanguage, provided that not all combinations of the entities occur, or at least that not all are equiprobable: e.g., *Arrive entered* does not occur at all in English, while *People entered* is more likely to occur than *Vacuum entered*. When the departures from equiprobability of combination are sufficiently massive, and when the utterances that one hears or reads are sufficiently numerous, there is no need for cryptographic calculations. Mere exposure to the utterances can familiarize a person with the entities of the language and with the constraints on their combination.

It is in this way that the structure of a language can be conformed to even without the speakers explicitly knowing the grammar. This experiential basis for knowing the structure of natural language leaves us with two responsibilities. One is that we have only limited freedom in describing a language: the essential grammar is forced upon us by the constraints on combination, and is not a matter of our inventing a model. The other is that the grammar is now seen to be a statement of the constraints on combination, i.e., a presentation of the redundancy of the system. It follows that the formulation of the grammar should add no redundancy of its own to that which is being described. Hence of any set of grammars adequate precisely for the utterances of a language, the least redundant grammatical formulation is the most pertinent.

2. Structural basis of information

We now consider the structural basis for the information carried by colloquial language. This case is the most relevant to science information, for most science information is carried by specializations of colloquial language. We begin with the least redundant grammar. This consists of a set of constraints on the equiprobability of entities, the most efficient set that is able to produce all utterances of the language and only these. The fundamental constraint, and each added constraint, creates a specific and unchanging contribution to the informational capabilities of the system. Before presenting this fundamental relation, we note that the phonemes (comparable to letters), words, and sentences, which satisfy the relations listed below, can be established by objective procedures and stochastic processes, without reliance on such intangibles as meanings [2]. Whence then come the meanings? Some words have fixed meanings, independent of their combinations. The other word-meanings and the grammatical meanings come from the constraint relations presented below.

The fundamental constraint, that uniquely creates language, appears when in a set of elements (symbols or words) the occurrence of each word in an utterance depends on the occurrence there of an element—any element—of some stated subset of words: the presence of arrive requires (depends on) the presence of a noun (*John arrived*); *rent* requires two nouns (*I rented a room*); *probable* requires a verb (e.g., *arrive* in *John's arriving is probable*; there is no *John is probable*). Similarly in mathematics, = requires two variables or constants. The requirement condition creates sentences as a partial order of words; it makes certain sequences of words or symbols in language, in programming languages, and in mathematics into well-formed sentences while other sequences are not. This dependence inherently admits of a meaning for the relation of a word to that which is under it in the partial order: "to operate on," "to be a predicate on," "to say about," thus, above, *arriving* is predicated about *John*, and *is probable* is said of the *arriving*.

In colloquial language (but not in the other systems) this dependence has a mathematical property: each class of words depends not on a particular list or meaning of other words, but on just the dependence property of the other words. "Zero-level" words (*John*, *room*) are those that depend on the null class, i.e., on nothing. "First-level" words such as *enter*, *arrive*, *rent* depend only on words that depend on null: one zero-level word under *enter* or *arrive*, two under *rent*. For "second-level" words at least one of their required words requires something: *probable* requires one first-level word, *entail* requires two (*John's arriving entails my renting a room*). Thus, it is not intrinsic properties of sounds and meanings that determine the possible word-sequences of sentences. Rather, the word-occurrences are characterized only by a stated relation among them, namely their depending on the dependence (of words), with anything that satisfies this relation being a possible sentence. The fundamental constraint of language thus creates a mathematical object. This last property cannot be fortuitous. Indeed, in the absence of an external metalanguage, natural language could only have arisen as a self-organizing system, creating sentences in a world that up to then had had no sentences but only words or variegated word-combinations.

The requirement relation states that for each word there are some words that have positive probability of occurring under or over it, while the other words have zero probability there. In language, but not in mathematics, a further property holds within this requirement. For each word, we find roughly stable inequalities of probability among the words in its required (positive probability) set: in the second requirement position (the "object") under *rent*, the probability (or, less formally, likelihood) of *room* is greater than that of *city* (in *I rented a city*), which in turn is greater than that of *universe*. For a given word we find that, of the words with non-zero probability under it, the probability of some is vanishingly small (under *rent*: e.g., *universe*), and of others larger than average (under *rent*: e.g., *room*); there may also be one with highest probability there. The meaning of a word is indicated, and in part created, by the meanings of the words in respect to which it has higher than average probability: *dog* has such likelihood under the set including *ate*, *attacked*, *died*, *barked*, *buried a bone*; *cat* has it under the set including *ate*, *attacked*, *died*, *climbed a tree*, *drank milk* [3].

On these roughly stable inequalities we find another operation, which is important only for the overt form of sentences. Words with highest probability in respect to another word, or which otherwise can be shown structurally to have highest expectancy, add little or no information. In this requirement-environment these words are reducible [4]. The reductions are, for example, to being an affix or even to zero: e.g., the zeroing of the second *John took* in *John took math before John took physics* reduced to *John took math before physics*. The status of the word in the sentence does not change in its requirement relations and its meaning, even when it is zeroed. The sentence is altered only in its physical shape and not in its information. These reductions constitute a set of paraphrastic partial transformations on the set of sentences.

The requirement relation and the likelihood inequalities and the reductions constitute the grammatical framework of colloquial language. On the sets introduced here, various mathematically formulated decompositions, partitions, and mappings yield further sets and relations that constitute the specific structures of each language. The whole analysis can be considered a type of applied mathematics, insofar as that field includes not only the developing of calculational methods for various sciences but also the finding of cases where mathematical structures are satisfied in the real world [5,6]. The information carried by this structure consists of predication and word-meaning (from the first two constraints), which gives the sentences their character as statements. Other grammatical meanings can be shown to be derived from these by the reductions (e.g., questions are derived from *I ask whether—*). In all of these form-content relations in sentences and sentence-sequences, we see a structured composition of meaning and information, as given by the contributions to syntactic structure.

In the mathematical theory of communication (Information Theory), what was investigated was the information capacity in a system or a channel, and its relation to the amount of actual or possible departures from randomness therein [7]. In the theory presented above, we again find information characterizable in terms of measurable departures from equiprobability; but here the specifying of constraints enables us to locate each such departure and so each contribution to the information. What we have here is thus an information-theoretic approach to the structure of information, as against solely the amount of information.

3. Science sublanguages

When a set of texts is taken in a single subject matter, especially a science, the texts reveal a shared structure differing in a principled way from that of the language as a whole. The basic difference lies in what it is that words require. What is common to the texts of a given subject matter is that first-level words of a given subset require zero-level words of only a particular subset. In biochemistry, *is synthesized* (and other words of its subset) can require a word in the subset of *antibody* while *undergoes mitosis* requires cell-names; other nouns are excluded from the requirement of these verbs. This differs from colloquial language, where a verb, e.g., *sleep*, accepts more-probable-than-average words from its requirement set, such as *John* and *dog*, but, with lesser likelihood, also *tree* and *earth* and (at cost of making non-sense) any other simple noun.

We thus obtain for the science several statement-types (e.g., antibody names with their predicates, cell names with theirs), instead of the single original sentential type created by satisfying the whole-language requirement. This difference has semantic effects. For one thing, irrelevance and non-sense (from the point of view of the given subject matter at the time) are largely excluded, though falsity is not. For another, it becomes possible to recognize fixed canonical forms for information, and more generally to find the informational structure of the science or subscience. There may even be possibilities of characterizing the causal relations that are relevant to the given science.

As an example, we give a brief sketch of what was found in analyzing representative articles of the early period of cellular immunology [8]. This summary applies to c. 1935–66, when a central problem was to determine which lymphatic cell produced antibody. There was a controversy as to whether it was the lymphocyte or the plasma cell; it ended with the evidence that both produced antibody, and with the realization that these were different stages of the same cell.

The following major word-sets were found, as having different requirement statuses. Zero-level words:

G: e.g. *antigen*, *bacteria*, *sheep blood cells*.
B: e.g. *ear*, *rabbit*.
A: e.g. *antibody*, *agglutinin*, *immune globulins*.
T: e.g. *lymph nodes*, *serum*, *adipose tissue*.
C: e.g. *lymphocytes*, *plasma cells*, *reticulum cells*.
S: various intracellular structures.

First-level words of the following sets require the above in various ways:

J: on **G–B** (*injected into*, as in *Antigen is injected into the ear*)
I: on **C–B–B** (*injected into... from*, as in *cells were injected into rats from non-immunized rats*)
U: on **G–T** (*reaches*, *concentration in*)
on **G–C** (*stimulates*, *uptake by*, *sensitizes*)
V: on **A–T** (*visible in*, *distributed in*; *formed in*; *drain into*; *pass through*)
on **A–C** (*found in, contained in*; *synthesized by*; *adsorbed to*; *secreted by*)
W: on **T–** (*react*, *affected*; *swollen*, *inflamed*)
on **C–** (*react*, *change*, *develop*; *enlarge*; *present*; *multiply*, *divide*, *undergo mitosis*)
on **C–T** (*present in*, *persist in*; *transferred from*, *drain from*; *pass through*)
on **S–** (*in parallel orientation*, *rough*, *clustered*, *basophilic*)
Y: on **C–C** (*is same as*, *has some similarity to*, *is called*; *formed from*, *derived from*; *develops into*)
on **C–C–C** (*bridges the gap between ... and*, *differentiates through ... to*)
on **S–S** (*is in the form of*, *intermingles with*)

If we insert the first-level word after its first required word, we see in the list above the major sentence-types of this material (e.g., **GJB** for *Antigen is injected into the ear*). There is also a strong constraint on short sequences of these sentence-types. Most occurrences of **V** and **W** sentences are explicitly (or implicitly, by zeroing) followers of **J** sentences, as in

GJB : AVC (*Antigen is injected into a rabbit. Thereafter antibody appears in the lymphocytes.*)
GJB : TW (*Antigen is injected into the left foot pad. Thereafter the homologous lymph node is inflamed.*)
GJB : GUT : TW (*Antigen is injected into the foot pad. Thereafter the antigen reaches the lymph node. Thereafter the lymph node is inflamed.*)

The colon indicates a set of inter-sentence connectives, mostly expressing time order. This sequence is so common, that although the **J**, **U**, **V**, and **W** constructions are each a sentence from the point of view of English grammar, it might be most appropriate to consider the **J–U–V/W** sequence to be the characteristic statement, or hyper-sentence, of this field. Like the constraint that creates the sentence-types, the additional constraint that creates this sequence clearly represents the chain of information dealt with in these articles.

Within these word-classes there are subsets distinguished by their requiring different subsets of their required classes, or by differences in their farther sentential environment.

For example:

- in **U**: $\mathbf{U_r}$ (*reach*), $\mathbf{U_s}$ (*adhere to*, *sensitize*), $\mathbf{U_i}$ (*found in*), $\mathbf{U_d}$ (*perish in*)
- in **V**: $\mathbf{V_i}$ (*contained in*), $\mathbf{V_p}$ (*synthesized by*), $\mathbf{V_t}$ (*stored in*), $\mathbf{V_s}$ (*secreted by*)
- in **W**: $\mathbf{W_f}$ (*inflamed*), $\mathbf{W_s}$ (*basophilic*), $\mathbf{W_y}$ (*oriented*), $\mathbf{W_p}$ (*multiply*), $\mathbf{W_u}$ (*flow*), $\mathbf{W_i}$ (*present in*), $\mathbf{W_c}$ (*develop*), $\mathbf{W_a}$ (*react*), $\mathbf{W_g}$ (*enlarge*), $\mathbf{W_m}$ (*mature*)
- in **Y**: $\mathbf{Y_a}$ (*classified as*), $\mathbf{Y_i}$ (*includes*), $\mathbf{Y_c}$ (*develops into*, *precursor of*)
- in **C**: $\mathbf{C_y}$ (*lymphocyte*), $\mathbf{C_z}$ (*plasma cell*)
- in **T**: subscripts for various relevant tissues.

Such subclassification can be made to any detail desired, either to some level useful for summarizing the information, or to the point where every word with relevant meaning difference is differently subclassified. (The relevant meaning, which can be checked by the environing words, is important because words can be used in less than their full meaning: for example, *agglutinin* means more than *antibody,* but in articles on the cellular site of antibody production it is used just to indicate antibody presence.) In addition, the sentences of the articles contain words and phrases (adjectives, adverbs, auxiliaries) which modify the meanings of the main words, and which grammatically are reductions of secondary sentences. Examples in the immunology articles are:

- on various noun (zero-level) classes: e.g., *large*, *distended*, *mature*, *active*, *homologous*, *family of.*
- on various verb and adjective (first-level) classes: e.g., *not*, *begin to*, *much*, *rapid*, *increased*, *receding*, *maximal*, *play a role in*, *in vitro*.
- on the colon conjunction (second-level words): e.g., various time intervals, e.g., *three days* (*after*).

Such secondary material can be indicated by superscripts on the symbols to which the modifier had been grammatically attached: $\mathbf{GJB} :^{t} \mathbf{AV_iTB}$ for *Antigen was injected into the ear*; *three days later antibody was found in the homologous lymph node*. (All grammatical terms, such as "secondary," "verb," can be defined in respect to the requirement relation introduced above.)

Finally, many sentences in the articles consist of an occurrence of a science sentence-type as above, grammatically under a metalinguistic predicate (marked **M**) which presents the scientist's relation to the science information: e.g., *We have found that* ...; *That* ... *was not expected*; etc. When the sentences of an article are represented by sentence-type formulas with subclassification, modifiers, and **M**, the result is a formulaic record of all the information in the article (Table 1).

When a subset of a system is closed under operations of the system, the subset constitutes a subsystem. If we take sentences such as are used in a science, and operate on them with the conjunctions or the transformations of the language, we obtain again a sentence such as is used in that science. The set of such sentences, as said or

Table 1
Formulaic representation of sentences

<table>
<tr><td>It seems clear from all the evidence that the cells responsible for the synthesis of antibody shortly after the injection of a second antigenic stimulus are members of a family which arise from some undifferentiated precursor as the direct result of the stimulus.</td><td>It seems clear from all the evidence that the cells | are | members of a family WH ||| antigen | the injection of the second stimulus of || shortly after || antibody | (are) responsible for the synthesis of | (cells) ← which ||| the stimulus || as the direct result of || (Members of a family) | arise from | some undifferentiated precursor</td><td>M
$C^{w}YC^{lw}$
$GJ^{2}{:}^{e}A\,V^{r}_{p}\,C$
$GJ^{2}{:}C^{l}\,Y^{f}_{c}\,C_{b}$</td></tr>
<tr><td>The first cells which demonstrably contained antibody and can therefore be assigned to this family are large cells with a thin rim of basophilic cytoplasm and large nuclei whose appearance is indistinguishable from that of other primitive hematogenous cells.</td><td>The cells | are | large cells Which ||| (antigenic stimulus) | (the second injection of) || first (after) | antibody | demonstrably contain | (cells) ← and therefore (which) ||| (cells) | can be assigned to | this family WH ||| (large cells) with a thin rim of cytoplasm (which) | (is) basophilic WH ||| (large cells with) nuclei (which) | (are) large whose ||| (large cells') | appearance is indistinguishable from that of | other primitive hematogenous cells</td><td>$C^{w}YC^{gw}$
$GJ^{2}{:}^{e}AV^{i}C$
CYC^{l}
$C^{g}S^{-}_{c}W_{s}$
$C^{g}S_{n}W_{g}$
$C^{g}YC_{b}$</td></tr>
<tr><td>During the 2 or 3 days after their first appearance they multiply, synthesize antibodies specific for the antigen which stimulated their development, and differentiate through immature to mature plasma cells.</td><td>the large cells | multiply, ||| (antigen) | (was twice injected) || WH ← || antibody specific for the antigen | synthesize | (the large cells) ← which ||| (antigen) || stimulated || the large cells' | development, and ||| (the large cells') | differentiate | through immature (plasma cells) | to mature plasma cells during the 2 or 3 days after ||| (antigenic stimulus) | (a second injection of) || (at a time which was) first (after) || the large cells' | appearance</td><td>$C^{g}W_{p}$
$G^{w}J^{2}{:}A^{G}V_{p}C^{g}$
$G{:}C^{g}W_{p}$
$C^{g}\,Y^{ft}_{c}C^{m}_{z}\,C^{m}_{z}$
$GJ2{:}^{e}C^{g}W_{i}$</td></tr>
</table>

The middle column is a grammatical transform of the left column. Brackets enclose elementary sublanguage sentences. Material between brackets is the sublanguage conjunction marked by colon. Material before a bracket is a general conjunction (not shown in formulas) to the preceding sentence; **WH** indicates a secondary sentence which has become relative clause or modifier. Vertical bars inside brackets separate the subject, verb, and object. Parentheses indicate zeroed material. ← indicates that the preceding material is to be read in English in reverse order; forward-readable transforms exist but are more complex. The right column gives the formulaic representation of the middle column, obtained directly by writing a sublanguage symbol for each segment between bars or brackets. Superscript **w** on a host letter indicates that the host is carrying a modifier which appears as a secondary sentence, below, introduced by **WH**. Other superscripts indicate a modifier that is written together with the host. Subscripts indicate subclasses of the class marked by the host letter. The sentences are from E.H. Leduc, A.H. Coons, J.M. Connolly, J. Exp. Med. 102, 66 Par. 4 sentences 1–3 (1955) and the analysis is given on pp. 360–361 of Harris, Gottfried, Ryckman, et al. op. cit.

written, is therefore a sublanguage—of English, French, or whatever [9]. However when, in a given science, articles written in different languages are analyzed, as was done for both French and English in the immunology analysis, we obtain the same sentence-types and structures, with only small differences due to the languages. The word class and subclass symbols, and the sentence-types, are therefore not just a sublanguage of a particular language, but an independent symbolic linguistic system. Its grammar is not the same as for colloquial language. Like language and mathematics, it has the requirement structure; but like mathematics this is based on membership lists and not on the property of a mathematical object. Like colloquial language and unlike mathematics, it has a large stock of words with real-world meanings. Unlike colloquial language, it apparently does not have significant likelihood inequalities; instead it has the unique system of subclasses creating a family of sentence-types. In informational capability it yields a controlled and advanced version of what language does—namely, indicating information about the real world. It cannot do what mathematics can, but in ways to be seen below it can take certain steps in that direction.

To clarify this last point, we have to note that a symbolic science-language is more than just a convenient presentation of the ordinary-language sentences from which it is mapped. For one thing, the symbols enable us to avoid many extraneous features such as grammatical demands which may be irrelevant to the given science (e.g., tense or plurality or the verb–adjective distinction). The essential difference, however, is that the canonical formulas create a structure for the information, one that is relevant because it grows out of the regularities of the science writing itself. The symbols and their sentential structures provide an index of what information is dealt with, and if an object or a fact is present in the material we know where to look for it in the formulas.

The fact that a structured representation is constituted here makes possible various inspections and critiques. In the immunology example, we can see how the field changed during the period investigated. First, $\mathbf{AV_{i}T}$ (*Antibody is found in the lymph nodes*) is replaced by $\mathbf{AV_{i}C}$ (*Antibody is found in lymphocytes*). Then $\mathbf{AV_{p}C}$ begins to appear (*Antibody is produced by the cell*). Then as more cell types and even cell stages are distinguished and named we find $\mathbf{C_{i}Y_{c}C_{j}}$ (*Cell$_i$ develops into cell$_j$*). The controversy appears when some articles have $\mathbf{AV_{p}C_{y}}$ (*Antibody is produced by lymphocytes*) matched against others stating $\mathbf{AV_{p}C_{z}}$ (*Antibody is produced by plasma cells*) and claiming $\mathbf{A}\ V^{r}_{p}\ \mathbf{C_{y}}$. (*Lymphocytes have a role in the production of antibody—rather than actually producing antibody*) and even entering a denial in **A**

$\mathrm{V}_{\mathrm{p}}^{\sim}$ $\mathbf{C_y}$ (*Antibody is not produced by lymphocytes*). The resolution is recognizable when an article contains both $\mathbf{AV_pC_y}$ and $\mathbf{AV_pC_z}$ and also the explanatory $\mathbf{C_yY_cC_z}$ (*Lymphocytes develop into plasma cells*) which recognizes that the $\mathbf{Y_c}$ relation applies also to the two cells of controversy.

In addition to such inspection of the argument, the formulaic structure forces the writer or the reader to be explicit, or to recognize lack of explicitness, at points where the difficulty would not be noticed in ordinary language use.

4. Survey languages

The sublanguage method is clearly workable in the natural sciences, where the terminology and its interrelations are both well-defined. In the social sciences it would seem questionable, because the structure of the field is less explicit, and colloquial material from daily life can be readily introduced. Nevertheless, it has been found possible to extend the method, so long as the framework is explicit and the vocabulary relatively closed. In particular, it has been found in analyzing survey instruments (questionnaires), that it is possible to map the questions, in a manner that can be executed by computer programs, onto a family of formulaic sentence-types which are then usable for processing the information contained in the questions themselves.

In a pilot study of a small sample of surveys of income and wealth, including the longitudinal ones, predominance was found for a few sentence-types in the sublanguage sense: e.g., for kind and duration of employment, for income and program participation, and for certain conditions affecting this such as health. Using sublanguage parsing programs, the survey questions can be mapped onto these canonical types, so as to form a database. On the database of the pilot study various queries were executed, such as:

In this survey what is the range of relations to a job that the respondent might have?

- List all questions that relate the respondent's non-wage income to the respondent's qualifying condition.
- Generate a keyword index of questions on income and program participation (or on stated other categories).
- Find all questions in which the respondent has some condition related to employment which qualifies the respondent for income or program participation.

The structuring of the information in the database made it possible to obtain complete and relevant answers to the queries. Once the sublanguage formulas have been established, computer programs can carry out various kinds of structured information storage and processing, including summarizing and comparing the questions, especially with regard to redundancy and alternative wordings, within one instrument or several.

5. Applications

Many possible applications arise from the sublanguage method, some primarily of a research nature, and others of practicable development. One research is to see what are the possibilities of obtaining standard notations for science languages, not by fiat but by boiling down from actual use, somewhat as happened for mathematics in the 16th century. Another is to relate the information structure of a science to anything else that characterizes the field, in order to reach if possible a "structure" of the science. A third is to critique the conceptual system of the science in respect to the formulas of its operative statements, in order to see if the concepts exceed the needs of the system: for example, there may be such excess in teleological vocabulary such as the "information" terminology of the genetic "code," which gives an end-point interpretation instead of describing the biochemical mechanisms.

Of greater applicability is the investigation of the structure of individual science languages. One kind of investigation is to spot trouble or the process of change, by seeking unclarities or inconsistencies in the interrelation of formulas in texts. Another is to see how tabular or other two-dimensional displays can represent the data (or the Result statements) of articles, for human inspection or for computer processing. As to the argument structure in articles, it may be open to regularization, because it consists primarily not of logical or other new statements, but of statements from the Result section; the Result sentences are modified in certain ways (e.g., by generalization), and combined via particular causal and other connectives and logical operators. This means that an argumentation is some kind of controlled sequence of Result statements. Hence one might investigate successful examples of such sequences so as to judge what conditions on the sequence permit one to assert the correctness or plausibility of the last sentence (the conclusion) given the correctness of the input sentence—all this as a weak analog to proof in mathematics. More simply, one could investigate science languages to see what sentence-types are common to all or many of them, such as the metalinguistic and the statements of quantity. One can show, for each science, what if any are its prior sciences, because these occupy bottom positions in the requirement hierarchy of the sentence. One can also investigate the differences between the formulas of neighboring science languages to see how distinct are the things they talk about or the way they talk about them; that is, to see to what extent they have become separate subsciences.

Since, in a least grammar, the information representation is very close to the word-combination analysis of sentences, which by its nature can be carried out by a computer program, a wide range of information processing from science records and articles is possible on a computer. Computer analysis of unedited texts, and mapping of the information into a database, has been carried out, chiefly on narrative patient records, at the Courant Institute of Mathematical Sciences of New York University [10]. Such work includes storing the information in tabular or otherwise standardized form on the basis of the regularities in the material itself. It can include preparation of summaries or of word-pair indexes, which can list for any given word, what required words appear with it: these are the word pairs that carry informational contributions. It can also include various forms of fact retrieval. Once a formulaic representation is developed for a sublanguage, computer processing could find instances of the individual formulas (with subclassifiers and modifiers) that represent whatever information is being sought—this because of the intrinsic locating of information in the structure of the formulas.

References

[1] Carnap R. The logical syntax of language. London: Kegan Paul; 1937. first German edition 1934.

[2] Harris Z. Mathematical structures of language. In: Interscience tracts in pure and applied mathematics, 21. New York: Wiley-Interscience; 1968.

[3] That the grammar of a language can be built up from the specific combinations of words in sentences is seen in the computer tabulation of word combinations and word forms in, e.g., Gross M. Methodes en syntaxe. Paris: Herman; 1975.

[4] That the reduced words are indeed the most probable in their partial-ordered position in their sentence is shown by two independent procedures: (i) informant eliciting, (ii) considerations of maximizing regularity, i.e. of least grammar.

[5] A presentation of the above theory, with a detailed English grammar derived therefrom, is given in Harris Z. A grammar of english on mathematical principles. New York: Wiley-Interscience; 1982.

[6] A parsing algorithm based on this theory has been developed: Johnson S. An analyzer for the information content of sentences. Doctoral Dissertation, New York University. (Dissertation Abstracts International, 1987:vol 48-11B; 3340–601).

[7] Shannon C. The mathematical theory of communication. With an essay by Warren Weaver. Urbana, Illinois: Univ. of Illinois Press; 1949.

[8] A full report, showing the procedure for representing the sentences of the articles in canonical formulas, is given in Harris Z, Gottfried M, Ryckman T, Mattick Jr P, Daladier A, Harris TN, Harris S. The form of information in science: analysis of an immunology sublanguage. Reidel, Dordrecht: Boston Studies in the Philosophy of Science; 1989.

[9] See, e.g.,Kittredge R, Lehrberger J, editors. Sublanguage: studies of language in restricted semantic domains. Berlin: de Gruyter; 1981.

[10] Sager N, Friedman C, Lyman M, MD, and members of the Linguistic String Project, Medical language processing: computer management of narrative data, Reading, MA: Addison-Wesley, 1987.