CHAPTER B

LANGUAGE AND DOCUMENTATION LANGUAGE

• Language
• Metalanguage
• Language of Science
• Documentation Language
• Documentation Language as Metalanguage
• Features of Document and Linguistic Analysis
• Components of Documentation Language
Language is a means of communication and thinking; for dissemination of information and to an extent the creation of it. "In information and documentation, language plays a very special role... However, it is not language which is the object of study here, but rather the use of language to formulate and communicate ideas and opinions" (10). A study of the salient features of language is necessary for the design of information systems in general and indexing systems in particular.

B1 LANGUAGE

Language is a system of symbols serving the means of human communication, thinking and expression. By means of language the world is cognized. In language the self-consciousness of an individual is objectivized.

B11 Functions of Language

Language has three or four roles or functions:

1. the biological - identification of strangers and "clansmen" by means of the spoken language;
2. cultural/historical - metaphors, special phrases linked to historical experiences and culture.
3 communicative — interhuman conveying of information; and
4 cybernetic — rationally constructed algorithmic languages for computer programming (11).

B11.1 The primary function of language is communication between individuals. "It is difficult to imagine any satisfactory definition of the term 'language' that did not incorporate some reference to the notion of communication" (12). Language is a specific social means of information storage and transfer, as well as of controlling human behaviour. It is a purely human and non-instantive method of communicating ideas, emotions and desires by means of a system of voluntarily produced symbols. Only the communicative aspect of language plays a part in documentation systems, "the essence of which is storage of information for use and manipulation by an individual" (13).

B11.2 One of the features of modern investigation of language is the recognition that 'ordinary language' — be it written or spoken — is only a member of a class of coherent symbolic communication systems. All conceivable symbol systems have been considered as languages. For instance: the language of music, the language of dance,
the language of biological code, the languages of computer programming etc.

But fulfilment of the communication function alone cannot be considered as a necessary and sufficient requirement for elevating a symbol system to the rank of 'ordinary language'. Exchange of information may be effectuated not only with 'words' but also with other symbols. Also exchange of information may take place between a human being and a computer and also between inanimate mechanisms - between two computers.

Many phenomena of the physical world can be regarded in terms of receipt and transfer of information. Even photoelectric effect can be regarded as an information exchange. But these do not become 'ordinary languages' and are not within the domain of linguistic studies. These phenomena do not fall within the purview of computational linguistics too, which is again concerned with ordinary languages and the computational methods applicable to them for specific purposes such as machine translation. The essential functional characteristic of ordinary language (hereinafter referred to as just 'language') is its role in information reduction or synthesis, abstraction, generalisation; in elaboration
or expansion, and particularisation or concretisation — in short thinking.

B12 Structure of Language

B121 Signs/Symbols

First, a language is composed of a plurality of signs — articulate sounds as signs of ideas.

B121.1 Second, in a language, each sign has a signification common to a number of interpreters.

B121.2 Third, signs constituting a language must be 'common signs', that is, producible by the members of the interpreter-family and have the same signification to the producers which they have to other interpreters.

B121.3 Fourth, the signs which constitute a language are plurisituational signs, that is, signs with a relative constancy of signification in every situation in which a sign of a sign-family in question appears. A sign in a language is a sign-family and not merely an unisituational sign-vehicle (14).

B121.4 A sign and its meaning do not completely cover each other, Their boundaries do not coincide in all points. One and the same sign has several functions, one and the same meaning is expressed by several signs. "If signs
were fixed and each of them fulfilled only one function, language would become a mere collection of labels" (15). It is not enough for the perfection of language that sounds can be made signs of ideas, unless those signs can be so made use of as to comprehend several particular things: for, the multiplication of signs would have perplexed their use, had every particular thing needed a distinct name to be signified by. To remedy this inconvenience language had yet a further improvement in the use of general terms, whereby one word was made to mark a multitude of particular existences.

B121.5 Fifth, the signs in a language must constitute a 'system' of interconnected signs combinable in some ways and not in others in order to form a variety of complex symbols of different levels (14). To be admitted as language, just signs and rules restricting the ways in which they may be combined are not sufficient. The different signs should form a hierarchy semantically. 'I see two things: a chair and some other furniture' is not an acceptable sentence, ('I see two things: a chair and some other furniture' is acceptable). Semantic hierarchy among the different signs may be viewed as one of the conditions necessary for regarding a system of signs/symbols as a language.
Symbol System

The structure of language starts with

1 subelementary linguistic symbols: the letters or alphabets, morphemes (a morpheme is a meaningful part of a word — the root and the affix, prefixes, suffixes etc.);

2 the elementary symbols: the words (a word is a symbol for one of the smallest completely satisfying bits of isolated 'meaning' — a fragment of a text between spaces);

3 phrases (a phrase is a fragment of a text between two punctuation marks);

4 clauses and sentences (a sentence is a fragment of a text between fullstops);

paragraphs and so on.

The complex symbols constructed from simpler ones form a hierarchical symbol system of several levels. Phenomenologically thinking is a process of constructing complicated symbol systems from simpler ones which is outwardly reflected in the hierarchical structure of language (15). It is reflected even in the sequence of sentences, paragraphs and even larger portions of text. A text does not consist of a sequence of unrelated sentences.
If it did, it would not convey the intended message properly. In fact, every preceding sentence thrives to fix the context for the next sentence, helping the precise understanding of the following sentences, infusing a kind of hierarchy among them. "It should be acknowledged that the logical hierarchy of statements exists in the language, but it is so concealed that in practice it cannot be directly observed" (16). The narration of events in chronological sequence in texts is an example of this kind of logical hierarchy.

B13 Grammar of Language

In order to make operative the rules for constructing complex logical structures using the simpler symbols, the symbols (words) are viewed as belonging to grammatical categories or parts-of-speech, such as noun, verb, adverb etc. The rules of syntax (grammar) are defined with respect to these categories. But Chinese makes no use of the formal categories devised for the Indo-European languages. So also the polysynthetic languages of American Indians.

"Symbols and grammar are, of course, the structural elements of language; they are clearly seen in the majority of the symbol systems perceived as
B14 Characteristics of Language

B141 Interpretablility

A sign system has a right to be called a language, if it can be interpreted into another language. Speaking to a foreigner we interpret our mother tongue in the system of another language, and actually it is not a translation but a mere interpretation. This is because of the metaphors, special phrases, idioms etc., linked to culture and historical experiences present in language.

B142 Polymorphism.

Consider the following three phrases:

1. This is a good strawberry.
2. This is a good lemon.
3. This is a good carving knife.

It is readily seen that the word "good" has quite different meanings in these phrases. A good lemon should be sour; a good strawberry should not be sour; and a good carving knife should be sharp which has nothing to do either with the quality of lemon or with
that of strawberry. Both in everyday language and in many other languages, every symbol is connected in a probabilistic manner with a variety of meanings. One of the characteristics of polymorphism is that of synonymy. This diversity of everyday language is considered its most essential characteristic and is no longer regarded as an index of its deficiency. It is due to its polymorphism that natural language is richer than any artificially created one.

B2 METALANGUAGE

Languages with highly developed logic emerge up, to constitute metalanguages whereby one may judge the correctness of statements made in the object-languages. Our everyday language is a metalanguage in relation to the "language" of things surrounding us. In terms of everyday language, we operate not with things but with their names. Making judgements about the things of the outer world, we try to arrange them in some consistent structures, which is equivalent to searching for logical foundations of the world of things.

B2.1 Generally the metalanguage is formalized to a greater extent than the object language. A metalanguage
should be so rich that everything stated in terms of
the object language could be said in the metalanguage;
particularly, it should have the means for constructing
names of the object-language. "The metalanguage of
everyday speech uses the same sign system and the same
logical means as the everyday language" (18).

B31 Components of the Language of Science

Science is a linguistic or symbolic representation
of experience. Scientific development is reflected
in the development of the language of science. In the
language of science the symbols and grammar of ordinary
language are used. New terms (or new combination of old
words) and new meanings ascribed to old words borrowed
from everyday vocabulary are introduced. This gives an
esoteric character to specific languages of science:
they prove comprehensible only to the initiated. Terms
in science are closely connected with the respective
theoretical concepts, though on the surface, many terms
seem to be more than names of some objects or phenomena.
For example, the 'Raman Effect' would seem to be a name
of some physically observed phenomenon. In fact, the
meaning of this term becomes clear only through the
understanding of the theory of this phenomenon.
No doubt, in the language of science, there are some terms which may be clearly defined. But the meaning changes with time together with the development of scientific concepts. The meaning ascribed to the word 'atom' now, differs considerably both from that ascribed by the ancient Greeks and from that used at the beginning of this century. The meaning ascribed to terms change in different theories though they may be closely related. Both 'Classical Mechanics' and 'Relativity Theory' make use of such terms as 'mass' and 'length', but they are interpreted differently.

Polymorphism of the Language of Science

Scientific terms have a more polymorphic character than the words of ordinary language. The term 'model' (quite a fashionable term now a days) has been studied by Chao Yuan-Ren (19). He has given a list of several synonyms which is characteristic of 'model' and non-synonyms which are notions contrasted to 'model'.

Scientists express new notions with the help of rather unusual combinations of old, well known, and familiar expressions. Scientists have permitted rather different senses for old terms with the emergence of new
theoretical concepts. In science, theories are continuously changing, but the change does not cause a waterfall of new words. The new phenomena are interpreted, through the old familiar ones, through old words whose meaning is slightly continuously changed. For instance, everybody knows the common meaning of the word 'replica' which is borrowed from French. In mathematical statistics, there are such terms as 'replica', 'fractional replica' and 'regular replica' which have specific meaning. In optics, 'replica' is a copy of a diffractional lattice prepared in a special way. In the languages of science, new concepts emerge, and old notions are often assigned new meaning. Because of its continuously changing nature, specific languages of science are accessible only to those working in the field and thus constantly interacting with the informational flow in science. It can even be said that, the emergence of a new independent discipline must be accompanied by the emergence of a new specific language or a dialect of the discipline.

B4 DOCUMENTATION LANGUAGE

The functions of a bibliographic information system are: collecting, organizing, storing, citing and disseminating of documents or the information/ideas.
recorded in the documents. Such information processing activity can be generally divided into the following stages:

1. Collection of information/documents embodying information with the purpose of supplying most fully, the necessary information/documents in accordance with the request of the users;

2. Information analysis (analytic and synthetic processing of information/document surrogates/documents themselves), categorization and systematization of the incoming flow of information and

3. Storage and retrieval of information/documents and the development and use of information retrieval procedures with the use of modern means for achieving the desired results.

B4.1 Of the above mentioned three stages, the second one, namely, information analysis or document analysis, decides the efficiency of the system. It deals with systematizing or organizing the information/documents accompanied by classification and indexing of the information content of documents; the creation of new
secondary documents; or document surrogates on the basis of certain rules and procedures depending on specific tasks of information practice, to form the index file or enquiry file; and the retrieval procedures constituting the interface in the information retrieval process.

B4.2 A model bibliographic information system is given in Exhibit-1. Documents are selected, received and their information content analysed to determine the subject categories into which they are to be classified as well as the index terms associated with the documents, according to a specific language having grammar, called information/documentation language. After the completion of this analysis, there is a branching which allows for more than one method of 'organizing' the 'files'. The documents themselves are stored physically in a pre-defined sequence constituting the 'storage file'. Since only one physical arrangement of documents is possible without expensive duplication of the documents, alternative paths of access to them are provided through the 'index file' which contains the index terms determined at the analysis step.
EXHIBIT - 1: GENERAL MODEL OF BIBLIOGRAPHIC INFORMATION SYSTEM
B4.3 The user requests information. He interacts with the index and/or storage file. To do so, he must convert his request for information into a well defined search question to which the system can respond. The formulation of the search question again has to be according to or in the information/documentation language. When the user is satisfied with the documents which have been located in response to his request, he reads them to obtain new facts and ideas. Presumably, he will now become a generator of new documents which will in turn find their way into the system.

B4.4 The nucleus of the bibliographic information system is the information/documentation language. According to documentation language, the documents are analysed. The creation and organization of the 'index file' and the setting up of the retrieval procedures are all based on the analysis of documents according to the prescriptions of the documentation language.

B4.5 Natural language is a tool for thinking and is a means of communication. In such a language there are synonymous correspondence between words and meanings.
The meaning of words, sentences and even speech may change with time. Hence, the use of a natural language as such, in information retrieval systems for the description of semantic contents of documents is associated with difficult problems. In order to overcome these problems, special artificial languages called documentation languages or information languages are used. These artificial languages are formalised auxiliary languages, created by information specialists for use in bibliographic information storage and retrieval systems. Because of their formalized structure they are regarded as metalanguages for information organization and retrieval.

B4.6 General classification schemes such as UDC, DDC, BC, CC etc., are documentation languages. Uni-term index coupled with the retrieval procedure using 'roles' and 'links' is also a documentation language. Key word indexes associated with thesauri and Boolean operators for retrieval is a documentation language. The different subject indexing languages such as, Chain Indexing, Preserved Context Indexing System, Postulate-based Permuted Subject Indexing System, Relational Indexing System etc., are all documentation languages.
In documentation languages, the meaning extracted from the text of a document (denoting the information content), is designated by concepts and their interrelations (which are not necessarily found in the text) by 'ad hoc' symbols (descriptors, role indicators, facet indicators, links, codes etc.), in a formalised way. Because these set of symbols are external to the object-language in which the documents under analysis are written and intended to facilitate the manipulation of these documents in various ways and for various purposes, they can be called 'metalanguages' (20). They are not only used in analysing 'what the information content of documents are' but also in constructing statements of the names of subjects of documents and also for verifying the correctness and completeness of these statements. In the formalised statements of these metalanguages 'some of the 'logical semantic' relations, specifically those of implication are specified, but not in the surface-structure of the object-language, i.e. natural language' (21). Moreover, these documentation languages, though developed under independent conditions for the analysis of documents
of many different kinds — not only scientific texts, but also sociological records, folklore, scripture etc. — have striking similarities (20). An account of these metalanguages has been given by Gardin (22).

B6 FEATURES OF DOCUMENT AND LINGUISTIC ANALYSIS

Document analysis in the context of a particular science, obviously implies some knowledge about the universe of discourse. It is left to be decided whether language analysis in a more general, less specialized, field of speech or writing, does not also imply some knowledge about that field.

B6.1 Documentation languages do not feel it an obligation to consider the natural language sentence as the proper unit of analysis, as a large number of linguists would have it. For one thing, the definitions of what a sentence is, in any given natural language are not so stable as to provide with a firm analytical framework for the kind of semantic analysis required in documentation languages. Conversely, infra-sentential units often provide a more convenient basis on which to conduct the derivation of documentation language units, which are later chained to one another through procedures that again overstep the boundaries of natural language sentence.
B6.2 In language analysis, the priority is given to syntactical analysis, which is considered as the necessary step by which to initiate the description or generation of natural language text and to the understanding of language. The standard unit of analysis is the sentence, as it is more or less intuitively defined in traditional grammar; any larger unit is felt to exceed the scope of syntactical analysis. The basic language units for syntactical analysis are grammatical categories (nouns, verbs, adjectives etc.) and grammatical functions (subject, object etc.), of former times. But they are being further refined as shown in the paragraphs below.

In accordance with the priority given to syntactical analysis, semantic analysis is confined to a subsidiary position as illustrated by Katz and Fodor (23), which still seems to form the basis of the idea that the properties of 'surface structure' (in grammatical terms) play a primary role in the determination of meaning (24).

B6.3 But linguists themselves have felt the need to broaden the basis of language analysis. Grimes has (25) pointed out that it is unwise to go on ignoring the
findings of other disciplines, also concerned with the analysis of text, such as rhetoric, criticism etc., and information science itself, especially since some of the better defined procedures recently set forth in these areas occasionally unveil facts that linguists would be ill-advised to neglect. A number of linguists have proposed the broader concept of 'paraphrase' as a basis upon which to decide the proper level where sentences should be assigned the same 'deep structure' (26). Also a notion of syntactic analysis as a less discriminating form of semantic analysis (27), and the notion that "syntactic structures are derived from semantic graphs" (28) point to the fact that syntax is no more considered to be the first step of language analysis. The derivation of syntactic structures on 'semantic graphs' is exactly the strategy adopted in documentation languages in connection with document analysis.

An investigation of 'prelexical' structures, equivalent to representations in documentation languages, prior to the development of grammatical transformations has also been advocated by linguists (29). Again the adoption of "semantic categories (that) can guide the
interpretation of sentences, independently of and in parallel with perceptual processing of the syntactic structure" (30) has also been advocated. Though some linguists have favoured units smaller than sentence (31), others on the contrary, have demonstrated that larger units are necessary to provide a proper understanding of 'coordination', 'pronominalization', 'subordination' etc. (32-34). More generally there have been many signs of a revival of interest for the analysis of 'discourse' over and above the analysis of sentences, as done in document analysis.

It has also been observed by linguists (35-37) that the 'proposition' defined in more or less logical terms, provides a more convenient analytical framework than the grammatically defined sentence. In the presentation of the basic structures that are supposed to underlie the formation of speech, linguists have tended to use logical concepts often very close to those which underlie document analysis. From the modest use of 'markers' (23) came the view of semantic representations as 'trees' from which has come the broader concept of relational or combinational information.
about the world' that have to be taken into account for a proper understanding of language. The function of such semantic networks is to provide for the enunciation and application of rules of 'well-formedness' that were presented as a necessary extension of syntactical theory. The 'presuppositions' have been given more and more importance later (38). Though the notion of 'presupposition' needs "some sort of conceptual straightening up" (39) it can still be observed that it is being used in the same way for theoretical purposes just as the notion of semantic or 'paradigmatic' organisation has been used for applied purposes in document analysis.

Another parallelism with the procedures of document analysis could be seen if the kind of 'categories' used in the formulation of presuppositions is considered. In brief, the presuppositions associated with a given word 'X' consists in a statement of some relational/combinational properties of the given word 'X' with respect to other words 'w_1, w_2, ..., w_n' or classes of other words 'W_1, W_2, ..., W_n'; that is, the role of 'W_i' with respect to the given word 'X'. The basic idea is
that 'role structures' can be attached to particular words (40) in order to express the range of their possible relations with other words, from a syntactico-semantic viewpoint. The 'case frames' of Fillmore (41), the 'basic relation structures' of Bever (42), the 'mental templates' of Wilks (43), all have much the same function. The nature of 'roles' (or 'cases', 'mental templates' etc.) naturally differ according to the particular word under consideration. However, identical role structures have been assigned to classes of words. By carrying the abstraction process further a set of 'universal roles' can be suggested. According to a well known linguist Bangendoen: "eventually it will be necessary for linguistics theory to come up with a universal inventory of roles" (44).

Such universal roles, then would have to be of a rather abstract nature (such as Agent, Instrument, Object, Process, Property etc.). They would thus tend to converge with logical operators of the kind used in 'syntactical' analysis, even though the initial purpose was to account for 'semantic' structures. Specific field-bound roles have been used in document analysis.
for a long time. For example, the various faceted classification schemes for different special fields. The process of abstraction leading to the more general 'role operators' or 'categories' too is a well known path used in theory of document analysis to relate semantics and syntax. For example, the Fundamental Categories Personality, Matter, Energy, Space and Time of Ranganathan's Colon Classification and his search for 'absolute syntax' of facets (45, 46). Linguists also have come up with similar proposals. For instance: Grimes' categories: States, Processes, Actions etc., (47) and Bach's (48) new category 'contentive' to cover nouns, verbs and adjectives.

Further, meaning representation systems such as Sager's automatic conversion of texts to a structured information base (49), Schank's computerised paraphrase and inference system (50, 51), Wilks' intelligent analyser and understander of English (52), question answering systems of Grishman and Hirschman (53) and of Lehner (54) and so on, all depend on some sort of underlying semantic categorisation and deep syntactic relations. Moreover, the suggestion that "deep structures
should be stated in terms of role relationships rather than syntactic relationships" (55) and the 'correlation of roles with prepositions' (56) suggest that the broad conceptual framework inferred from the practice of document analysis bears a definite relation to linguistic theory and language analysis.

B6.9 Formally stated, the common thread between the theories of document analysis and that of language analysis is, the role assigned to n-place predicates (the term 'predicate' in this context designates any relation holding between two or more entities or any property of an entity (57)) to express semantic as well as syntactic representations with:

1. the need to categorise the symbols of the vocabulary (words, descriptors) in such a way that formation rules equivalent to the phrase-structure rules of grammar can be stated adequately with no regard to the grammatical status customarily assigned to the words concerned; and

2. the need to account for the derivation of propositions from one another, as a necessary component, in the understanding of language behaviour.
In documentation languages, there is the concept of 'summary' or 'aboutness' which may not be considered in other languages. Documentation languages are used to formulate statements as answers to questions of the type "What the information content of particular documents are about?". These statements stand as names of the topics or subjects of the documents. In short they are languages for naming subjects for the special purpose of storage and retrieval. As already mentioned in paragraph B4.6, they may be classification schemes, indexing languages, thesaurus based indexing and retrieval languages and so on. As mentioned in paragraph B4.5 and section B5, they are also regarded as metalanguages for information organisation and retrieval.

B7.1 The minimum components of any such metalanguage (58) are given below:

1. The 'lexicon', a list of content terms, either extracted as they occur in a given natural language (keywords) or redefined for the purpose of the analysis (descriptors).
If the list of content terms carry no relational information of any kind, then the metalanguage is said to be 'unorganized' (uniterm lists).

2 A set of relational data provided 'a priori' with the lexicon, irrespective of the way in which they are expressed (cross references, hierarchy, tree, factoring etc.), and whether they be regarded as semantic data as in taxonomies or as syntactic templates as in some faceted classification schemes. This forms the paradigmatic structure of the metalanguage.

3 To take into account the immensely diverse relational data observed in documents, a set of rules of syntax, constituting the syntagmatic structure of the metalanguage, which is contrasted to the paradigmatic structure, not in essence but in use.

Subject indexing languages such as, Preserved Context Indexing System (PRECIS), Chain Indexing System and Postulate based Permuted Subject Indexing System (POPSI) all have the components mentioned above.