LANGUAGE AND INDEXING LANGUAGE: NALIMOV AND GARDIN REVISED

F J DEVADASON
Documentation Research Training Centre
Indian Statistical Institute
Bangalore-560059

M R KUMBHAR
Librarian
Manipur University Library
Canchipur
Imphal-795003

A study of the salient features of ordinary language is necessary for the design of information storage and retrieval systems in general and indexing systems in particular. This paper, based on the works of Nalimov, and Gardin, attempts to discuss some of these salient features. Only the relevant characteristics of ordinary language are mentioned, which are reflected in indexing languages. The important features of indexing language and its meta-language character are also mentioned. Common analogical features of both subject analysis and linguistic analysis are presented along with a gauge. Indicators that the Postulate based Permuted Subject Indexing (POPSI) language, together with its vocabulary control tool Classarius, has also the important components of an indexing language.

INTRODUCTION

Language is a means of communication and thinking — for dissemination of information and to an extent the creation of it. In indexing, language plays a very special role in information organisation and documentation. 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 [1]. A study of the salient features of language is helpful in the design of information storage and retrieval systems in general and indexing systems in particular [2]. This paper based on the works of Nalimov [3] and Bardin [4] presents some of the salient features of language and indexing language.

LANGUAGE

Language is system of symbols serving the means of human communication, thinking and expression. 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” [5]. 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-instinctive method of communicating ideas, emotions and desires by means of a system of voluntarily produced symbols. Only the communicative aspect of language plays an important part in information systems, "the essence of which is storage of information for use and manipulation by an individual” [6].

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 languages used for 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 mechanism — 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 information transfer. But these do not become ‘ordinary languages’. The essential functional characteristic of ordinary language (referred to as just ‘language’ hereafter) is its role in information reduction or synthesis, abstraction, generalisation; in elaboration or expansion; and particularisation or concretisation — in short, thinking.
STRUCTURE OF LANGUAGE

Signs/Symbols

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

Second, in a language each sign/symbol has a signification common to a number of interpreters.

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.

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 uni-situational vehicle [7]. 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” [3]. 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 make a multitude of particular existences.

Fifth, the different signs in a language should fit into a semantic hierarchic structure. “I see two things: a chair and a furniture” is not an acceptable sentence. 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. 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 [7].

Symbol system

The structure of language starts with

1. subelementary linguistic symbols: the letters or alphabets, morphemes (a morpheme is a meaningful part of 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 full stops); 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 symbols from simpler ones which is outwardly reflected in the hierarchical structure of language [3]. 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 strives 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 logical hierarchy of statements exists in language, but it is so concealed that in practice, it cannot be directly observed” [3, p.31]. The narration of events in chronological sequence in text is an example of this kind of ‘hierarchy’.

Grammar

In order to make operative the rules for constructing complex logical structures using the simpler symbols, the symbols 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 structural elements of language; they are clearly seen in the majority of the symbol systems perceived as languages [3, p.28].
Characteristics of Language

Interpretability

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, idoms etc., linked to culture and historical experiences present in language.

Polymorphism

Consider the following three phrases:

1. This a good mango,
2. This 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 mango 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 mango. Both in everyday language and in many other languages every special 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.

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. Generally the metalanguage is formalised to a greater extent than the object language. A metalanguage should be so rich that every thing 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 guage [3, p.39]."

LANGUAGE OF SCIENCE

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 sciences: 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 object or phenomena. For example the term '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' has been studied by Cho Yuan-Ren [8]. 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, every body knows the common meaning of the word ‘isolate’. In Colon Classification its meaning is completely changed (that which in isolation is not fit to form the name of subject and has to have a Basic Subject component along with it to represent a subject). In the language 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 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. "In the realm of professional activities too there evolves a restricted and accommodated language, a so called professional language. A professional language is developed and used by people for some professional purpose, that is, performing professional tasks in a specific field and communicating about them. A professional language consists of vocabulary (terminology and concepts) and types of communicative acts (including typical intentions). A professional language is not only used for talking about an activity field (a universe of discourse); it is also a part constituting it" [9]. Of course Library and Information Processing profession has its own professional language and dialects too.

BIBLIOGRAPHIC INFORMATION SYSTEM

The functions of a bibliographic Information System are: collecting, organising, 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), categorisation and systematisation 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.

Of the above mentioned three stages, the second one namely, information analysis or document analysis or subject analysis or content analysis decides the efficiency of the system. It deals with systematising or organizing the information/documents accompanied by classification and indexing of the information content of the 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. A model Bibliographic Information System is given in figure-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 specified language having grammar, called information/documentation/indexing language. After the completion of this analysis, there is a branching which allows for more than one method of 'organising' 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.

The user requests information. She/He interacts with the index and/or storage file. To do so, she/he must convert his/her 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/indexing language. When the user is satisfied with the documents which have been located in response to this
request, she/he reads them to obtain new facts and ideas. Presumably, she/he will now become a generator of new documents which will in turn find their way into the system.

The nucleus of the Bibliographic Information System is the Information/Indexing language. According to it the documents are analysed. The creation and organisation of the 'index file' and the setting up of the retrieval procedures are all based on the analysis of the documents according to the prescriptions of the information/indexing language.

Natural language is a tool for thinking and a means of communication. In such a language there are synonymous correspondences 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 information languages/indexing languages/documentation languages are used. These artificial languages are formalised auxiliary languages, created by information professionals for use in bibliographic information storage and retrieval systems. Because of their formalised structure they are regarded as metalanguages for information organisation and retrieval. They are used in:

1. analysing the information content of documents;
2. formulating names for the information content or names for the subject of the documents;
3. verifying the correctness and completeness of the statements of the names of subjects of documents; and
4. constructing other information handling tools.

INFORMATION/INDEXING LANGUAGE AS METALANGUAGE

In information/indexing languages, the meaning extracted from the text of a document (denoting the information content), is designated by concepts and their inter-relation (which are not necessarily found in the text) by 'adhoc' symbols (descriptors, role indicators, facet indicators, links 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’ [4]. 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, that is, natural language” [10]. Moreover, these information/indexing languages, though developed under independent conditions for the analysis of documents of many different kinds – not only scientific texts, but also sociological records, scripture, folklore etc., – have striking similarities [4]. An account of these metalanguages has been given by Gardin [11].

FEATURES OF INFORMATION/SUBJECT AND LINGUISTIC ANALYSIS

Information/document analysis in the context of a particular science, obviously implies some knowledge ‘about’ the universe of discourse. There is a need to have a knowledge to write summarising statements about the content of the document/information being analysed. It is left to be decided whether language analysis in a more general, less specialised field of speech or writing, does not also imply some knowledge about that field.

Information/indexing 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. 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 information/indexing languages. Conversely, infra-sentential units often provide a more convenient basis on which to conduct the derivation of information/indexing language units, which are later chained to one another through procedures that again overstep the boundaries of natural language sentence.

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. In accordance with the priority given to syntactical analysis, semantic analysis is confined to a subsidiary position as illustrated by Katz and Fodor [12], which still seems to form the basis of the idea that the properties of ‘surface structures’ (in general terms) play a primary role in the determination of meaning [13].

But linguists themselves have felt the need to broaden the basis of language analysis. Grimes [14] has 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 setforth 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 structures’ [15]. Also a notion of syntactic analysis as a less discriminating form of semantic analysis [16], and the notion that “syntactic structures are derived from semantic graphs” [17] 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 information/indexing languages in connection with subject/document analysis.

An investigation of ‘prelexical’ structures, equivalent to representations in information/indexing languages, prior to the development of grammatical transformations has also been advocated by linguists [18]. 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” [19] has also been advocated. Though some linguists have favoured units smaller than sentence [16, p. 101-102], others on the contrary, have demonstrated that larger units are necessary to provide a proper
understanding of ‘coordination’, ‘pronominalisation’, ‘subordination’ etc. [20-22]. 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/subject analysis.

It has also been observed by linguists [20, p. 32; 199p 297; 14, p.141] 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/subject analysis. From the modest use of ‘markers’ [12] came the view of semantic representations as ‘trees’ from which has come the broader concept of relation of combination 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 [23]. Though the notion of ‘presupposition’ needs “some sort of conceptual straightening up” [24] 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 information/subject analysis.

Another parallelism with the procedures of document/subject 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 consists in a statement of some relational/combina

EDVADASON

suggested. According to a well known linguist Langendoen: “eventually it will be necessary for linguistics theory to come up with a universal inventory of roles” [20, p. 79]. Vermier [26] too has come up with such conclusions.

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/subject 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 categories of Kaiser’s systematic indexing, the Fundamental Categories Personality, Matter, Energy, Space and Time of Ranganathan’s Colon Classification and his search for ‘absolute syntax’ of facets [27, 28]. Moreover the suggestion that “deep structures should be stated in terms of role relationships rather than synthetic relationships” [19, p. 279; 32, P.62] suggested that broad conceptual framework inferred from the practice of document/subject analysis should bear some (at least analogical) relation to linguistic theory and language analysis. Linguists also have come up with similar proposals. For instance, Grimes’ categories: States, Processes, Actions etc., [14, p. 167-8] and Bach’s new category ‘Contentive’ to cover nouns, verbs and adjectives [29].

Further, meaning representation systems such as Sager’s automatic conversion of texts to a structured information base [30], Schank’s computerised paraphrase and inference system [31, 32], Wilk’s intelligent analyser and understander of English [33], question answering systems of Grishman and Hirschman [34], and of Lehnert [35] and so on, all depend on some sort of underlying semantic categorisation and deep syntactic relations.

Formally stated, the common thread between the theories of document/subject 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 any
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.

**POSTULATE BASED PERMUTED SUBJECT INDEXING LANGUAGE**

The Postulate-based Permuted Subject Indexing language of Bhattacharyya [36-42] has all the essential features mentioned above. It is based on

1. a set of postulated Elementary Categories of the elements fit to form components of names of subjects;
2. a set of syntax rules with reference to the Elementary Categories for formulating admissible names of subjects; and
3. a faceted hierarchic scheme of terms for vocabulary control called Classaurus.

Each of the terms occurring in POPSI language statements have their appropriate broader terms prefixed to them. Hence the POPSI language has been shown as a source language for producing and organising associated indexes [43]. It has also been shown as a metalanguage for computer aided generation of information retrieval thesaurus [44], as well as Classaurus [45, 46]. POPSI language is amenable for computerisation of index generation, especially in producing different types of indexes including chain index and PRECIS-format index entries [47, 48]. It is also possible to use POPSI language in computer based online information retrieval. In such an online information retrieval system the searcher need not know any query language and the system would have built-in vocabulary control mechanism [48, 49].

**CONCLUSIONS**

A study of the salient features of language and language analysis is necessary for the design of information retrieval systems in general and indexing system in particular [2]. It would help in incorporating the required features though not exactly but atleast analogically in indexing systems. POPSI has incorporated all the essential
features necessary for making it an universal indexing system. Its resilience and amenability to computerisation has paved the way for research in this direction, the ultimate aim of which is to develop an online information retrieval system which would be user friendly.

BIBLIOGRAPHICAL REFERENCES


21. Lakoff, R. If’s, and’s and But’s about conjunction. In above cited ref. 14; p. 55-70.


