GENERATION OF THESAURUS IN DIFFERENT LANGUAGES
A COMPUTER BASED SYSTEM

F J DEVADASON
Documentation Research and Training Centre
Indian Statistical Institute
31 Church Street
Bangalore - 560001
INDIA

The development of the theory of library classification and of subject indexing, for the organisation, storage and retrieval of subjects embodied in documents has a striking parallelism to the search for 'universal forms' and deep structure'in language and linguistic studies. The significant contributions of the theories of classification and subject indexing are the subject analysis techniques of Ranganathan and Bhattacharyya's POPSI. A computer based system, for generating an information retrieval thesaurus, from modulated subject-propositions, formulated according to the subject analysis techniques, enriched with certain codes for relating the terms in the subject-propositions has been developed. The system generates hierarchic, associative, coordinate and synonymous relationships between terms and presents them as an alphabetical thesaurus. Also, once a thesaurus is generated in one language it is possible to produce the same thesaurus in different languages by just forming a table of equivalent terms in the required language.

Information Retrieval Thesaurus

An information retrieval thesaurus could be defined as "a controlled dynamic vocabulary of semantically related terms offering comprehensive coverage of a domain of knowledge". Its main use is in the subject characterization of documents and queries in information storage and retrieval systems based on concept coordination [26].

The application of computers for updating, testing, editing and printing thesaurus [25] has gained much importance due to the use of thesaurus as a vocabulary control device in bibliographic information storage and retrieval systems, at the input stage for controlled indexing and at the retrieval stage for expanding the 'search query' to increase recall - both in batch and on-line modes of processing [18, 22, 62, 64].

Automatic Generation of Thesaurus

Several experiments in automatic generation of thesaurus have been carried out in which relationships between terms have been determined by taking into account the number of documents in which the respective terms occur jointly [27, 65]. Various clustering techniques have been investigated out of a range of similarity criteria. The role played by similarity criteria in obtaining the environment of each term and the use of this environment for retrieval have been explored [57]. Computational procedures for generating thesaurus include keyword statistics, calculation of Tanimoto coefficient [61], matrix inversion, formation of similarity matrix, automatic cluster analysis using minimal tree procedure and compilation of groups and main groups of descriptors [1, 29, 54, 56, 58, 65, 66].

But, "the difficulty, however, is that text-scanning is more effective in syntactic and morphological analysis where there is sufficient repetition to justify the belief that a particular fact is significant" [31, 32]. Further, all these techniques use a lot of computer time and are capable of producing a list of selected and grouped keywords. But it has also been observed that, although a large variety of clusters and associated query expansions have been obtained, no significant improvements in the document retrieval performance have been achieved [39]. In other words, an information retrieval thesaurus is something more than a list of grouped and ranked keywords [44].

Basic Aspects of Thesaurus

There are two basic aspects of thesaurus construction. They are:

1 selection of keywords/descriptors of the subject for which the thesaurus is constructed; and

2 establishment of interrelationships among these selected keywords as to whether the terms form a broader or narrower or related or synonymous or 'use' relation.

Using computers alone for both the above mentioned aspects of thesaurus construction is not practicable for, using computers for selecting the keywords from free language text is not an economic approach, and it is not feasible to make a computer automatically distinguish the relationship between terms as to broader, narrower, synonymous etc.

Metalanguage for Information Organisation

It was realised that the failure of experiments with automatic abstracting, indexing etc., "should be sought above all, in an insufficient knowledge of the structure of the text from the standpoint of relationships between an apparent formal linguistic representation on one hand and on the other hand, the informational content involved in the text ... As the result of such investigations we can arrive, among others, at various descriptive formulas of the structure of scientific texts"[28]. One way of arriving at structures that reflect specific textual content is to make use of the restrictions in language usage which are characteristic of the texts in a particular subject matter that is, to exploit the fact that on a particular topic, only certain words in certain combinations actually appear [55]. In other words, it was realised that what is required is a special purpose artificial language, to cater to the needs of information storage, processing and retrieval. The Automatic Language Processing Advisory Committee (1966) realised and reported that "A deeper knowledge of language could help ... to enable us to engineer artificial languages for special purposes... and to use machines as aids in translation and in information retrieval"[40].

Subsequently, suggestions for a metatheory of linguistics and information science, with a metalanguage having all the properties of a classification schema have been proposed. The term 'metalanguage' specifies a 'public' metalanguage, such as a document classification system, as distinguished from the 'object language' represented by the documents. The written record of a document classification schema is not really parallel to the surface structure of the object language - the natural language sentences of a document. A classification schema is intended to classify, and, therefore the language of the schema is mainly classificatory. In other words, the metalanguage does not explicitly include all relevant terms in the object language, but the object language does include all terms in the metalanguage. Moreover, superset-subset (class inclusion) relations are usually explicitly given by the structure of the classification [38]. Thus some of the'logical semantic' relations, specifically those of implication [34] are specified in the so-called 'surface structure' of the metalanguage, but not in the surface structure of the object language [38].

Universal Forms and Subject Representation

Parallel to the search for universal linguistic forms such as that expounded by Chomsky, Fodor and others (the discovery that certain features of given languages can be reduced to universal properties of language and explained in terms of deeper aspects of linguistic form, [11, 12, 37]; and that such deep structure of sentences determine the semantic content while their surface structures determine the phonetic interpretation), steps towards the formulation of generic framework for structuring the representation of the name of a subject for the development of classification schemes and subject indexing languages were investigated [21, 43, 45-48]. Such universals are being arrived at and used in various other areas dealing with information and information processing. For instance, in the area of data modelling, now the basic problem is to identify the world as a domain of objects with properties and relations [10].

Such categorisation of objects of study is not new to the library profession. As early as 1930s, the use of categorisation of component ideas forming the name of a subject into Personality/core object of study, Matter/property/method, Energy/action, Space/place and Time, and defining an order of these categories to form a 'logical, classificatory language' resulting in 'faceted' library classification schemes was known in India [45, 47].

It is interesting to note that it has been realised now that the above mentioned Ranganathan's categories Personality, Matter and Energy, are "general categories building the system's structure as a spatiotemporal neighbourhood relationship " useful in deriving meta informational, for a process of automatic analysis too [13, 14].

The order of the component ideas denoting the different categories in the name of a subject as prescribed is context-dependent order. More specifically it is context-specifying order. Every component category sets the context for the next and following ones. Also in this classificatory language, every category should explicitly have the corresponding superordinate component ideas preceding it. The reason for fixing the superordinates before the component elements concerned is to render the component elements denote precisely the ideas they represent.

Further, it has been conjectured that [46, 52] the syntax (order) of representation of the component elements in the name of a subject as prescribed by the principles for sequence - facet sequence [49] - is more or less parallel to the Absolute Syntax - ie., the sequence in which the component ideas of subjects falling in a subject-field arrange themselves in the minds of a majority of normal intellectuals. If the syntax of the representation of the component ideas of subjects is made to conform to, or parallel to the Absolute Syntax, then the pattern of linking of the component ideas - ie., the resulting knowledge structure is likely to be [42]

1 More helpful in organising subjects in a logical sequence for efficient storage and retrieval;
2 Free from the aberrations due to variations in linguistic syntax from the use of the verbal plane in naming subjects; and
3 Helpful in probing deeper into the pattern of human thinking and modes of combination of ideas.

Subject Indexing and Thesaurus

Due to the development of techniques for structuring of subjects and for classification of subjects, several experiments were conducted at the Documentation Research and Training Centre to use them for thesaurus construction. To begin with, a faceted library classification scheme for a specific subject field was used in the computer generation of thesaurus [59] in which it was possible to incorporate the hierarchic relationships of terms. But it was not possible to incorporate the generation of non-hierarchic associative relationship of terms.

Terms that have associative relationship to each other have to be established only by consensus of experts in the field concerned. But the validity of the assumption that, knowledge based on the consensus of experts in a field is different from the knowledge expressed in the literature of the field has been challenged, as the two lists of keywords, one given by experts and the other formed by analysis of published literature were not significantly different [33]. In other words, terms that are related to each other associatively could be easily ascertained by an analysis of the statement of the name of the subject of a document or of a reader's query. For instance, whether "x-ray treatment" is associatively related to "cancer", or not, could be established if there exists a document on "x-ray treatment of cancer". In other words, a published document on "x-ray treatment of cancer" brings into associative relationship both "x-ray treatment" and "cancer". Also it is unimportant which terms co-occur frequently in the names of subjects for, any term that is used once in the statement of the name of a subject is enough to be admitted into the thesaurus for that subject and is related with other terms in that name of the subject in some particular way. In order to incorporate associatively related terms in thesauri, experiments were conducted [35, 53] using subject representations formulated for the purpose of developing classification schedules, which were arrived at by Ranganathan's facet analysis [21, 49] for thesaurus construction. With certain limitations it was possible to generate broader, narrower and associative relationships but not coordinate relationships. Further, it was realised that [2, 17] selection of candidate terms and ascertaining of multiple linkage of relationships among terms can be done in several ways such as by

1 the analysis of user's query specifications;
2 the analysis of summarised statements of the subjects of documents; and
3 the analysis of sentences in the text of dictionary, glossary, encyclopaedia and even text books and treatises.

Artificial Language for Thesaurus

Further research into the fundamentals of subject indexing languages resulting in the development of a general theory of subject indexing languages [4] and the development of the Postulate-based Permuted Subject Indexing (POPSI) language [3, 8] has provided a basis for a more efficient and flexible system for thesaurus construction.

According to the general theory of subject indexing languages; information is the message conveyed or intended to be conveyed by a systematised body of ideas, or its accepted or acceptable substitutes. Information in general, is of two types: discursive information and non-discursive information or unit facts. Non-discursive information or unit facts may be either qualitative or quantitative [5, 16]. The name of a subject is essentially a piece of non-discursive information and it is conveyed by an indicative formulation that summarises in its message, 'what a particular body of information is about'. "The language for indicating what a body of information is about, need not necessarily be in terms of sentences of the natural language. It can be an artificial language of indicative formulation used to indicate what a body of information is about" [6].

The essential ingredients of a language - natural or artificial - are the elementary constituents; and rules for the formulation of admissible expressions using the elementary constituents. A Subject Indexing Language consists of elementary constituents and rules for the formulation of admissible subject-propositions. It is used to summarise in indicative formulations what the contents of a source of information are about. The purpose of these summarising indicative formulations is to create groups of sources of information to facilitate expeditious retrieval of information about them by providing necessary and sufficient access points.

The component ideas in the name of a subject can be deemed to fall in any one of the elementary categories: Discipline, Entity, Action and Property. The term 'manifestation' is used to denote an idea or a term denoting an idea, falling in any one of the elementary categories. Apart from the elementary categories there are Modifiers to the elementary categories. A modifier refers to an idea or a term denoting an idea, used or intended to be used to qualify the manifestation without disturbing the conceptual wholeness of the latter. A modifier can modify a manifestation of any one of the elementary categories, as well as a combination of two or more manifestations of two or more elementary categories. Modifiers can be common modifiers like time, place etc. or special modifiers which can be entity based or action based or property based. Apart from the elementary categories and modifiers there is a Base and Core. Due to the fact that recent research work is generally project-oriented, mission-oriented and inter-disciplinary and not generally discipline-oriented, there may be a need to bring together all or major portion of information pertaining to a manifestation or manifestations of a particular elementary category. This manifestation or elementary category is the Base. Similarly, need may arise to bring together within a recognised Base, all or major portion of information pertaining to manifestations of one or more elementary categories, the category or categories concerned are the Core of the concerned Base. Also the elementary categories may admit of Species (genus-species) and Parts (Whole -Part).

The elementary constituents of a specific Subject Indexing Language - POPSI [3, 7, 8] are given below:

2	Relation		
2.1	General		
2.2	Bias		
2.3	Comparison		
2.4	Similarity		
2.5	Difference		
2.6	Application		
2.7	Influence		
	Common Modifier		
3	Time Modifier		
4	Environment Modifier		
5	Place Modifier		
	Elementary Category		
6	Entity	(E)	,Part
7	Discipline	(D)	.Species/Type -Special Modifier
.1	Action	(A)	A and P can go with another A and P also
.2	Property	(P)	
8	Core	(C)	Features analogous to D, E, A and P.
9	Base	(B)	

The rules of syntax of POPSI prescribed for the subject-propositions is D followed by E (both modified or unmodified) appropriately interpolated or extrapolated wherever warranted, by A and/or P (both modified or un-modified). A manifestation of Action (A) follows immediately the manifestation in relation to which it is an A. A manifestation of Property (P) follows immediately the manifestations in relation to which it is a P. A Species (type)/Part follows immediately the manifestation in relation to which it is a Species/ Part. A Modifier follows immediately the manifestation in relation to which it is a modifier. Generally a modifier gives rise to a species. Also if necessary auxiliary words within brackets could be inserted in between terms if found necessary. These form the basis of the POPSI language.

While examining whether a classification scheme could form a 'metalanguage' of a metatheory of linguistics and information science, it has been observed that "all relational information necessary for the explication of an object language" are not present in classification schema, especially role notions and presuppositions [38]. Such 'relational modifiers' or 'role indicators', [15, 20, 63] that describe the role of the concept in context, representing basic 'role notions' such as the cause of the event, the effect of the event etc., similar to that of the case relations - nominative, accusative, instrumental [19] etc. - if incorporated in the subject-propositions, formulated according to the 'subject analysis' techniques mentioned above [3-8, 45-52], then it could form a 'metalanguage' for thesaurus, from which thesaurus could be generated automatically.

Input Subject-propositions for Thesaurus

The preparation of input to the thesaurus construction system starts with writing out sentences such as, "this book is about ... , this report is about ... , this paper is about ..., this query is about ... " 23, 36. "To tell what is the subject or topic of a play, a picture, a story, a lecture, a book etc., forms part of the individuals mastery of a natural language ... They are the starting point of most requesters when approaching a bibliographic information retrieval system or in a dialogue with a librarian or documentalist" [60]. To aid in such an indicative formulation that summarises in its message what a particular body of information is about, the title of the document or the raw specification of the readers' query or even sentence or sentences in the text of dictionary, glossary, abstract and even text-books is taken as the starting point. Each of the specific subjects dealt within the document or specified in the reader's query or text statements are determined and expressed in natural language.

Let one of the names of subjects be expressed as "Re-tanning of chrome tanned leather using chestnut". Each of the component ideas such as the name of the discipline (base) the core object of study (entity) etc., that are implied in the expressed statement of the subject are explicitly stated to form an 'expressive title' [48, 50, 51] as follows: "In Leather Technology, re-tanning of chrome tanned leather by vegetable tanning using chestnut".

The 'expressive title' is then analysed to identify the 'elementary categories' and 'modifiers' and the component terms are written down removing irrelevant auxiliaries, as a formalised representation, following the principles of sequence of components 9, 49. The analysed and formalised subject-proposition is given below:

(Discipline) Leather Technology,
(Core Entity) Chrome Tanned Leather,
(Action on Entity) Re-tanning, [by]
(Action based Modifier) Vegetable Tanning, [Using] (Entity based Modifier) Chestnut.

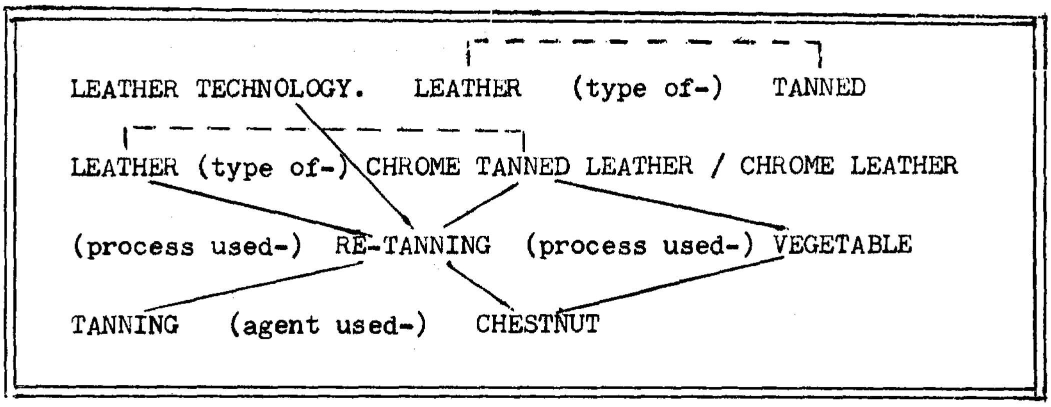
The subject-proposition is then modulated by augmenting it by interpolating and extrapolating as the case may be, by the successive superordinates of each elementary category by finding out 'of which it is a species (type) or part'. The synonymous terms if any are attached to the corresponding standard terms. The modulated subject-proposition is given below:

Leather Technology. Leather, Tanned Leather, Chrome-tanned Leather/ Chrome Leather. Re-tanning (by) Vegetable Tanning (Using) Chestnut.

The auxiliary words (even if relevant are removed from the subject-proposition and phrases enclosed within brackets indicating 'role notions' or 'role indicators' are inserted between the kernal terms. The resulting subject-proposition is given below:

Leather Technology. Leather (type of-) Tanned Leather (type of-) Chrome Tanned Leather/Chrome Leather. (process used-) Re-tanning (process used-) Vegetable Tanning (agent used-) chestnut.

The subject-proposition is further analysed to determine which terms are associatively related to each other specifically. For instance, in the above subject-proposition 'chestnut' is related to 'Vegetable tanning' and also to 'Re-tanning', as an agent used in both the processes. 'Chrome tanned leather' is related to 'Re-tanning' as it admits of being re-tanned, and also to 'Vegetable tanning' as it admits of being vegetable (re) tanned. After this analysis, the subject-proposition is formulated as a relation map showing the 'links'. The relation map for the above subject-proposition is given in the figure below:


LEATHER TECHNOLOGY. LEATHER (type of-) TANNED
LEATHER (type of-) CHROME TANNED LEATHER / CHROME LEATHER
(process used-) RE-TANNING (process used-) VEGETABLE
TANNING (agent used-) CHESTNUT

In the relation map given above, the dotted lines indicate NT/BT relationship, continuous lines indicate RT relationship and slash indicates synonym/use relationship.

The relationship between pairs of terms NT or RT as indicated by dotted lines and continuous lines respectively as shown in the example, are replaced by appropriate codes to form the input to the thesaurus generation system.

The codes used in the subject-propositions for generating entries for a thesaurus are of the following types:

1 those that indicate which terms are to be related (codes for relating terms) and whether the relation is NT or RT or SYN; and
2 those that denote the role indicators.

The codes for relating terms are of the following three types:

1 those that indicate NT relation;
2 those that indicate RT relation; and
3 that which indicates Synonymous relation.

The codes for generating NT relation and the associated computer manipulation are: '$2' -- Generate a NT relation with the immediately succeeding term using the role indicator code of the term being manipulated and generate a reverse BT relation changing the position of '-' in the role indicator code (genus - species relation); and '$3' -- Generate NT relation with the immediately succeeding term and generate a reverse BT relation. No role indicator code is used (whole - part relation).

The codes for generating RT relation and the associated computer manipulation are: '$1' -- Generate a RT relation with the immediately succeeding term using the role indicator code and generate a reverse RT relation changing the position of '-' in the role indicator code; and '$5, $6, $7, $8, and $0' -- Generate a RT relation with the immediately preceding term with the same '$ code' taking the role indicator code of the term being manipulated and generate a reverse RT relation changing the position of '-' in the role indicator code.

The code for generating Synonymous relation and the associated computer manipulation is: '/' -- Generate a Synonymous relation with the immediately preceding term and generate a reverse 'Use' relation.

It is to be noted that the role indicators are used specifically for further categorisation of RTs, as they are expected to be numerous. But representation of genus-species relations could also be categorised to achieve better display format and for proper generation of coordinate RTs out of NTs to a particular term. The following is an extract of role indicators used in our experimental thesaurus on Leather Technology:

01	- Source of;
07	- Property of;
08	- Process used;
12	- Agent used;
13	- Device used;
16	- Type of;
19	- Constituent of.

The subject-proposition drawn as a relation map is augmented with the codes described above to reflect the different NT and RT links as given below:

```
$0 LEATHER TECHNOLOGY $4 LEATHER
$2 (16-) $5 TANNED LEATHER $2 (16-)
$6 $7 CHROME TANNED LEATHER / CHROME
LEATHER $0 (08-) $5 (08-) $6 (08-)
RE-TANNING $7 (08-) $6 (08-) VEGE-
TABLE TANNING $1 (12-) $5 (12-)
CHESTNUT.
```

Computer Coding of Subject-propositions

An assorted number of subject-propositions from a specific subject field, augmented with codes for relating terms and codes for role indicators are read by a program 'CODER'. Each of the unique terms in the subject-propositions is internally serial numbered uniquely and the respective terms in the subject-propositions are replaced by their serial numbers. As and when a term is encountered in a subject-proposition, it is matched with existing terms and its serial number is picked if the term is available, if not the term is entered as the last entry with appropriate serial number and the given serial number is replaced in the subject-proposition. The term dictionary thus built, and the translated subject-propositions, are written separately as two different files for further processing. A sample of the dictionary is given below:

0001	SKIN
0002	BEND
0003	BELLY
0004	OFFAL
0005	HALF BACK
0006	SPLIT
0007	FLESH SPLIT

Manipulation of Subject-propositions

The coded subject-propositions are manipulated to generate term-pairs (terms denoted by serial numbers) following the links indicated by the codes. Once an entry is prepared its reverse entry is automatically generated by changing the position of the 'lead term' and the 'context term'. In hierarchic relationships the relation NT is changed to BT in reversing the entry. In RT entries the relation does not change in the reversal. In the case of entries having the role indicator codes, the position of '-' is changed from prefix to suffix and vice versa as appropriate. In the case of Synonymous relationship indicated by '/' in the input, a SYN and a reverse USE entries are generated. These processes are done by a program named 'GENTHES'. The entries for the thesaurus at this stage are in the form of serial numbers standing for the 'lead' and the 'context' terms with the role indicator code in between them. The entries look as shown below:

```
0009RT(08-)0433
0433RT(-08)0009
0010NT(16-)0011
0011BT(-16)0010
```

Generation of Coordinate Term-pairs

The term-pairs so for generated are the hierarchic and non-hierarchic associative types. Terms coordinate to a particular term are not present in them. In order to generate coordinate entries, the generated entries are sorted in ascending sequence so that, 'context' serial numbers for the same 'lead term' (having the same serial number in the lead term position and having the same role indicator code) that are NTs, are formed as a separate table and coordinate RT term-pairs are generated among them. These coordinate entries are merged with the earlier generated entries, and passed as a file for further processing. The generation of coordinate entries is done by a program named 'GENCORD'.

Translation of Thesaurus Entries

The file of generated entries for thesaurus is retranslated back into natural language terms by a program named 'TRANSLAT'. The term dictionary created as a file by the program CODER is read together with role indicator codes and their corresponding descriptive phrases. The file of thesaurus entries, passed on by the program GENCORD, is read record by record. The serial number of both the 'lead' and 'context' terms are translated into natural language terms using the term dictionary. The role indicator code is also translated into the corresponding descriptive phrase. The translated entries are written as a file for further sorting and printing.

Translation to Different Languages

In order to translate the generated thesaurus into another language the term dictionary and the descriptive phrases denoting the role indicator codes are replaced by equivalent terms in the required language. Incompatibility of terms though pose some problems it is possible to form these two files easily [24] . But care must be taken to choose the correct standard terms and synonyms. The term dictionary if dumped out has an indication as to which terms are taken as synonyms, which must be taken care of in preparing the 'translation table'.

Sorting and Printing Thesaurus

The file of thesaurus entries in natural language terms, output of the program 'TRANSLAT', is sorted alphabetically using the SORT program available in the computer system. It is then printed out in double column format with proper indention for 'lead' term, relation, role indicator, and 'context term'.

Programs Developed for Thesaurus

The programs developed for generating thesaurus as outlined in this paper are written in COBOL and ASSEMBLER languages for IBM System/370 series computers and require a 256K partition, two tape drives, one disk drive and a line printer. The programs have been used to generate a thesaurus of Leather Technology terms using test data of about 1500 subject-propositions. The number of unique terms were 1851 , the total number of entries were 13,717. The thesaurus generation work took about 3 months of input preparation by two persons and 10min 26.73secs of CPU time at an IBM System/370-155. The programs were kept as load modules and were executed.

Conclusion

The study of linguistics in general, and the theories of universal grammar and structure of languages in particular, provide a frame-work for the development of scientific languages - artificial languages for specific purposes - relevant to applications in the different links in the'communication chain' that links creators of information and users of the same. The development of the theory of Subject Indexing Languages and its applications in the field of information storage and retrieval is a clear indication of this development.

Bibliographical References

1 Auguston, J G and Minker, Jack. Deriving term relations for a corpus by graph theoretical clusters. (Jl of Amer Soc for Inf Sc. V 21, no 2; 1970, Mar; p 101-111).

2 Balasubramanian, V. Computer-based thesaurus generation from modulated subject structures. (Guide: F J Devadason). Research Project. DRTC, 1978.

3 Bhattacharyya, G. A general theory of SIL, POPSI and Classaurus: Results of current classification research in India. Paper presented at the International Classification Research Forum, organised by SIG(CR) of American Society for Information Science. Minneapolis. Oct 1979.

4 ---. Fundamentals of subject indexing languages. In Neelameghan, A, Ed. Ordering systems for global information networks: Proceedings of the Third International Study Conference on Classification Research, Bombay, India. 6-11 Jan 1975. FID/CR publication 553. DRTC, Bangalore. 1979. p 83-99.

5 ---. ---. ---. ---. ---. p 87.

6 ---. ---. ---. ---. ---. p 88.

7 ---. Intellectual foundation of POPSI. (mimeographed). DRTC, Bangalore. 1979.

8 ---. POPSI: Its fundamentals and procedure based on a general theory of subject indexing language. (Lib Sc with slant Doc. V 16, no 1; 1979, Mar; p 1-34).

9 ---. ---. ---. ---. ---. p 23-24.

10 Biller, H and Neuhold, E J. Semantics of data bases: the semantics of data models. (Inf Systems. v 3; 1978; p 11-30).

11 Chomsky, Noam. Aspects of the theory of syntax. The MIT Press. 1965. p 35.

12 ---. Reflections on language. Fontana. Collins. 1976. p 4.

13 Ciganik, Marek. Meta informational approach to the theory of integrated information retrieval systems. (Inf Processing and Mgt. v 11; 1975; p 1-10).

14 ---. Meta informational in action in the process of the automatic semantic analysis. (Inf Processing and Mgt. v 15; 1979; p 196 - 198).

15 Costello, J C. A basic theory of roles as syntactical control devices in coordinate indexes. (Jl of Chem Doc. v 4; 1964; p 116-123).

16 Costello, J C. Coordinate indexing. Rutgers, the State University. New Jersey. 1966. p 14-15.

17 Devadason, F J. Using taxonomy of concepts by subject specialists for thesaurus construction: A case study. DRTC Annual Seminar. v 15; 1977, Dec; p 179.

18 Documentation Research and Training Centre. Seminar on thesaurus in information systems. Bangalore, India. 1-5 Dec 1975. 315 p.

19 Fillmore, D J. "The case for case". In Bach, Emmon and Harms, R T, Ed. Universals in linguistic theory. Holt, Rinehart and Winston Inc. New York. 1968.

20 Foskett, A C. The Subject approach to information. Clive Bingeley, London. 1977. p 387, 412, 426.

21 Foskett, D J. Systems theory and its relevance to documentary classification. (Intl Classification. v 7, no 1; 1980; p 2).

22 Gilchrist, A. Thesaurus in retrieval. ASLIB, London. 1971.

23 Hutchins, W J. The concept of 'aboutness' in subject indexing. (Aslib Proceedings. v 30, no 5; 1978, May; p 172-181).

24 Iljon, A. Creation of thesauri for Euronet. In Overcoming the language barrier: Third European Congress on Information Systems and Networks. Luxembourg, 3-6 May 1977. v 1. Ed 2. K G Saur, New York 1978. p 426.

25 ---. ---. ---. ---. ---. p 427.

26 International Atomic Energy Agency. INIS: Thesaurus. IAEA-INIS-13. IAEA. Vienna. 1970. p 5.

27 Ivanova, N S. Automatic compiling of thesauri on the basis of statistical data. 7th Annual Meeting of the Committee for International Cooperation in Information Retrieval among Examining Patent Offices Stockholm, 18-29 Sept 1967. ICIREPAT, BIRPI, Geneva. 1968. p 92-107.

28 Janos, Jiri. Theory of functional sentence perspective and its application for the purpose of automatic extracting. (Inf Process and Mgt. v 15, no 1; 1979; p 22).

29 Jardine, N and Van Rijsbergen, C J. The use of hierarchic clustering in information retrieval. (Inf Storage and Retrieval. v 7; 1971; p 217-240).

30 Jones, Karen Sparck. Synonymy and semantic classification. Ph.D. Thesis, University of Cambridge. 1964. p 1.7-1.9.

31 ---. ---. ---. ---. ---. p 3.8.

32 ---. ---. ---. ---. ---. p 6.21.

33 Kim, Chai and Kim, Soon D. Consensus Vs frequency: An emprical investigation of the theories for identifying descriptors in designing retrieval thesauri. (Inf Processing and Mgt. v 13, no 4; 1977; p 253-258).

34 Leech, G N. Towards a semantic description of English. Bloomington, Indiana University Press. 1970.

35 Maitra, Ranjita. Semi-automatic method of generating micro-thesaurus. Project Report. DRTC, Bangalore. 1977.

36 Maron, M E. On indexing, retrieval and the meaning of about. (Jl of Amer Soc for Inf Sc. v 28, no 1; 1977, Jan; p 38-43).

37 Mcneill, D. Empiricist and naturist theories of language: George Berkeley and Samuel Bailey in the 20th century. In Koestler, A and Smythies, J R Ed. Beyond reductionism: New perspectives in the life sciences. Alpback Symposium, 1968. Hutchinson, 1969. p 291-292.

38 Montgomery, Christine A. Linguistics and information science. (Jl of Amer Soc for Inf Sc. v 23, no 3; 1972, June; p 214-215).

39 Minker, Jack, Wilson, G A and Zimmerman, B H. An evaluation of query expansion by the addition of clustered terms for a document retrieval system. (Inf Storage and Retrieval. v 8; 1972; p 329-348).

40 National Academy of Sciences, National Research Council, Automatic Language Processing Advisory Committee. (Chairman: Pierce, John R). Language and machines: Computers in translation and linguistics-A report. National Research Council Publication 1416, Washington D C. 1966. p 30.

41 Neelameghan, A. Absolute syntax and structure of an indexing and switching language. In Above cited ref.no. 4, p 170.

42 ---. ---. ---. ---. ---. p 173.

43 ---. Sequence of component ideas

in a subject. (Lib Sc with slant to Doc. v 8; 1971; p 323-334).

44 Pickford, A G A. Some problems of using an unstructured information retrieval language in a co-ordinate indexing system. (Aslib Proceedings. v 23, no 3; 1971, Mar; p 133-138).

45 Ranganathan, S R. Colon Classification. Ed 1, 1933. .. Ed 6, 1960. Asia Publishing House, Bombay.

46 ---. Hidden roots of classification. (Inf Storage and Retrieval. v 3; 1967; p 399-410).

47 ---. Prolegomena to library classification. Ed 1, 1937. Ed 2, 1957. Asia Pub House, Bombay.

48 ---. ---. Ed 3. Asia Pub House, Bombay. 1967. p 86-87.

49 ---. ---. ---. ---. ---. p 395-434.

50 ---. ---. ---. ---. ---. p 404-405.

51 ---. ---. ---. ---. ---. p 439.

52 ---. ---. ---. ---. ---. p 579-582.

53 Ravichandra Rao, I K. Semi-automatic method of construction of thesaurus. In Above cited ref. no.18, p B16-B30.

54 Rolling, L N. Compilation of thesauri for use in computer systems. (Inf Storage and Retrl. v 6, no 4; 1970, June; p 341-350).

55 Sager, Naomi. Computational linguistics: Steps towards application. Paper prepared for the Workshop in Linguistics and Information Science. FID/LD. Stockholm, 3-5 May 1976. p 18.

56 Salton, G. Automatic information organisation and retrieval. McGraw Hill, New York. 1968.p 40-48.

57 Sasmori, Katsunosuke. Software design for vocabulary control (DOCTOR) system. In North, Jeanne B, Ed. The information conscious society: Proceedings of the American Society for Information Science. v 7; 33rd Annual Meeting. Philadelphia. 11-15 Oct 1970. ASIS, 1970. p 195-197.

58 Schwanhausser, G. An automatically produced thesaurus. In Above cited ref. no. 18. p B1-B15.

59 Shephard, Michael and Watters, Caroline. Computer generation of thesaurus. (Lib Sc with slant to Doc. v 12; 1975; Paper E).

60 Spang-Hanssen, H. Are classification systems similar to natural languages? In Above cited ref. no. 4. p 15.

61 Tanimoto, T T. An elementary mathematical theory of classification and prediction. IBM Research Yorktown Heights. New York. 1958.

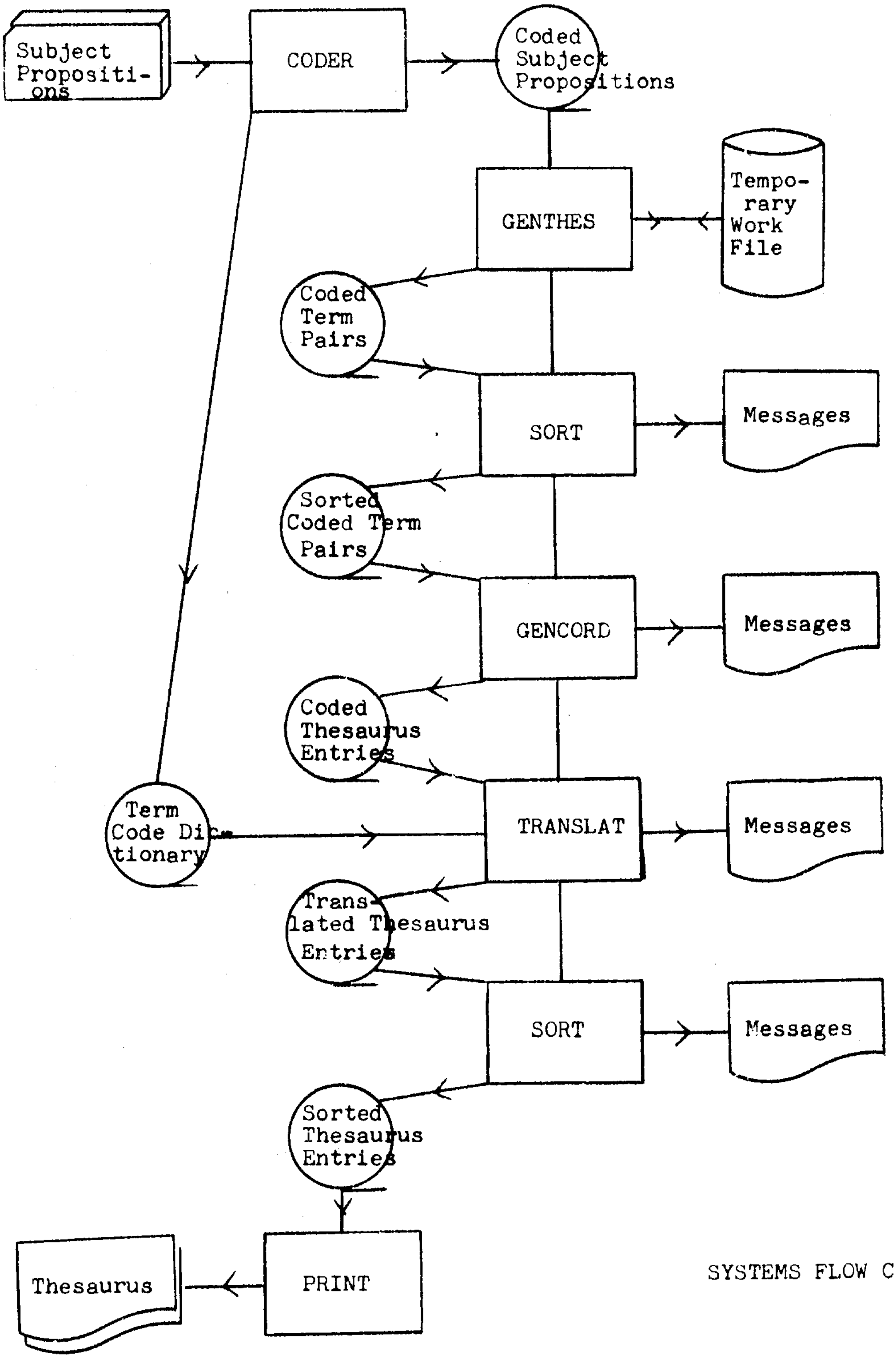
62 Thompson, David A. Interface design for an interactive information retrieval system: A literature survey and a research system. (Jl of Amer Soc for Inf Sc. v 22, no 6; 1971, Nov-Dec; p 361-373).

63 Vickery, B C. On retrieval system theory. Ed 2. Butterworths, London. 1965. p 58-60, 97.

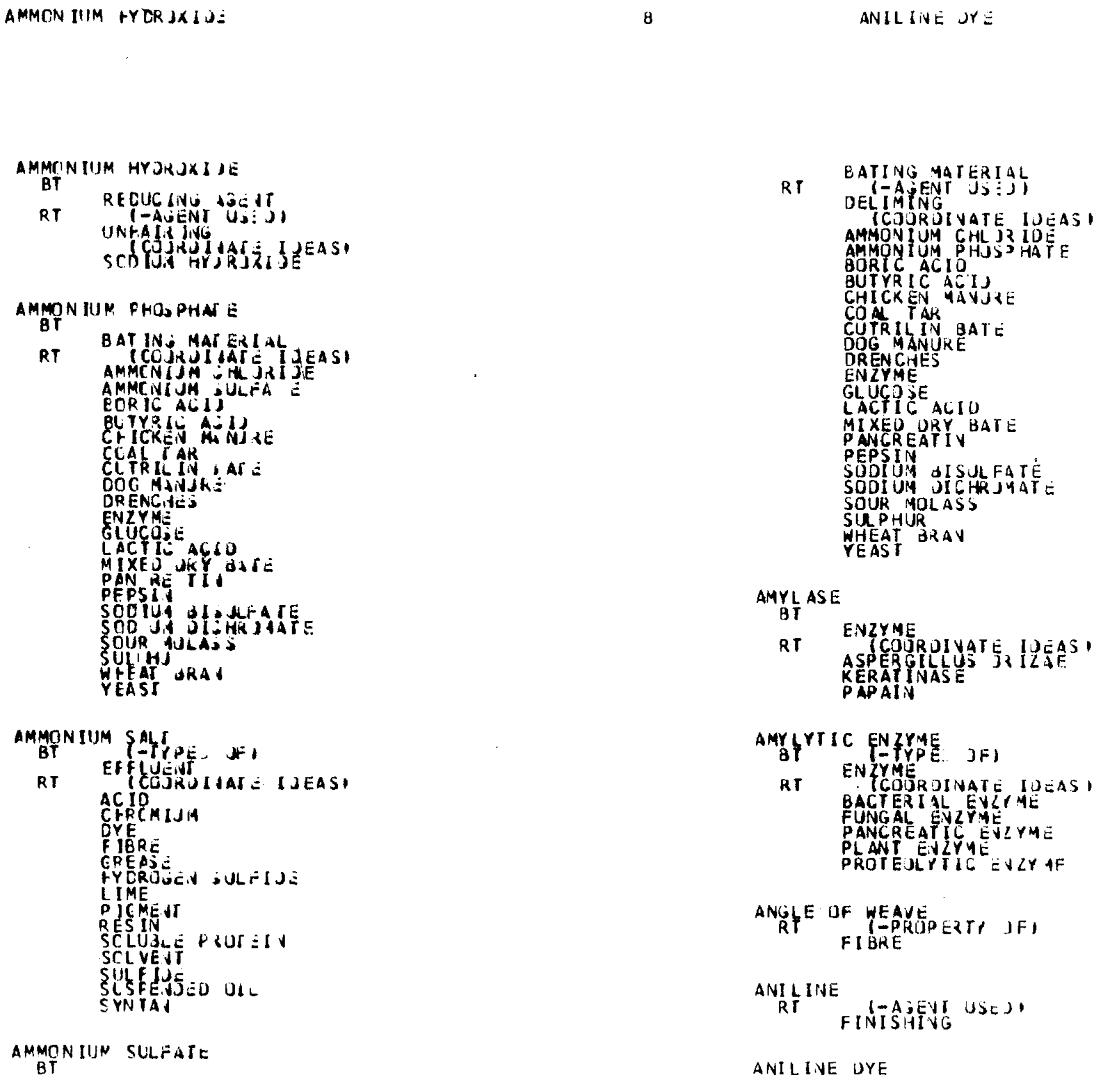
64 Wall, Eugene. Symbiotic development of thesauri and information systems: A case history. (Jl of Amer Soc for Inf Sc. v 26, no 2; 1975, Mar-Apr; p 71-79).

65 Wolf-Terroine, M, Rimbert, D and Rouault, B. Improved statistical methods for automatic construction of a medical thesaurus. (Methods of Inf in Medicine. v 11, no 2; 1972, Apr; p 104-113).

66 Yu, Clement T. A methodology for the construction of term classes. (Inf Storage and Retrieval. v 10; 1974; p 243-251).



SYSTEMS FLOW CHART



```
AMMONIUM HYDROXIDE                 8                  ANILINE DYE


AMMONIUM HYDROXIDE
  BT
        REDUCING AGENT
  RT       (-AGENT USED)
        UNHAIRING
           (COORDINATE IDEAS)
        SODIUM HYDROXIDE

AMMONIUM PHOSPHATE
  BT
        BATING MATERIAL
  RT       (COORDINATE IDEAS)
        AMMONIUM CHLORIDE
        AMMONIUM SULFATE
        BORIC ACID
        BUTYRIC ACID
        CHICKEN MANURE
        COAL TAR
        CUTRILIN BATE
        DOG MANURE
        DRENCHES
        ENZYME
        GLUCOSE
        LACTIC ACID
        MIXED DRY BATE
        PANCREATIN
        PEPSIN
        SODIUM BISULFATE
        SODIUM DICHROMATE
        SOUR MOLASS
        SULPHUR
        WHEAT BRAN
        YEAST

AMMONIUM SALT
  BT       (-TYPES OF)
        EFFLUENT
  RT       (COORDINATE IDEAS)
        ACID
        CHROMIUM
        DYE
        FIBRE
        GREASE
        HYDROGEN SULFIDE
        LIME
        PIGMENT
        RESIN
        SOLUBLE PROTEIN
        SOLVENT
        SULFIDE
        SUSPENDED OIL
        SYNTAN

AMMONIUM SULFATE
  BT

        BATING MATERIAL
  RT       (-AGENT USED)
        DELIMING
           (COORDINATE IDEAS)
        AMMONIUM CHLORIDE
        AMMONIUM PHOSPHATE
        BORIC ACID
        BUTYRIC ACID
        CHICKEN MANURE
        COAL TAR
        CUTRILIN BATE
        DOG MANURE
        DRENCHES
        ENZYME
        GLUCOSE
        LACTIC ACID
        MIXED DRY BATE
        PANCREATIN
        PEPSIN
        SODIUM BISULFATE
        SODIUM DICHROMATE
        SOUR MOLASS
        SULPHUR
        WHEAT BRAN
        YEAST

AMYLASE
  BT
        ENZYME
  RT       (COORDINATE IDEAS)
        ASPERGILLUS ORIZAE
        KERATINASE
        PAPAIN

AMYLYTIC ENZYME
  BT       (-TYPES OF)
        ENZYME
  RT       (COORDINATE IDEAS)
        BACTERIAL ENZYME
        FUNGAL ENZYME
        PANCREATIC ENZYME
        PLANT ENZYME
        PROTEOLYTIC ENZYME

ANGLE OF WEAVE
  RT       (-PROPERTY OF)
        FIBRE

ANILINE
  RT       (-AGENT USED)
        FINISHING

ANILINE DYE
```

THESAURUS SAMPLE PAGE