SEMINAR ON THESAURUS (1975). Paper AJ

TERM-CONCEPT RELATIONSHIP IN AN INFORMATION RETRIEVAL THESAURUS

S SEETHARAMA
Documentation Research & Training Centre, Indian Statistical Institute, Bangalore

Concept representation in Thesauri, especially the representation of compound concepts, has been studied. The criteria and relative advantages in representing a compound concept either as a precombined descriptor or as a combination of elemental descriptors, are discussed with illustrative examples taken from Medical Subject Headings (MeSH) of the National Library of Medicine.

0 INTRODUCTION

Information Storage and Retrieval Systems (IRS) are primarily concerned with the retrieval of a document (or a surrogate) relevant to a concept or concepts. This, however, is possible only by referring concepts to the appropriate terms used in the system. This, in turn, presupposes the identification and recognition of inter-relationships between concepts and terms. For this purpose, it is essential that one distinguishes between the plane of concepts and plane of terms, as ignoring of this distinction would result in confusion. Concepts belong to the Idea Plane and the terms belong to the Verbal Plane. The terms used to designate concepts may be either single-worded or multi-worded. This is complicated further by the existence of synonyms, quasi-synonyms, homonyms, eponyms, acronyms, abbreviations, etc. in the written and spoken language. Therefore, it becomes necessary to consider terminological problems and to introduce terminological control in the practical construction of a thesaurus.

1 SCOPE OF THE PAPER

In this paper, an attempt is made to study concept representation in Thesauri, especially the representation of compound concepts. The criteria and relative advantages in representing a compound concept either as a precombined descriptor or as a combination of elemental descriptors, are discussed with illustrative examples taken from Medical Subject Headings (MeSH) of NLM.

2 DEFINITIONS

The operational definitions of some of the terms used in the paper are given below

a) Term - A word or word-group denoting an idea and enumerated in a thesaurus or scheme for classification of subject.

i) Uniterm - A single term representing a basic idea is a uniterm and is a 'free' term.

ii) Compound terms- Terms which have been pre-coordinated.

iii) Composite term - A term which is single or multi-worded and representing a compound concept.

iv) Preferred term - A term that is selected from a class of synonymous and quasi-synonymous terms to designate unequivocally the concept underlying the class.

b) Concept - A concept is a theoretical construct or organised thought about a phenomenon or a set of phenomena combining all of its essential characteristics and is a component of a given field of study or a subject. It represents an abstractable, public essential, and agreed form of an entity (concrete or conceptual) and is expressed in the 'formal context' of any discourse.

i) Elemental concepts - Concepts that cannot be decomposed or split further in a given system.

ii) Compound concept - A concept that can be mentally split or factored or decomposed into separate concepts, the mental addition of which would generally lead back to the initial concept.

c) Descriptor - A formalised, standardised or controlled term or symbol representing one/ or a combination of concepts in an unambiguous or univocal way, and used

a) in the selection and/or arrangement of documents (in indexing) and
b) in the selection and/or arrangement of documents or their substitutes in a given system by a given mechanism.

i) Elemental descriptor - A descriptor designating an elemental concept.

ii] Precombined descriptor - A descriptor designating a compound concept.

iii) Non-descriptor - A forbidden term in the controlled vocabulary, i. e., any thesaurus term or symbol not considered as a descriptor.

d) Thesaurus - A tool of information systems, is an alphabetically and/or systematically ordered collection of descriptors and non-descriptors having hierarchical and/or non-hierarchical interrelationships.

CONCEPT REPRESENTATION IN THESAURUS

Concepts -- unitary and compound -- are represented in a thesaurus by use of terms which may be single-word terms or multi-word terms, the relationship between the two (concepts and terms) governed by the rules of terminology. Synonyms, homonyms, eponyms, acronyms, etc used in language complicate the problem of concept representation by terms in a thesaurus. The problems thus encountered are overcome by the application of some rules for the form of terms in the process of thesaurus building.

31 Uniterms and Multi-warded/Compound Terms

Concepts -- unitary and compound -- can be represented either by single-worded terms or multi-worded terms. For example

i) Unitary Concept - Uniterm.
DISEASE. TOOTH.
ii) Compound Concept - Uniterm.
TUBERCULOSIS. APPENDICITIS
iii) Unitary Concept - Multi-worded term. RARE EARTH METALS.
VATER'S AMPULLA
] Compound Concept - Multi-worded term. ROCKY MOUNTAIN SPOTTED FEVER.
HEMORRHAGIC DISEASE OF NEW BORN.

However, some of the compound concepts can preferably be expressed as a combination of elemental concepts, The conditions under *which* compound concepts should be represented either as combination of elemental concepts or as pre-coordinated concepts are discussed in Sec 4.

32 Synonym and Quasi-Synonym

True synonyms are rarely met with. However, due to the presence and usage of trade names and popular names in scientific literature, synonyms have become problematic in IR systems. For example:

Trade name : CELIN = VITAMIN C

Popular usage : DANCING DISEASE = EPIDEMIC CHOREA
ATHELETE'S FOOT = TINEA PEDIS
CHRISTMAS DISEASE _ HAEMOPHILIA B

Eponyms : ADDISON-BIERMER DISEASE _ PERNICIOUS ANEMIA

The problem of synonyms is overcome by selecting one as the preferred term with provision of cross referencing from the other synonymous terms to the preferred term. The "See References" in a thesaurus are good examples of this kind. The number of such "See References" will generally be few in a thesaurus. For example :

DEFORMITIES	See ABNORMALITIES
SWALLOWING	See DEGLUTITION
ENAMEL	See DENTAL ENAMEL
HYPERCORTICISM	See ADRENAL GLAND HYPERFUNCTION

Quasi-synonyms which are closely related are treated as synonymous for retrieval language purposes.

GENETICS - HEREDITY
ACCURACY - PRECISION

Note : - The true synonyms and quasi-synonyms in a thesaurus represent Equivalence Relationship.

b) Post-coordinated synthesised concept-Linguistic: The concept may be represented by the combination of its constituent words. For example,

STAPHYLOCOCCAL ENTEROTOXIN Use STAPHYLOCOCCUS and ENTERS-TOXIN

BONE DISEASE Use BONE and DISEASE

C) Post-coordinated synthesised concept - Semantic factoring : The concept may be represented by the combination of its semantic factors. For example,

AUTOPSY Use POST and MORTALITY and EXAMINATION

PHTHISIS Use LUNG and TUBERCULOSIS

PERCEPTUAL COMPLETION PHENOMENA Use ILLUSIONS and VISION

CARDIAC FAILURE Use HEART and OUTPUT and BELOW and NORMAL

Note : In methods (III and (c) it is implied that compound concepts are represented in thesaurus by elemental descriptors. On the other hand, in method (a) it is implied that compound concepts are represented in thesaurus by precombined descriptors which maybe either single-worded or multi-worded terms.

42 Criteria for Choice

421 Precombined Descriptors

Soergel has summarised the following as some of. the reasons in using precombined descriptors in the practical construction of thesaurus :

a) To prevent "Fapse drops" and for increasing precision in retrieval.

b) To serve for the arrangement of documents or catalogue cards.

c) To decrease the number of descriptors needed to index a document.

However, semantic factoring of compound concepts according to their meaning is helpful in

i} Rendering explicit all essential aspects of a concept ;

ii) Detection of general concepts which later on might be useful for retrieval ;

iii) Arriving at a comparatively small set of elemental concepts which can be used in combination to represent all compound concepts, especially if syntax is used; and

iv) Searching for any combination of elemental concepts.

Therefore, a need arises for the identification of a set or criteria in deciding as to when a compound concept should be represented as a pre-combined descriptor in a thesaurus in preference to a combination of elemental descriptors. The criteria mentioned by different authors (1, 2, 6, 7) are summarised below

1) Frequency of use of a compound concept as a precombined descriptor in literature as well as in indexing and/or searching.

ALVEOLAR PROCESS, ACTINSMYCOSIS, TUBERCULOSIS, CARCINOMA, SPLENSMEGALY.

ii) When meaning of one of the terms would be changed as a result of combination.

E. g : LANDING LIGHTS, YELLOW FEVER, THYROID CRISIS, CHRISTMAS DISEASE, BLIND LOOP SYNDROME, BUNDLE BRANCH BLOCK.

iii) When each term of combination falls into a generic cpass which differs from the specific precombined term.

E. g.: BLIND LOOP SYNDROME

iv) When elemental descriptors of a compound concept are used as in the original uniterm approach, unwanted combinations are likely to occur at the retrieval stage resulting in what are known as 'false drops'.

E. g : FISH and TOXINS can represent *both TOXINS* PRODUCED BY FISH as also TOXINS AFFECTING FISH. In such a case, FISH TOXINS can be made to represent the former, while the mupti-worded precombined descriptors TOXINS AFFECTING FISH can be made to represent the latter. Other examples are : YELLOW FEVER, CHRISTMAS DISEASE, etc.

v) To maintain syntactical relationship of

components. For example, when two compound concepts have the same elemental concepts.

E. g: LIBRARY SCHOOLS and
SCHOOL LIBRARIES,
ADMINISTRATIVE PERSONNEL
and PERSONNEL ADMINISTRATION

FOOD POISONING and
POISONED FOOD

vi) To achieve logical completeness at a location in the hierarchy or if it is to be used in the checklist technique of indexing.

E. g : AGRARIAN REFORM
INSECT VIRUS

vii) When a compound concept has broader, narrower, or related concepts that cannot be seen from the derivation rules, a precombined descriptor should be used.

E. g : PRIMATES. This is narrower than the compound concept MAMMAL which can be expressed by the combination VERTEBRATES : SUCKLING FORMS. But, PRIMATES cannot be derived from VERTEBRATES : SUCKLING FORMS by adding a component or narrowing down a component. Other examples are : HELICOPTER, FOXES.

viii) A compound concept should be used as a precombined descriptor if such use does not lead to an increase in the indexing language. The precombined descriptor HELICOPTER should be used rather than AIRCRAFT : ROTARY WING, since the elemental descriptor ROTARY WING is not useful in any other context. Therefore, the introduction of ROTARY WING rather than HELICOPTER will not reduce the overall number of descriptors in the indexing panguage.

ix) In doubtful cases, a precombined descriptor should be used as at a later stage, it may easily be reduced to its elemental descriptors, whilst the coordination of elemental descriptors is *not easily* ***accomplished*** **retrospectively.**

4211 Advantages and Disadvantages

The advantages in using precombined descriptors (1) Ensures high precision because they are specific and (2) Ensures that commonly used singleworded/multi-worded compound concepts by the scientific community appears in indexing and retrieval language, the relationships between the concepts being shown in the alphabetical thesaurus or classification display.

However, there are three disadvantages in using precombined descriptors : (1) Adds to indexing costs by inflating the vocabulary size; (ii) Causes recall failure : and (iii) Binds terms unnecessarily.

422 Elemental Descriptors

When none of the criteria mentioned in the previous section apply, it is helpfup to represent a compound concept by a combination of elemental descriptors in preferance to precombined descriptors. In addtion, the following criteria are applicable :

1 Compound concepts descriptive of an object made up of a certain material can be represented by a combination of elemental descriptors, one for the object and the other for the material. For example, METAL: TUBS

2 Many chemical compounds can be expressed by a combination of descriptors, eac standing for a specific group of atoms. For example

AMMONIUM + SULPHATE

BARIUM + SULPHATE

MAGNESIUM + SULPHATE

SODIUM + CHLORIDE

SODIUM + HYDROXIDE

SODIUM + HYPOCHLORIDE

4221 Advantages and Disadvantages

The use of combination of elemental descriptors for the representation of compound concepts gives better recall, but not high precision because they are less specific. However the use of elemental descriptors has the disadvantage in that the indexers and searchers may use a different combination of terms to indicate the same concept, resulting in a recall failure. For example

a) FUEL STORATE TANKS

Indexer : FUELS and STORAGE TANKS
Searcher : FUEL STORAGE and TANKS

b) DRUG WITHDRAWAL SYMPTOMS

Indexer . DRUG and WITHDRAWAL SYMPTOMS
Searcher: DRUG WITHDRAWAL and SYMPTOMS

c) FEMORAL NECK FRACTUP L'S

Indexer : FEMUR and NECK FRACTURES
Searcher : FEMORAL NECK and FRACTURES

However, this recall failure can be avoided by the provision of an adequate entry vocabulary. One other limitation with elemental descriptors is when species or subdivisions of the term are required. For example, if 'Bones of the fore-arm" is synthesized from 'Bones' and "Fore-arm" , it is not possible to construct the index term for the narrower term 'Lateral bone of the fore arm" (Radius), if there is no index term "Lateral bone" listed as a narrower term under "Bone".

5 MEDICAL SUBJECT HEADINGS

The Medical Subject Headings (=MeSH) appeared for the first time in 196G with the inception of Index Medicus. This was based on the Subject Heading Authority List (1954) of NLM which was based on the internal authority list that had been used for publication of Current List of Medical Literature. This, in turn, had incorporated headings from the Library's Index Catalogue and from the 1940 Quarterly Cumulative Index Medicus Subject Headings (5). MeSH is the authority list of technical terms_ used for the subject analysis of the biomedical literature in the NLM. The purpose of MeSH is to

i) Provide a list of Subject Headings for use in the assignment of appropriate Subject Headings for any main entry in the Index Medicos ;

ii) Serve as the basis for search formulations in retrieval by computer of bibliographical citations stored on the MEDLARS ; and

iii Be the authoritative Subject Heading List for the subject cataloguing of books and periodicals in NLM.

For the achievement of the objectives mentioned above, it has been found essential for NLM to keep track of developments in the biomedical literature and modify the MeSH as and when necessary. The 1974 Medical Subject Headings contains about 9, 700 descriptors (Main Headings) compared with 8, 500 in 1968, 5, 70C in 1963 and 4, 400 in the 1960 edition. The subheadings which can be used in combination with the main headings number 60.

51 McSH as a Thesaurus

MeSH can be considered as a thesaurus for information retrieval purposes, since it provides a conceptual structure as well as terminological control. It contains a set of descriptors which are indicative of conceptual relationships -- such as hierarchical and non-hierarchical relationships among concepts.

511 Hierarchical Relationships

The hierarchical relationships among concepts can be seen in the categorized lists section of MeSH which displays the terms in separate categories arranged hierarchically and semantically with an alphanumeric designation for each category and subcategory. However, the categorized lists are not a complete classification system for every biomedical discipline since they contain only terms which have been selected for inclusion in the vocabulary (3). Reference to this listing from the alphabetical list is obtained by using the category number attached to each term. In the alphabetical listing of subject headings, the usage of cross reference "See also specific" (XS) enables to recognise the hierarchical relationship between concepts. However, these are sparingly used since the categorized listing provides for the more specific subterm to be listed under the general term. Further, the provision of wholly integrated "Tree structures" of MeSH, make MeSH an alphabetical thesaurus with hierarchical classification. "The terms/ descriptors are arranged in subject groups and within each group terms are ordered hierarchically. The link between the alphabetical thesaurus and the location of the term in the classification is provided by the detailed notation. In the alphabetical MeSH, there are few BT/NT references, as these are displayed in the tree structures, but some related RT terms are shown in the alphabetical thesaurus. When a term occurs in more than one hierarchy, it is listed in all appropriate places in the tree structures, and all the class numbers are indicated against the term in the alphabetical thesaurus" (1). Examples of "See also specific" entries and Alphabetical MeSH with tree structures are given below:

"See also specific" entries

i) BIGUANIDES (D2)

See also specific
Metformin (D8)
Phenformin (D8)

ii) FISTULA (C17)

See also specific
Arteriovenous fistula (C8)
Biliary fistula (C4)
Bladder fistula (C6)
Urinary fistula (C6)
Vaginal fistula (C6)
Vesicovaginal fistula (C6)

iii) SALICYLIC ACID (D2)

See also specific
Arninosalicylic acid (D3)
Aspirin (n6)

iv) VISCOCITY (H)

See also specific
Blood viscosity (G1)

Alphabetical MeSH with tree structures
Alphabetical MeSH

TINEA FAVOSA

C1. 40. 27. i; C12. 14. 55. 1
X Favus (Cl, C12)

TINEA PEDIS

C1. 40. 27. i ; C1 Z. 14. 55. 1
C12. 31. 32 ; C17. 25.24. 1
X Athlete's foot (C1, C12, Cl7)

TISSUE BANKS

N2. 18. 54
XU Eye banks (N2)

Tissue compatability See Histocompativity bility (G1)

Tree Structures

DERMATOMYCOSIS	C12. 14	01.40.27	012.94.8
ACTINOMYCOSIS, CERVICOFACIAL	C12.14.11 012.14.16	01.10.16.1 01.40.27.1	
TINEA	C 12. 14. 55	C 1. 40. 27. 1	
TINEA CAPITIS	C12.14.55.1	C 1. 40. 27. 1	C12. 76. 23
TINEA FAVOSA	C 12. 14. 55. 1	C 1. 40. 27. 1	
TINEA PEDIS	012.14.55.1	C1.40.27.1	C12.31.32 C 17. 25.24
TINEA UNGUIUM	C 12. 14. 55. 1	C 1. 40. 27. 1	
DERMATOMYOSITIS	012.16	03.80.35.1	C17.16.18
ERYTHEMA	C12. 19		
FAVRE-RACOUCHOT SYNDROME	012.30		
FOOT DERMATOSIS	C12.31	C17.25.24.1	
TINEA PEDIS	C12.31.32	C1.40.27.1	C12.14. 55.1 C17.25.24.
HAND DERMATOSES	C12.34		

512 Non-Hierarchical Relationships

The non-hierarchical relationships can be said to include the associative and equivalence relationships. In the MeSH, these relationships among concepts are represented by the usage of cross references "See also related" (XR) for associate relationship, and"See, and See under" (X) (XU) for equivalence relationships. The "See also related" reference is used primarily to indicate related terms not occurring in the same category that may contain citations of direct pertinence to the area of interest. However, the most obvious references, as those from an organ to its diseases, have been avoided. On the other hand, the "See" reference directs from a synonym to the preferred term appearing in MeSH. The "See under" reference is used to refer from a specific term not in the list to a more general heading under which it is indexed. (The specific term is considered as quasi-synonymous, and therefore the specific term representing a specific concept is subsumed under broader terms). (1). Examples from MeSH of the above mentioned cross-referencing are given below.

"See also related" entries

i) ANGIOMATOSIS (C2)

See also related
Arteriovenous malformations (C16)
Telangiectasia, Hereditary hemorrhagic (C8, C9)

ii) JURISPRUDENCE (N3)

See also related
Forensic dentistry (G2, I)
Forensic medicine (G2, I)
Malpractice (N3)

"See" entries

i) KALA-AZAR see Leishmaniasis, Visceral (C1)
ii) NERVE CONDUCTION see Neural conduction (GI)
iii) PERIADENITIS see Stomatitis, Aphthous (C 1, C4)

"See under" entries

i) PERISTALSIS see under Gastrointestinal motility (G l)
ii) PSEUDONYMS see under Anonyms and pseudonyms (L)
iii) PSYCHICAL RESEARCH see under Parapsychology (Fl, Fe)

52 Term-Concept Relationship in MeSH

Term-Concept relationship in MeSH follows the general pattern outlined in Sec 3 and 4. Almost all the examples cited in those sections have been taken from MeSH. "

6 CONCLUSION

In the development of an effective and efficient thesaurus for information retrieval purpose, considerable intellectual effort is necessary. The intellectual problems encountered in thesaurus building maybe summarised as follows

i) Delineation of concepts ;
ii) Definition of concepts ;
iii) Concept representation in thesaurus by use of elemental and precombined descriptors; and
iv) Arrangement of concepts in a structured system or network, that is, establishing for each concept its hierarchical and non-hierarchical relationships.

One other point that needs to be mentioned is that a thesaurus is never complete. It has necessarily to be updated continuously based on practical experience gained in application. In other-words, thesaurus should reflect the most recent developments in the subject field concerned (6).

7 BIBLIOGRAPHICAL REFERENCES

1 Sec 4 421 511 512 AITCHISON (3) and GILCHRIST (A). Thesaurus construction A practicap manuap. 197Z

2 Sec 421 GILCHRIST (A). Thesaurus in retrieval. 1 971

3 Sec 511 NATIONAL LIBRARY OF MEDICINE. Medical Subject Headings. 1974-

4 Sec 35 NEELAMEGHAN (A). Non-hierarchical associative relationships. Their types and computer generation of RT links, (Seminar on Thesaurus in Information Systems. (Bangalore) (1975). Paper AC)

5 Sec 5 SEWELL (W). Medical Subject headings in MEDLARS. (Bull Medical Lib Assn. 5Z, 1 q64 ;

6 Sec 421 6 SOERGEL (D). Indexing languages and thesauri : Construction and maintenance. 1974,

7 Sec 421 THESAURUS BUILDING. (International Conference on general principles of --) (Warsaw) (1970). Proceedings. 1970.