Theoretical Problems of Thesaurus Building with Particular Reference to Concept Formation.

Dagobert Soergel
Associate Professor
College for Library and Information Services
University of Maryland

Preface

This paper draws together, for the use of this conference, materials from the books "Indexing Languages and Thesauri: Construction and Maintenance" and "Dokumentation und Organisation des Wissens". I saw my task in the synthesis of ideas for a specific audience rather than in the presentation of new ideas.

Outline

A Definitions

B Functions of a thesaurus

B 1 Functions of the indexing language

B 1.1 Further remarks on the functions of an indexing language in the research process

B 1.2 Multimodel indexing

B 2 Functions of the lead-in structure

B 3 "User's" or "author's[11] vocabulary versus logical structure and request-oriented indexing through the checklist technique

B 4 Complex thesaurus structures for natural language searching and automatic indexing

C Steps in the construction of indexing languages and thesauri

I) Problems of concept formation in the different steps of thesaurus building

A Definitions

Since in the field of information science in general and in the area of classification theory in particular concepts and, partly as a consequence, terminology are as muddled as in any field, it is appropriate to define first some basic concepts to be used throughout this paper.

The first two definitions have to do with the relationship between concepts and terms (or other symbols):

An ISAR (Information Storage And Retrieval) concept is a concept that has been formed by confounding or consolidating several widely overlapping or very similar concepts, e.g., Attorney, Barrister, and Solicitor.

A preferred term is a term that has been chosen to unequivocally designate the ISAR concept, e.g., Attorney.

By the use of the preferred terms, terminological control is introduced. (In a wider sense, terminological control includes any provisions that prevent or decrease retrieval failures stemming from terminological problems.)

The next four definitions have to do with the functioning of an ISAR system:

Descriptor 1 (retrieval cue) = any string of symbols or other marks used (1) in the description and representation of documents and (2) in the selection and/or arrangement of documents (retrieval objects) or their substitutes in a given system by a given mechanism.

This definition includes author and title (in an author/title catalog), date of publication (in a mechanized system or in a card catalog if sequential scanning of the card by the human eye is the "given mechanism"), subject headings (in a catalog arranged by subject headings), a term or notation causing the appropriate peek-a-boo card to be selected for searching or to be punched in file building (if the ISAR system uses peek-a-boo cards), the words and phrases or whole sentences in an abstract (if sequential scanning by a human searcher is considered to be the "given mechanism").

I have given this general definition because "descriptor" is sometimes used in this general sense, especially in mechanized systems. The following definition corresponds more nearly to the common use of the term "descriptor".

Descriptor 2 (subject descriptor) = Descriptor 1 narrowed to include only strings of symbols designating ISAR concepts used in subject indexing and searching. Descriptor 2 may be a term or a notation or another string of symbols used to designate the ISAR concept. In the rest of this paper, the term "descriptor" will be used with the meaning of descriptor 2; whenever descriptor 1 is meant, it will be so designated.

Terminological control has not been included in the definition. This has consequences for the concept of an indexing language to be defined shortly. However, in the remainder of this paper terminological control will be assumed unless specified otherwise. A descriptor then designates unequivocally an ISAR concept actually used (or intended to be used) for indexing and retrieval purposes. In other words, every descriptor is a preferred term (but not vice-versa). Most commonly, the term "descriptor" is used with these connotations. (If it is necessary to refer specifically to the string of characters that constitutes a preferred term, I shall speak of the text of the descriptor. Notation has been mentioned in the definition because in many systems the notation is the string of symbols used in indexing and searching. Again, if it is necessary to refer specifically to the notation, I shall speak of the notation of the descriptor.) Based on the definition of "descriptor" it is now possible to define "indexing language".

Indexing language (documentary language), as used in this paper = any language (broadly defined) for the representation and/or for the arrangement of retrieval objects and/or their substitutes with the objective of making the items retrievable.

An indexing language comprises:

(a) (mandatory) a list of descriptors, the system vocabulary or lexicon, relationships among descriptors (such as hierarchical relationships or associative relationships) may be indicated. The system vocabulary or lexicon is often called classification scheme, especially if the descriptors are broader than some sort of classified arrangement.

Examples of descriptors are: Hammer; Form of government; Theory; France; Graduate-level text; 621.21 (Water wheels, from DDC); SB 211 .P8 (Potatoes, LCC); NQCL.MACH.003 (Reactor, Semantic Code); Ph Rab Ssd Zb (Dimensional stabi1ity of plastic at high temperature, a faceted classification)

(b) (optional) a list of role indicators and/or relators (relational descriptors). Again, hierarchical or associative relationships among role indicators or among relators may be indicated. Examples (role indicators or relators, respectively, in bold face):

role indicators:

(Effect : Noise) : (Cause : Children);

(Starting materials : Hydrogen and Oxygen): (Final product : Water);

Relators:

Noise : Caused by : Children;

Hydrogen and Oxygen : Produce : Water ;

One could say that the role indicators and/or the relators have a syntactical function and therefore belong to the syntax of the indexing language (as distinguished from the vocabulary given in (a)). However, there is no clear-cut frontier between descriptors and role indicators. The distinction is made more for practical than for theoretical purposes, and I shall not attempt to define the distinction beyond the illustration provided by the examples.

(C) a set of formation rules for constructing expressions in the indexing language as detailed below. In more elaborate indexing languages these rules are made explicit. More often than not, however, the rules are not spelled out explicitly, but assumed to be obvious.

(c 1) (mandatory) A set of rules for the construction of more or less compound expressions, using descriptors from the lexicon and syntactical elements. These compound expressions may either be document representations or formulations of search requests.

(c 2) (optional) A set of rules for the deduction of relationships between compound expressions and descriptors and between compound expressions themselves.

(c 3) (optional) A set of rules for the arrangement of compound expressions in a linear sequence. These rules may be used for the filing of catalog cards or for the shelving of documents.

A thesaurus is a set of terms or other symbols that consists of an indexing language and a lead-in vocabulary. The lead-in vocabulary contains terms (or other symbols) that are not part of the indexing language (that are not descriptors) but are included for the purpose of leading to the appropriate descriptor to be used, e.g.,

Lawyer

USE Synonymous Term Attorney

This definition applies to the most commonly used thesaurus type. To accommodate thesauri with a more complex structure, the definition must be generalized as follows: A thesaurus in the field of information storage and retrieval is a list of terms and/or other signs (or symbols) indicating relationships among these elements, provided that the following criteria hold:

(a) the **list** contains a significant proportion of non-preferred terms and/or preferred terms not used as descriptors;

(b) terminological control (in the broader sense) is intended.

B Functions of a Thesaurus

B 1 Functions of the Indexing Language

It is the purpose of an Information Storage And Retrieval (ISAR) system to retrieve "retrieval objects" according to criteria that the user specifies at the time of the search. Preferably, the retrieval objects or their substitutes are returned in a form suitable for further processing, again according to criteria specified by the user. It is the function of the indexing language to communicate to the indexer the set of criteria that are likely to be used in searching. For this purpose, all criteria to be expected in searching must be collected and brought into an orderly arrangement. This then enables the indexer to analyze the retrieval objects according to these criteria. The indexing language should therefore be constructed as a list of present or anticipated search requests or components of such search requests. This list must be used as a checklist in indexing a retrieval object so that all aspects for which the retrieval object is relevant are elicited. I call the underlying idea request-oriented indexing, and the technique implementing it the checklist technique of indexing. Obviously, the development of the checklist of criteria is of paramount importance. This is a problem of concept formation. Some examples are in order to illustrate these considerations.

(1) The reference storage and retrieval system of a company. ("A reference storage and retrieval system is an ISAR system retrieving references to documents.) In such a system it is very useful to have the following descriptors:

Technological developments that might put us out of business;

Technological developments that might be used to improve our products;

Market gaps for our products.

Indeed, these might prove to be the most important descriptors in the system. The indexer analyzing a document must compare it with each of these descriptors and make a judgment whether or not the document is relevant.

(2) An ISAR system for curriculum development. The purpose of such a system is to retrieve topics that contribute to specified educational **objectives**. The retrieval objects are therefore topics, and each topic has to be indexed by the educational objectives **it** serves. But how could this be done without drawing up a **list** of all educational objectives in the first place? As it happens, **this** list will be hierarchically structured because there are objectives, sub-objectives, sub-sub-objectives, and so on.

Again, as the indexer examines a topic, he must confront it with each of the objectives and ask the question: Does this topic contribute to the educational objective? This obviously requires a considerable amount of judgment and even didactic creativity on the part of the indexer,

(3) An ISAR system of alternative paths of action to be considered in decision making. Each alternative is indexed according to a list of criteria that has been established for deciding between alternatives. Alternatives could then be retrieved in response to a specific objective formulated in terms of these criteria. The alternatives could in turn be used as descriptors in a reference storage and retrieval system to retrieve references dealing with the alternative. Indexing in this case requires a careful analysis of the outcomes of the alternative being considered under various circumstances - a usually rather involved task.

(4) Information storage, retrieval and processing system for objects, states, and events occurring in the real world. This includes collection of data (or, to use J. D. Singer's term, "making of data") and processing these data to obtain more generalized data and/or inferences. Data collection can

be viewed as indexing the objects, states and events occurring in the real world using a checklist of criteria that are drawn from a more or less developed theoretical framework. (In data collection through observation, the observer is the "indexer", comparing objects, states, or events to a pre-established list of criteria. Often the observer is an agent for the researcher. In data collection through questionnaires, the respondent himself is the indexer, comparing his own attitudes, opinions, etc. to a pre-established set of questions.) The checklist of criteria is in fact a list of questions to be asked of the real world, and to that extent it determines the data collected. The data collected are inextricably related to the theoretical framework used. This is all the more true if the criteria used for indexing are general rather than specific. The criteria used in data collection enter into the very definition of the data. This holds even more for derived data, such as averages for population or the value of a disposition variable or index variable for an individual case. The important point is, then, that the objects, states, and events occurring in the real world are looked at from the point of view of a checklist of criteria and that the selection of these criteria should be determined by the *need* for later retrieval and processing. In the words of Karl Deutsch: "We may define theory as an information code for the storage, retrieval, and processing of new items of information, and for search for new items of information." (1969, p.22)

Data already collected can often be used for purposes other than the ones originally intended. They might be used for testing hypotheses established in a theoretical framework different from the theoretical framework used in the original data collection. For example, in a survey made for the purpose of determining optimal store locations the question might be asked whether people shop in their residential area o downtown. Later on, the set of data consisting of the

answers to that question might be used in an analysis of the relationship of people to the community in which they live and the community in which they work. However, in order that this set of data can be retrieved properly for the second researcher, it must be indexed properly in the first place. For this purpose it must he known to the indexer that somebody might be interested in doing a study on the relationship of people to the community in which they live and the indexer must be able to judge the relevance of the set of data to this problem.

Since many documents contain data that might be of relevance for testing new hypotheses, these considerations apply to the indexing of documents, too. Based on these examples we can now elaborate on the principle sketched at the beginning of this section, the principle of request-oriented indexing. This principle is implemented by the checklist technique of indexing. In the example of documents as retrieval objects; this technique works as follows: having read and understood the document (or at least what the document is about) the indexer looks at each descriptor (each criterion) in turn and decides whether the document is relevant to that descriptor (that criterion). The indexer thus acts as the user's agent, looking at the documents with the user's eyes, so to speak, and selecting relevant documents. In fact, he acts as the agent of many users, since the checklist of criteria has been constructed to include the viewpoints of many users.

Deciding whether or not a document might be useful for a researcher or other user dealing with the concept or problem expressed by a descriptor is a task that requires a good deal of judgment on the part of the indexer. Arnold Bergstraesser used the term "wissenschaftliches Vordenken" in this connection.

In order to appreciate the significance of the checklist technique of indexing, one must contrast it with the extractions and conversion method, which consists of two steps:

Step 1: Determine the important terms in the full text or in the abstract of the document. This can be done by a human indexer or by automatic indexing methods.

Step 2: Convert those terms into descriptors using the lead-in vocabulary.

This method is extremely document-oriented. In a less extreme variant, the indexer selects a number of descriptors from an indexing language until he feels that the content of the document is expressed adequately by the descriptor set. It is obvious that important descriptors like Technological developments that might put us out of business will never be used in document-oriented indexing. However, it can also be argued that request-oriented indexing with a checklist of criteria that is too narrow also fails in creating an adequate representation of the document. Both approaches to indexing should therefore be used simultaneously.

Successful application of the checklist technique of indexing hinges on the successful communication of the anticipated search criteria to the indexer. Successful communication can be achieved first of all through clarification of the individual criteria. This is obviously a task in concept formation involving lengthy negotiations of the maker of the indexing language and the potential user of the systems. Second, it is necessary to structure the set of criteria or descriptors properly as follows: The set of descriptors (criteria) is subdivided into subject fields (usually overlapping) in such a way that the indexer looking at the heading of a subject field can already decide whether or not there is an expectation that the document is relevant to any of the descriptors contained in the subject field. For example:

Education
Communication and language
Society
Politics
International politics
Law
Economics
Technology
Problems of developing nations
Socio-cultural change

In that way the indexer discards many of the subject fields, thus narrowing down considerably the number of descriptors to be looked at. Within a field the same procedure is applied: The field is sub-divided into subfields (usually overlapping) and the indexer again starts by looking at the headings of the subfields.

For example:

Politics
System of government
State and organs of the State
...
Constitution
...
Political process
Internal politics
Public administration

This leads to a polyhierarchical structure. Constitution, for example, is also narrower than Public law, a subdivision of Law. In this way, the indexer is lead to consider Constitution whether he approaches that subject from the viewpoint of Politics or from the viewpoint of Law. This structure is complemented by the introduction of numerous Related Term cross-references, so that the indexer is led almost automatically to the descriptors for which the document is relevant.

Since descriptors are, in practice, not search requests but components of search requests, the checklist technique is of equal importance in search

request formulation. By browsing through an indexing language that is properly structured and displayed, the user can clar**ify** and pinpoint more precisely his own image or concept of what it is he needs, and how he should formulate his need for best retrieval. By thinking through a field and ordering its concepts, the builder of the indexing language helps the user to think through his problem.

B 1.1 Further remarks on the functions of an indexing language in the research process.

In example (4) above the research process was viewed as indexing objects, states, and events occurring in the real world and processing the data so collected. Indexing languages for this purpose are usually called taxonomies. The variables defined in a taxonomy and their values or their operationalizations are used in formulating and testing hypotheses about the real world. Request-oriented indexing is particularly pronounced in this case. It is a well-established tenet in the theory of research methods that the real world will answer only the questions that we ask from it. The criteria used in asking questions from the real world (in collecting data to test a hypothesis) may be inspired through knowledge of the real world, but they ultimately derive from theoretical constructs. The real world and empirical data already collected alone are not sufficient to derive a complete indexing language. It is essential to draw on theoretical considerations. The following quote from Singer serves to illustrate this point further:

"Theoretical Relevance. Since one's choice of constructs will determine which variables are to be examined in seeking to explain the dependent variable phenomena, it is essential that the taxonomy reflect whatever relevant knowledge exists, in order that it not only embrace all the plausible predictor variables,

but also not exclude any reasonable candidates. To put it more programmatically, a taxonomy should serve during our entire search for the explanation of the phenomena which interest us, and should not have to be replaced each time we shift our empirical gaze from one set of independent and intervening variables to another. And if it can serve several scholars representing various research strategies, so much the better. If it does not meet this requirement, there is no adequate framework for comparing the results of a series of interdependent investigations, and we thus lose that cumulativeness which is so essential to science. In sum, one's theoretical predilections must influence one's taxonomy, and the latter can, in turn, have a profound effect on the efficiency of our inquiry." (1968.6, p.2)

If a theoretical criterion has not been considered in collecting data or indexing data already collected, then it is not possible to retrieve data relevant to the testing of a hypothesis involving this theoretical criterion, let alone to process the data as required. by the testing procedure. Instead the researcher must sift through large amounts of data and recode them or even collect his own data.

Related to these considerations is the function of hierarchy in an indexing language, especially for social science research. In order to test a hypothesis about the association of general variables, the researcher must retrieve objects, states, or events occurring in the real world that are described by the values of these variables or of more specific variables, especially variables that are used as operationalizations of or indicators for the general variables. All these relationships between general and specific variables must be included in the indexing language. Different levels of the hierarchy allow for different levels of aggregation. Again,

the hierarchical relationships depend heavily on one's theoretical viewpoint. An indexing language for general use must include hierarchical relationships from many different theoretical viewpoints. This leads to the following section on multimodel indexing.

The indexing language also has an important role in the integration of several disciplines (compare Section 1)).

B 1.2 Multimodel indexing

Multimodel indexing is a highly developed form of request-oriented indexing. It is particularly important for analyzing the role of concept formation in the building of an indexing language. The idea behind multimodel indexing is to make data usable and retrievable in different theoretical frameworks or models. This can be achieved either through using very specific concepts that can be related to more general concepts of several theories or models, or through including concepts from various theories or models and using all of them in data collection (indexing objects, states, and events) or in indexing data and documents.

For example, in the demographic section of a questionnaire one might ask for the occupation of the respondent from the following viewpoints:

level of education;

leadership function in administration, business, or society;

occupation by legal criteria;

occupation by field.

The data collected from such a questionnaire would be more useful than if a single typology of occupations (based on only one of these viewpoints or a mixed typology) is used. ("Of course, data collection would also be more expensive.) (Multimodel ndexing was proposed by Friedrich et al. 1964.3; the term is due to Karl Deutsch.)

B 2 Functions of the lead-in structure

Firstly, the lead-in structure serves to standardize terminology. Secondly, it serves to consolidate widely overlapping; concepts into newly formed ISAR concepts. And thirdly, it serves to lead from ISAR concepts not used as descriptor to the descriptors to be used, for example: Domestic trade

USE Trade and Domestic economic affairs

B 3 "User's" or "author's" vocabulary versus logical structure and request-oriented indexing as implemented through the checklist technique

It is appropriate at this point to sum up the considerations of sections B 1 and B 2. It has been stated that << in building a thesaurus, the user's vocabulary should be followed up as nearly as possible and that every term that is not contained explicitly in the user's vocabulary should be omitted even if it is necessary for logical coherence>> (Gillum 1969.10) Conversely, it has also been stated that indexing of documents should use, insofar as possible or even exclusively, the vocabulary of the author and that only terms appearing in the literature should be included in the thesaurus. I submit that these positions reflect a failure to perceive the necessity of a tool to solve the problems of communication that have been outlined in the previous sections and that are amply covered in the literature on classification. It is the task of a thesaurus to provide optimal service in indexing and retrieving documents. I do not believe that this task can be achieved by following up the user's or the author's vocabulary as nearly as possible. There is no such thing as the user, and users' viewpoints often contradict each other. There is no such thing as the author either, and different authors use different terminology, and users again use different terminology. The use that a user makes of a document is often quite different from what the author thought the document would be useful for. The indexer has to serve as the agent of all users by

indicating possible uses of each incoming document. In order that the indexer can fulfill thin role, he has to have a clear picture of what uses his clients are interested in. Were it not for these reasons, we would not need a thesaurus at all in reference storage and retrieval. (In ISAR systems for retrieval objects other than documents or other textual units, the controversy is pointless*)

B 4 Complex thesaurus structures for natural language searching and automatic indexing

In the previous sections I have argued the case that in order that a retrieval object be properly retrieved, it must be analyzed first with respect to the search requests to be expected, and that indexing is an eminently intellectual task that cannot be automated. If the retrieval objects are units of text (documents), such as transcripts of interviews or publications (possibly represented by their title and/or abstracts which are also units of text), it is possible to build a retrieval system that does not require prior indexing. Rather, such a system takes as input the units of text and the search request and performs an algorithmic comparison of each unit of text with the search request. In this comparison the terms (and possibly phrases, sentences, and groups of sentences) contained in the document are used as indicators for the relevance of the document to the concept *or* problem stated in the search request. The result is a coefficient of relevance of the document to the search request. It has sometimes been assumed that a thesaurus would not be needed in such a retrieval system. This assumption is not tenable. On the contrary, retrieval systems based on searching natural language text need very sophisticated thesauri, as I shall show in the following.

First, it is important to recognize that in all ISAR systems for documents, including those using search of natural language text, the purpose is to retrieve

documents that are relevant for a stated concept or problem. The terms used in the search request statement describe that concept or problem, and the terms occurring in the documents are used as indicators for relevance to the concept. If the search request states that documents on Attorney are sought, documents containing in their text the term Lawyer are of equal relevance, and documents containing the terms Barrister or Solicitor should also be found. This can be achieved by expanding the search request formulation to

Attorney OR Lawyer OK Barrister OR Solicitor

The use of a properly structured thesaurus will make sure that all these terms are included in the search request formulation, no matter whether the starting term is Attorney or Lawyer or .... Such a thesaurus could also be used to automate the process of expanding the search request formulation.

The thesaurus discussed so far contains essentially the same information as a thesaurus leading from non-preferred to preferred terms. In either case, classes of synonymous and quasi-synonymous terms are defined. The only difference is that in the thesaurus for natural language searching no preferred term is selected, I call a thesaurus structure that is based on the definition of classes of synonymous and quasi-synonymous terms simple in contrast to a more complex structure to be discussed next.

Putting synonymous and quasi-synonymous terms together in one class, thereby treating them as if they had exactly the same meaning, is a rather crude procedure which fails to take into account the complexities of language, the shades of meanings, the network of associations. The following example may serve to illustrate a structure that is more adequate to reflect these complexities of language.

Assume again that the search request statement as turned in by the user consists of one term only, namely, Lawyer. By expanding the search request formulation as described above, documents which contain in their text Attorney or Barrister or Solicitor are judged by the system as being of equal relevance as documents which contain in their text the term Lawyer. However, documents containing Lawyer or Attorney are probably more relevant than documents containing Barrister or Solicitor. We could quantify this judgment by assigning the relevance coefficient 1.0 to documents containing Lawyer or Attorney and the relevance coefficient .8 to documents containing Barrister or Solicitor. Furthermore, we might assign the relevance coefficient .5 to documents containing Judge. In the "simple" thesaurus we might have the cross-reference Attorney Related Term Judge, which could be used to expand the search request formulation to Lawyer OR Attorney OR Barrister OR Solicitor OR Judge. Again, in the simple system a document containing Judge is not distinguished in its relevance from a document containing Lawyer or Attorney. The information contained in a complex thesaurus structure can be represented in the form of the following table:

Term in search request	Term in, document	Relevance coefficient
1st example:		
Lawyer	Lawyer	1.0
	Attorney	1.0
	Barrister	.8
	Solicitor	.8
	Judge	.5
2nd example:		
International politics	International politics	1.0
	World politics	1.0
	Global politics	1.0
	International relations	.7
	Foreign policy	.6
	Foreign relations	.5

If the search request combines two terms, e.g., Lawyer AND International politics, the coefficients for both have to be considered to determine the degree of relevance of a document containing, say, Barrister and Foreign policy. The figure .8 in column 3 thus indicates the strength of a relationship between the terms Lawyer and Barrister. We call this relationship a "relevance relationship" because it is used in determining the coefficient of relevance of a document containing the term Barrister to a search request containing the term Lawyer. Note that these relevance relationships are often not symmetric. The problem of how these relevance relationships between terms can be determined will be discussed briefly in Section D.

Within the restrictions of natural language searching, the complex thesaurus structure offers the user the possibility of tuning his search request rather finely. If the search request formulation has the term Barrister (rather than Lawyer), the relevance coefficients of the documents (and therefore the rank ordering of the documents) will be somewhat different. The same is true for International relations versus International politics. This situation poses interesting problems of concept formation and definition and of the relationship of concepts and terms. These problems are outside my area of competence, so I shall confine myself to a short remark. Whereas in the foregoing it was assumed that concepts are somehow known and that terms are expressions of or indicators for these known concepts, the complex thesaurus structure might suggest another view. A concept can be thought of as being defined by a set of relevance relationships to the terms in the thesaurus. Since each term in the thesaurus has associated with

it a set of relevance relationships to other terms, each term has associated with it a concept in this sense. However, the user could also define a new concept by giving a set of relevance relationships that does not correspond to any term in the thesaurus.

Even though in an ISAR system using natural language searching it is not necessary to have a logical arrangement of concepts for the purpose of indexing, such an arrangement would still help the user to structure his own problem more clearly. Furthermore, such a logical structure might be helpful in the determination of relevance relationships.

The discussion of searching natural language text leads naturally to the problem of automatic indexing. In the checklist technique of indexing, indexing is viewed as a search done in advance by an indexer who can judge the relevance of documents (or other retrieval objects) to anticipated search requests or components thereof. Automatic indexing should accordingly be viewed as a search through natural language text done in advance by an algorithm that can determine the expected relevance of documents to anticipated search requests or components thereof. In such an algorithm, judgment is replaced by the identification of indicators for concepts and computing an index or coefficient of expected relevance of that document to the concept. Most commonly, the indicators used are terms; but more sophisticated algorithms also use phrases, sentences, or groups of sentences.

This view of automatic indexing (which again must be contrasted to the extraction method or the extraction—and—conversion method of automatic indexing) has major implications for the functions of a thesaurus in automatic indexing. First of all, an indexing language (a checklist of anticipated search criteria) must be built in the same way as has been described in

Section B1. Next, the thesaurus must give a set of indicators for each concept in the indexing language. (I can only mention the very thorny problem of deciding which of the relevance relations are based on the relationships between language and concepts and which are based on the hierarchical structure among the concepts themselves.)

The kind of automatic indexing described here is used in automated content analysis. The results of indexing are then further processed to produce derived data or test hypotheses.

Natural language searching and automatic indexing both have disadvantages as compared with indexing by a human indexer using the checklist technique. This should be clear from the discussion in Section B1. It should also be clear from this section what considerations one should use in order to determine just how serious the disadvantages are in a given situation. Natural language searching offers the advantage of fine tuning, which in some situations might outweigh its disadvantages. A simultaneous use of both manual indexing and natural language searching would give best performance (and is also most costly). Automatic indexing has the same disadvantages as natural language searching but does not offer the advantage of fine tuning. (In most situations it will also be cheaper than natural language searching.) A more detailed discussion would be beyond the scope of this paper.

C Steps in the Construction of indexing languages and thesauri

The standard procedure of constructing a thesaurus consists of the following steps:

Step 1: Collect and record material (terms, concepts, relationships between and among them).

The following sources, among others, are used in this step: other indexing languages/classification schemes and thesauri, tables of contents and indexes, taxonomies and list of variables used in empirical research, term association lists compiled from answers of human respondents, term association lists produced by-text analysis (all pre-arranged sources); and lists of search requests (an extremely important source), list of terms used in free indexing, abstracts (all open-ended sources).

Search requests are particularly useful if they give both a statement of the problem and the formulation of the search request; if the statement of the problem calls for the retrieval of data or documents relevant for a general variable and if the search request formulation gives the operationalizations or indicators used by the searcher for the general variable, then the appropriate relationships should be introduced in the indexing language. A similar remark holds for the statement of the problem and the formulation of the search request for searching natural language text. Such pairs are extremely valuable for deriving the relevance relations discussed in Section B4

Step 2: Sort into alphabetical order and merge information on identical terms on one card.

Step 3: Work out the preliminary structure of the thesaurus: disambiguate homonyms, identify classes of synonymous and quasi-synonymous terms -form ISAR concepts, and determine hierarchical and associative relationships between ISAR concepts (the classificatory structure).

A subtle case of consolidating widely overlapping concepts occurs when all concepts are named by the same term. Example: the term Intelligence (in Psychology) means slightly different things to different people. In our indexing language we should have a concept Intelligence (broadly defined) that includes all these meanings. (Again, this does not preclude the retaining of Intelligence 1, Intelligence 2, Intelligence 3, each carefully defined.)

If the thesaurus being built is intended for manual indexing, the thesaurus builder must select a preferred term from a class of synonymous and quasi-synonymous terms. If the thesaurus is intended for automatic indexing, the thesaurus builder must establish relevance relationships of the ISAR concepts to terms occurring in documents. If the thesaurus is intended for natural language searching, it is not necessary to establish classes of synonymous and quasi-synonymous terms. Instead, relevance relationships between all the terms must be determined - an amazingly complex task. (From considerations very similar to the discussion of the checklist techniques of indexing, it follows that term associations derived from documents are not at all a sufficient basis for determining relevance relationships.)

The major task of concept formation occurs in the structuring of the set of ISAR concepts. This task will be discussed in the following with particular reference to developing an indexing language for manual use. But structuring a set of concepts is also very important in developing a thesaurus for natural language searching or for automatic indexing as has been discussed in Section B4.

There are two interdependent principles used in structuring a set of concepts, namely semantic factoring/concept combination and hierarchy building. The following is an example of semantic factoring:

Railroad Stations = Traffic stations: Rail transport

Harbors = Traffic stations: Water transport

Airports = Traffic stations: Air transport

Bus Stations = Traffic stations: Road transport:
Passenger transport: Public transport

The new concept Traffic Stations has been formed here through semantic factoring.

This new concept can also be used to express Parking Garage as follows:

Parking Garage = Traffic stations: Road transport:
Passenger transport: Private transport.

The principle of semantic factoring is very powerful for the formation of new concepts, especially for detecting concepts that are applicable in several disciplines and that might thus be useful in a unification of knowledge from different disciplines. Examples of such concepts are:

Centralized organizational structure

Decentralized organizational structure

Precision

Reliability

Validity

Error

Input

Output

Balance, homeostasis

Imbalance

Potential, capacity

Threshold.

(One might say that general systems theory would have to be invented by a thesaurus builder if it did not already exist for him to use. A similar remark applies to decision making theory.)

In hierarchy building it is usually best to start in the set of elemental concepts derived by semantic factoring because the hierarchical relationships between compound concepts are immensely more complicated and can be derived once the hierarchical relationships between the elemental concepts are known. Hierarchy building should he search-oriented. New broader concepts useful for searching should be introduced, for example:

Broader concept to include all the following:

Relation to own culture (Culture)
Relation to other culture (Culture)
Informal education (Education)
Socialization of the individual (Sociology)
Adaptation – re-adaptation (Sociology)
Culture and personality (Social psychology)
Attitudes, opinions (Social psychology)

This is an example of a broader concept including concepts from different fields as indicated in "()".

Hierarchy building should also be oriented towards the checklist technique of indexing. As I have said before, it is necessary to arrange the concepts in a logically coherent structure to communicate effectively to the indexer what criteria are going to be used in searching. For this purpose, the hierarchical structure must be as explicit as possible. This requires that the thesaurus builder have a thorough understanding of the discipline(s) involved and be prepared to engage in the consideration and thinking through of theoretical problems in the discipline. In this process, many concepts are clarified and redefined since they are seen in context. New concepts emerge as gaps in the logical structure are seen. Often it will be necessary to create new concepts to serve as headings and thus to clarify the arrangement (organizational headings).

It is important to recognize that only a polyhierarchical structure, which allows a concept to have more than one broader concept, can adequately accommodate all hierarchical relationships necessary for searching and for the checklist technique of indexing.

An important approach to hierarchy building and a special case of semantic factoring is facet analysis. This approach is illustrated by the typology of international organizations shown in Figure 1. In facet analysis one does not try to develop a one-dimensional listing of, in this case, types of international organizations, but one examines the different dimensions of the subject and makes them explicit. Composite types can then be constructed by choosing the appropriate descriptor from each facet and combining them.

In some cases, the thesaurus maker is confronted with concepts that come from different disciplines but that share commonalities. In this case he might form a new ISAR concept as discussed above. Such concepts "should be pitched at the level of abstraction permitting them to embrace concepts that are substantially identical and whose differences are largely a consequence of the idiosyncrasies of the field in which they have been used". Such concepts would contribute both to efficiency in ISAR systems and to the "transferability of knowledge across disciplines" (Singer 1968.6,p.2.) Or a common semantic factor might be extracted; for example, from Cultural change and Social change the common semantic factor Sociocultural change could be extracted. This concept and its narrower concepts, such as Endogenous change, Exogenous change and Innovation can be used in a variety of fields, such as Ethnology, Sociology, Political Science, Administration, and Social psychology.

Figure 1: Typology of international organizations

Facet 1: International organizations by level

Private international organizations

Quasi-governmental international organizations

Governmental international organizations

Facet 2: International organizations by membership

Universal membership

SN (Scope Note)

No restrictions as to geographical location, political system, main religion, or other characteristics of member countries

Limited membership

SN Members only from one region or from, say, Islamic countries, or industrial countries

Facet 3: International organizations by scope and orientation

Covers entire range of politics

SN E.g., United Nations; International Federation of Christian Democratic Parties

Covers only specific function

SN E.g., World Health Organization; International Federation for Documentation

Facet 4: International organizations by internal cohesion

SN Basic tendency, not momentary developments

Loose groupings

Cohesive organizations

Facet 5: International organizations by organizational structure

Centralized structure

Decentralized structure

Other problems of concept formation occur if the ISAR system deals with socio-economic information from different countries. Say the ISAR system deals with information on education in the United States, France, and Germany. For each country we have a list of types of educational institutions. We want to derive one list that is applicable to all countries. This would allow us to reduce the number of descriptors. But even if we retain the original lists in the indexing language, the newly developed common list allows a searcher to search easily for, say, Elementary schools in all countries. Such a common list is also essential for the gathering of comparative educational statistics. A lot of careful work on definitions has to be done in establishing the entries in the common list, as anybody having worked in or with comparative statistics can testify. The thesaurus-builder should rely on work done by experts in comparative education in this instance.

If the thesaurus is intended for automatic indexing or natural language searching, the relevance relations should be revised based on the insights gained through the scrutinizing study of the field(s) covered by the thesaurus.

In summary we can say that the builder of an indexing language or thesaurus is charged with the task of rendering explicit and laying down on paper the structural relationships among the concepts of a field or several fields. For this purpose he applies the principles of classification theory, such as semantic factoring/concept combination, polyhierarchy, and facet analysis. It is therefore not surprising that he sometimes comes up with the formation of concepts that have not been thought of before in that form by the experts in the particular field or fields. Very often, these concepts cannot be expressed by a term in the scientific language, let alone a term

in everyday language. The thesaurus-maker is thus confronted with the additional problem of inventing proper terms for the newly created concepts.

Bibliography

Deutsch, Karl W. 1969:
On methodological problems of quantitative research.
In: Dogan, Mattei,ed. 1969; Rokkan, Stein, ed.: Quantitative ecological analysis in the social sciences. Cambridge, Mass.; MIT Press, 1969, p. 19-39.
Section B1(4).

Gillian, Terry L. 1969.10:
Comments on the TEST conventions.
In: *ASIS 1969.10: The thesaurus in action. Washington, D. C.: Department of the Army/Information Systems Office 1969 (=AD 694 590} ED 038 983), p. 7-13.
Section B3.

Friedrich, Carl J. 1964.9; Horwitz, Morton; Rothschild, M.:
Adapting the Human Relations Area Files for use by political scientists.
New Haven?: *HRAF? 1964.9, 53 p.
Section B1.2.

Singer, J; David 1968.6:
A general systems taxonomy.
Ann Arbor, Mich,: Mental Health Research Institute 1968.6. 41 *P*
Section B1.1, Section D (p.29).

Soergel, Dagobert 1971:
Dokumentation und Organisation des Wissens.- Versuch einer methodischen und theoretischen Grundlegung am Beispiel der Sozialwissenschaften.
Berlin: Duncker & Humblot, 1971, 380 p.
Preface.

Soergel, Dagobert 1974:
Indexing languages and thesauri: construction and maintenance.
Los Angeles, Calif.:Melville/New York: Wiley.
To be published in April, 1974, ca. 600 p.
(Based on: Klassifikationssysteme und Thesauri. Frankfurt: DGD 1969, 224 p.)
Preface

*Zentralarchiv fur empirische Sozialforschung, Cologne (Director: E. K. Scheuch)
I became aware of the example used in B1(4) and of the use of problem-indicator relationships as stated by the user while working in the Zentralarchiv, especially with D. Klingemann and E. Mochmann.

Soergel 1971 and Soergel 1974 contain extensive bibliographies.