Robert Fugmann

An interactive Classaurus on the PC

Fugmann, R.: **An interactive classaurus on the PC.**
Int. Classif. 17 (1990) No. 3/4, p. 133 – 137, 6 refs.

Both classification systems and thesauri have their specific strengths and weaknesses. Through properly combining both approaches one can eliminate the latter and largely preserve the strengths. "Classauri" which originate in this well-known way are most effective if they are constructed and applied during computer-aided indexing. A special variety of classaurus is described which is characterized by the employment of simple but highly effective conceptual and technical devices and by the renunciation of attempts to generate the wording of index entries algorithmically.

(Author)

1. Introduction

When searching for literature relevant to a certain concept, one must know in advance how this concept might have been expressed in the search file. All these modes of expression must be phrased as alternative search parameters, regardless of whether one uses a printed index or a mechanized information retrieval system.

In our natural, uncontrolled language *general concepts* are encountered in an unlimited variety of expressions and, hence, it is *unpredictable*, which of the *conceiveable ones* might have entered the search file (cf. Fig. 1). Either

* they are encountered there in the definition-like, *paraphrasing, "non-lexical" mode of expression*[1]

or

* a general concept comprises *so many subordinate concepts*, that they cannot possibly be enumerated as search parameters and be processed by a computer program.

An example of the first phenomenon is "fear of entering any kind of sea vessel" or "...felt a pathological fear of sea voyages..." etc., which is equivalent to the *lexical expression* "thalassophobia". A country, a continent or a landscape can be implied or expressed by so many expressions (specific towns, rivers, mountains, people, landscapes etc.), that it is again impossible to enumerate them comprehensively and to use them as alternative search parameters.

In our natural language only *individual concepts* are normally expressed in the lexical, i.e. non-paraphrasing, manner and in lingual monotony and, hence, in a fairly predictable manner.

These are the reasons for which we need a language other than the natural uncontrolled one, when we set out on a search for literature on general concepts. The most prominent feature of this "index language" must be that *it supplies modes of expression for concepts which are more predictable* than those of uncontrolled natural language. We base this claim on a more recent definition of "indexing" and "index language", in which representational predictability constitutes the core[2, 3, p.14].

Such an index language always consists of a *vocabulary*, which may be a thesaurus or a classification system. An index language *grammar*, and expecially the syntax of such a grammar, is rarely used although its employment could serve to keep the vocabulary small and transparent. Grammar could thus contribute much to the reliable use of the vocabulary and, hence, to the quality of retrieval. We shall refer to this topic later.

1.1. Strengths of classification systems

It is inherent in classification systems that they use systematic notations as descriptors (cf. Figure 1). Similar notations are thus assigned to related concepts. Hence, the phrasing of a general query, which is to include all the subordinate concepts of a concept in question (e.g. more than one million species of insect) is facilitated or even made possible.

Furthermore, it is possible in classification systems to coin descriptors for important concepts, namely notations, long before a corresponding lexical expression emerges in natural language. In this respect, classification systems can be more up todate than a thesaurus. For example, we know several causes for the slipperiness of a road. But besides "aquaplaning" there are so far no lexical expressions such as "oilplaning", "leafplaning" or "clayplaning". In a classification system such descriptors could easily be introduced.

Classification systems, especially the faceted ones, are conceptually transparent and easy to view because they are *categorized* according to a set of predetermined, conceptual categories. Text analysis is thus rendered more reliable (predictable), and a useful guideline for the factoring of composite concepts is provided, apart from several other significant advantages.

Furthermore, in a classification scheme the concepts that are subordinate to a more general concept can be grouped most lucidly according to which *character of subdivision* is reflected in them. Characters of subdivision render a vocabulary particularly transparent and thus facilitate (or even make possible) the search for the

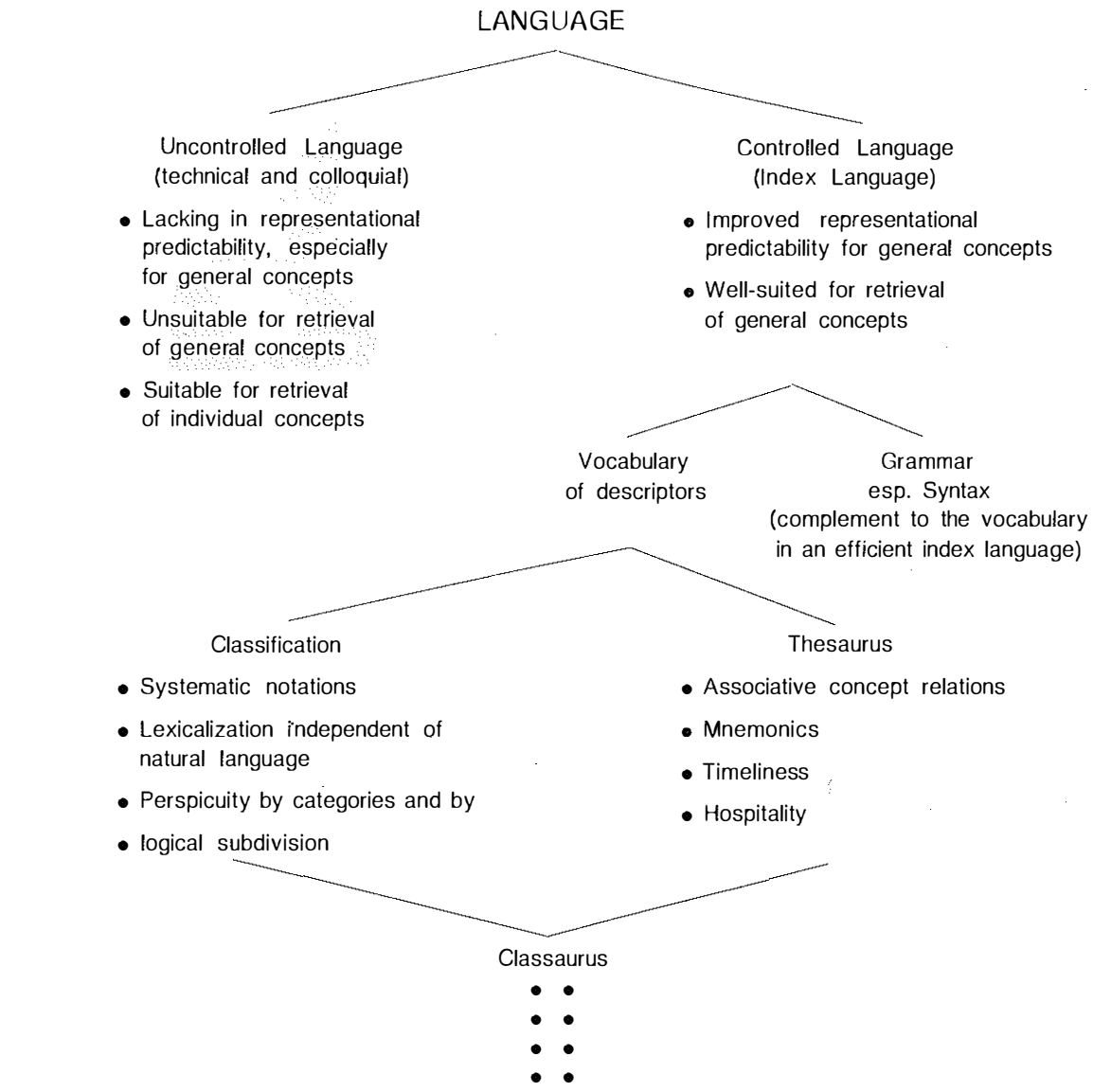

Fig. 1: Typical features of various kinds of language

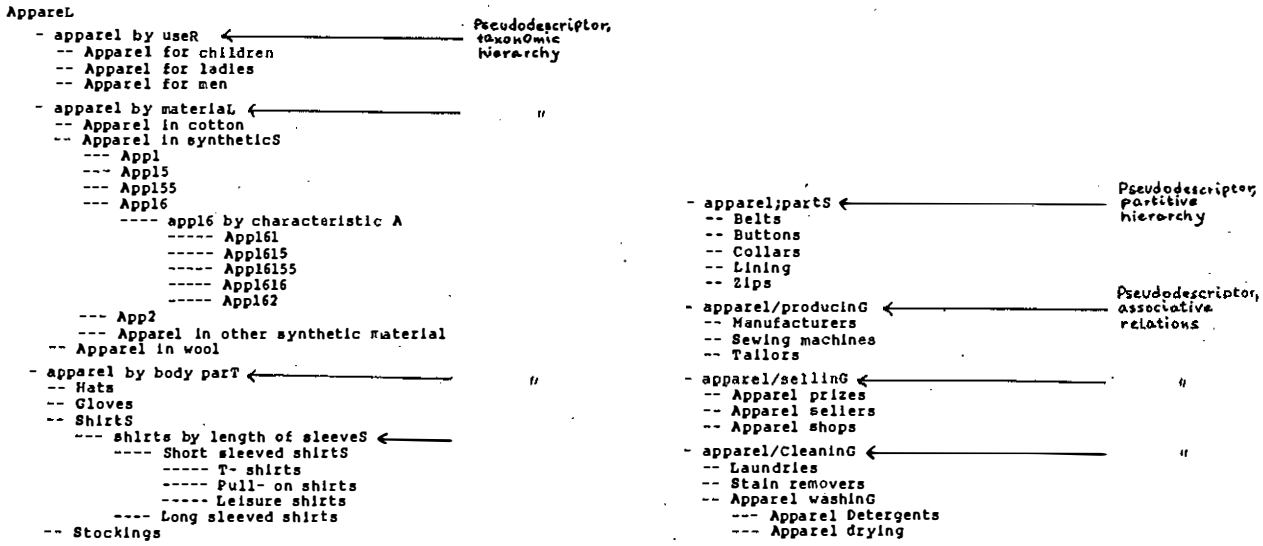

Fig. 2: Model hierarchy with taxonomic, partitive, and associative relations

most appropriate descriptor, which should always be performed by the indexers and questioners.

Vocabulary transparency (and the indexer's competence in the subject field) is crucial to reliable indexing and, hence, to the quality of the searches. If, for example, we are informed of a case of "illegitimate use of a pharmaceutical of the roborantia group with the aim of winning physical superiority in sport competition", then the descriptor "doping" is more appropriate than the descriptors "superiority", "sport" or "competition",

although just these descriptors are even suggested by the wording of the text.

It is the task of an orderly, systematic arrangement of the vocabulary to lead the indexer and searcher from less appropriate, less specific descriptors to the most appropriate ones, even if the less appropriate ones may be suggested by the wording of the original text. In other words,

any successful search in a file of documents is preceded by another search, namely by one in the file of descriptors. Any failure in or even omission of the search on this level will inevitably impair the quality of the subsequent search on the document level.

1.2 Strengths of thesauri

One of the strengths of a thesaurus is the possibility of expressing the entire set of the *associative concept relations* and not merely the hierachical relations.

Furthermore, an indexer or inquirer is more familiar with the *natural-language descriptors* of a thesaurus than with the notations of a classification system, the meaning of which must in most cases be looked up.

In those cases in which natural language has already coined a term for a new and important concept, it can immediately be adopted in the vocabulary, and no special decision about the phrasing of such a descriptor must be agreed upon, in contrast to classification systems.

This possibility, however, should only be used with caution. It is the *precombining* descriptors, i.e. descriptors which comprise the meaning of two or more descriptors, which are often too readily introduced into the vocabulary. The transparency of the vocabulary will decline concomitantly with the proliferation of these descriptors, as will the reliable use of the vocabulary by the indexers.

It should be kept in mind that any vocabulary and a classaurus, too, constitutes only *half of an index language*. If an information system is large and is incessantly growing with respect to file size and use frequency, then an index language *grammar*, especially the syntax of such a grammar, should be available. This would largely eliminate the necessity of introducing descriptors of the composite type and help to preserve the survival power of the information system.

This enumeration of the essential features of both classification system and thesaurus shows that they can well *complement each other* while eliminating their typical weaknesses. Proposals to this end have been submitted and in part already implemented. Examples are "Thesaurofacet"[4] and "Classaurus"[5,6]. A specific variant of classaurus is described in the following. It is restricted to simple but particularly effective conceptual devices and it utilizes an efficient program for an efficient personal computer.

2. A new variant of classaurus

It is specific to our variant of classaurus that

* it is grouped by conceptual categories,
* it works with characteristics of subdivision and these appear in the vocabulary as "pseudo-descriptors",
* it contains systematic notations at those locations in which this is desirable,
* it permits the presentation and mechanized usage of any kind of associative relations.

A section of a model classaurus of this kind is depicted in Figure 2.

These features make possible "mechanized grouping" of related concepts even in those cases in which the corresponding descriptors have no string of characters in common, as is usual in natural language. No "manipulative grouping" is necessary, i.e. the (sometimes very elaborate or even impossible) exhaustive and explicit enumeration of all the subordinate descriptors under a more general descriptor for a concept of the inquiry. Any kind of relationship, the partitive and associative included, can thus be expressed and used as a search parameter.

To consider a characteristic of subdivision as a kind of descriptor is justified because such a pseudo-descriptor in fact represents a general feature of all its subordinate concepts, namely *the kind of difference* prevailing among them. For example, "apparel by user" collects all concepts in which some specific statement about the kind of user is expressed or implied.

The kind of notation depicted in Figure 2 permits the insertion of any number of concepts at any location in an array[7]. This prevents hierarchies from becoming chaotic in the course of time, when new concepts have to be inserted and when no appropriate location is available forthem (the end of the array is not always the most appropriate location). This notational device is similar to that used in Dewey's Decimal Classification and is easily understood if the notations proposed for this purpose are considered as decimal numbers with the preceding (and here omitted) string "0." Through the recurrence of a common string of characters in all conceptually related decriptors they are also susceptible to mechanized grouping.

Practically unlimited hospitality in chain (i.e. hospitality with respect to the number of hierarchies available) should be provided by the capability of the software. In the case of LIDOS, up to 30 hierarchies can be used. Within these limits any number of hierarchies can be inserted at any location in the hierarchy. Again, this is a valuable device for preserving a meaningful systematic structure of the vocabulary over the course of time.

Analogously to the characters of subdivision, any other kind of relationship can be phrased as a pseudo-

descriptor. This makes it possible to represent these relationships in a predictable way and, hence, in a way useful for retrieval and for mechanized grouping.

In contrast to other, similar classaurus approaches it is not intended to generate the wording of index entries algorithmically. Rather, they are phrased intellectually, and descriptors are assigned to them so that they will occur at all locations in the index where they can be expected.

3. The interactive use of the classaurus

The possibility of interactively working with such a classaurus facilitates

* its construction,
* descriptor choice for indexing and retrieval,
* presentation of the vocabulary in print.

The assistance provided by a computerized classaurus is greatest in the construction of a book index. Here, simultaneously with the progress of indexing, the vocabulary of subject headings must be established from scratch. Its structure and contents are constantly subject to change, depending on the concepts which are newly encountered or which with hindsight prove important enough to be represented in the vocabulary, and also depending on the emerging desirability of changing the characters of subdivision and, hence, the arrangement of subject headings. All these procedures can be performed interactively and with relatively little expenditure of time and effort with the aid of a good computer program. Some of the desirable manipulations even *require* this computer assistance.

3.1 The compilation of the vocabulary

In the first step of scrutinizing a book for the first time, those *concepts are collected* which might be of interest for a searcher. They are phrased as subject headings for the index. In the catalogue of a department store this might apply, for example, to "shoes", "watches", "bicycles" etc.

Sufficiently *concise and expressive terms* will not always immediately come to mind when one encounters *paraphrasing expressions* for a concept in a text, i.e. expressions which are not suitable for inclusion in an index, due to their non-lexical (1) nature. For example, in one passage a "spray for combatting plantlice on indoor plants" may be mentioned. Later, one becomes aware of expressions such as "insecticides" and "plant protection" or still more appropriate lexical expressions and would assign these subject headings to this concept *when it recurs* and *when its importance has become more obvious*.

Only in a second perusal of the text, *after* the vocabulary has largely been established, can one go about *consistently assigning the subject headings* of the index to *all* passages, to which they apply. This is true particulary for the early passages, read in the first stage, when the vocabulary was still largely incomplete. For example, it is only in this second stage, that the above-mentioned passage would receive the subject headings "insecticides" and "plant protection".

Hence, subject indexing by no means consists merely of extracting meaningful words from the texts, as is often erroneously assumed.

3.2 Structuring the vocabulary

In the first step of scrutinizing a book for the first time, the indexer also obtains an overview of the *conceptual categories* according to which the subject headings should be systematically grouped. During the progress of indexing such an overview will soon become necessary in order to insert the new accessions of subject headings correctly, to avoid unintentionally introducing synonymous ones and in order to have an overview of the concepts that are not yet (or should not be) represented in the vocabulary. Only after the indexer has attained a good overview of the entire contents of a book, will he or she be able to make the best choice of categories. *Frequent re-grouping* of the subject headings, i.e. frequent shifting of subject headings (see below) within their systematic arrangement, will therefore be necessary.

3.3 Shifting of concepts

Additional vocabulary transparency can be achieved through the introduction of *characteristics of subdivision*. Here, too, it will only be after some experimentation and after considerable revision of the work already performed that one arrives at the best choice, so that the class of residual concepts, which could not yet be located under a characteristic of subdivision, has become as small as possible and so that the characteristics of subdivision can be named in a lingually acceptable form. Here again, the shifting of subject headings to more and more appropriate locations is necessary.

In the *categorization* of articles in a catalogue of a department store, subject heading shifting will often become necessary for still another reason. For example, one may have begun with the categories APPAREL, TOOLS, JEWELLERY, FURNITURE. Later on, one finds that there is no place for lawn mowers, bicycles and books. When one introduces corresponding categories, one will have to check whether some of the already established subject headings ought to be located in the newly established categories. For example, a book on tools for carpentry, originally located among TOOLS, will then, at least additionally, have to be located in the category BOOKS, as well.

In such a shift, not only the postings of the subject heading in question, but also all subordinate subject headings, together with all their appertaining postings, will have to be transferred. The machine program must perform these shifts without requiring the indexer to enumerate all the corresponding subordinate subject headings and all the appertaining postings.

Since only trial and error will lead to a satisfactory solution as to the best choice both of categories and of the characters of subdivision, one will frequently have to re-organize the vocabulary by repeated and extensive shifting.

3.4 Subdividing general subject headings with hindsight

In the initial stage of indexing textiles in a catalogue one may have contented oneself with fairly general subject headings, for example with "synthetics". Later on, it becomes apparent that the various kinds of synthetics (polyester, polyurethane, acrylics etc.) should be differentiated. This is demonstrated in Fig. 2 by the notations "App1", App15" etc.. However, one of the newly introduced subject headings may be implied by other subject headings already in use. For example, Spandex™ always means textiles in polyurethane. Then it is desirable to retrieve all the postings under this brand and to assign *to this group of postings* in one single step the subject heading "polyurethanes", too.

3.5 Collecting specific subject headings under a more general one

During the development of the vocabulary the indexer may also perceive that it will eventually become too large and specific in certain sections. For example, initially he may have differentiated the various kinds of bicycle (mountain, racing, children's bicycles etc.). He may then decide to bring them all together under the more general subject heading "bicycles". In this case, he will retrieve all the specific postings with one single query, assign them the more general subject heading in a single step and finally erase the more specific headings in the vocabulary.

3.6 Lead-in terms

During indexing, one will encounter again and again new lexical expressions for a concept that has already been represented by a subject heading with different wording. One common example is acronyms or near synonyms i.e. expressions for concepts so closely related to another one that they need not be differentiated in the index. In the alphabetical part of the index all these synonymous terms should be listed and lead the searcher to the preferred subject heading (or be directly available as search parameters in a mechanized index).

4. Printing the index

It is much in the interest of good indexing that the latest stage of vocabulary development should be available also in print. Display on the screen normally lacks the degree of transparency which is desirable for indexing work. This holds true for the alphabetical as well as for the systematic version of the vocabulary. In the alphabetical version, for each subject heading at least its immediately superordinate one should be displayed for better navigation.

5. Conclusion

Conventional, approved conceptual devices such as categorization and logical subdivision can effectively be used with the aid of simple computer technology. This facilitates and accelerates the look-up procedure in an index, renders the intellectual indexing procedure more reliable and thus drastically improves retrieval, regardless of whether it takes place in a printed or in a computerized index. The mere extraction of meaningful words from the texts cannot lead to good subject indexing

because it has not been shown so far that the equivalence of paraphrasing expressions with their corresponding subject heading or descriptor can be algorithmically recognized.

Acknowledgement: The author thanks Software Land for their prompt advice in the author's adaptation of LIDOS™ to the task described here.

Notes and references:

(1) *Lexical* expressions are deemed those which are commonly agreed upon and in which the sequence of the characters (spaces included) is fixed. They can therefore easily be located in an alphabetical arrangement of words. *Non-lexical* expressions are consequently those which lack this property.

(2) FID/Classification Research, News No. 2: "By 'indexing' is meant the description of the essential contents of a document, by extraction and/or assignment of significant terms with or without syntactical relationships with a sufficient degree of fidelity and predictability for retrieval demands". Int. Classification 8 (1981) p. 96.

(3) Fugmann, R.: "Indexing is the translation of the essence of a document into an indexing-lingual mode of expression." and "The task of an indexing language is to represent concepts and statements with a sufficient degree of predictability and fidelity." Int. Classification 6(1979)p.3-15.

(4) Aitchison, J.; Gomersall, R.; Ireland, R.: Thesaurofacet: a thesaurus and faceted classification for engineering and related subjects. English Electrical Company Ltd.; Whetstone, Leicester, England. Cited from Aitchison, J.; Gilchrist, A.: Thesaurus construction. A practical manual. ASLIB 1987.

(5) Bhattacharyya, G.: Classaurus: Its fundamentals, design, and use. Studien zur Klassifikation 11 (1982), p. 139-148. Frankfurt: INDEKS Verlag.

(6) Devadason, F. J.: Online Construction of Alphabetical Thesaurus: A Vocabulary Control and Indexing Tool. Inform. Process. and Management 21(1985),p.11-26.

(7) Descriptors with a capital letter at their end indicate on the screen that they are generic to other descriptors.

Dr. R. Fugmann, Alte Poststraße 13, D 6270 Idstein