LANCASTER, Wilfred: *Vocabulary Control for Information Retrieval*. Washington D. C.: Information Resources Press 1972. 233 p., 74 figs., 58 tables, chapter bibliographies, index. $ 17.50

This review is in two parts. Part 1 gives an assessment of the book as a whole and discusses some basic issues raised in it. Part 2 is a more detailed account and discussion of the contents of the book.

Part I – *General comments*

"This book deals with properties of vocabularies for indexing and searching document collections; the construction, organization, display, and maintenance of these vocabularies; and the vocabulary as a factor affecting the performance of retrieval systems." "I have attempted to cover the entire range of vocabulary possibilities, from highly structured lists to free text (no control), and to point out advantages and limitations of various approaches. The book was specifically compiled as a text to accompany a course in vocabulary control presented at the Graduate School of Library Science, University of Illinois." (From the Preface)

It seems useful to restate the topics dealt with in the book in a list arranged by decreasing emphasis:

Conceptual structure of vocabularies (approx. 85 p.)
Use and function of vocabularies in reference (document) storage and retrieval systems and effects of the vocabulary on system performance (approx. 70 p.)
Presentation and display of vocabularies (approx. 45 p.)
Generation and updating of vocabularies (approx. 30 p.)

The book provides many useful insights on these topics presented in a lucid style. It also provides a wealth of examples in the form of sample pages of existing vocabularies, excerpts and/or summaries of the rules and procedures used in building existing vocabularies, including forms used, concise, clear and informative descriptions of various systems of particular interest and summaries of the results of relevant papers. The book serves a useful function in bringing together all these materials. Several chapters, especially those having to do with generating and organizing vocabularies, consist mainly of relevant examples, which can be useful to the reader wishing to perform his or her own analysis and to derive his or her own principles and generalizations.

The book develops concepts gradually, moving back and forth between discussion of the structure and presentation of vocabularies and discussion of their use in reference storage and retrieval systems. This makes it easy for the reader to get familiar with the subject. But it also leads to some duplication and in some cases a more rigorous organization would have been beneficial. For example, Cutter's criteria for the selection of preferred synonyms are given in Chapter 4, *Vocabulary control by subject heading*. The criteria used in selecting terms for the Thesaurus of Engineering and Scientific Terms (TEST) are given in Chapter 6, *Generating the controlled vocabulary*, those used in SMART are quoted in Chapter 16 *Searching natural language data bases.* None of these places is accessible through the index under either Criteria or Selection or Term selection. The treatment of faceted classification/facet analysis and of synonym control, to give two other examples, is scattered over more places than necessary, but all these places are accessible through the alphabetical index.

This reviewer differs from Lancaster in his view of the role of indexing languages in information storage and retrieval systems. Lancaster states "However, the controlled vocabulary does not, or at least should not, influence the conceptual analysis of documents and the conceptual analysis of requests. The conceptual analysis stage is separate from the translation stage." (p. 2) This reviewer could not disagree more. Indexing must be viewed as a relevance judgment by the indexer with respect to an interest profile, an anticipated search request or a number of anticipated search requests. The indexer acts as the user's agent in scanning the literature and deciding which of that literature is relevant for a certain request. In order to serve in this role the indexer must know what the user's interests are. The indexing language is a communication device that should communicate to the indexer the interests of a group of users. The indexing language thus provides the framework within which the indexer analyses documents. Thus the indexing language is central in the conceptual analysis of documents and not only in the translation stage, (request-oriented indexing) Lancaster is very well aware of this when he says: "It is more important that the *warrant* of a vocabulary be established by the language of requesters than by the language of documents." (p. 32, emphasis in original).

This basic point provides a proper perspective for reviewing Lancaster's assessment of the use of natural language as an indexing language. In Chapter 16, *Searching natural language data bases*, Lancaster argues in favor of less sophisticated reference storage and retrieval systems, using only free terms (either all terms from the title and/or an abstract, or terms picked by an indexer from the title and/or an abstract and possibly some free terms added by the indexer) or free terms and very broad subject descriptors assigned by the indexer from a very small controlled vocabulary. Such systems, Lancaster says, work effectively if search requests are formulated appropriately, especially with regard to inclusion of all synonymous terms designating a search concept i. e., vocabulary control in searching. If there is any difference in performance as compared to systems using a sophisticated controlled vocabulary , such differences can be made small. This assessment is based largely on Cranfield 2 and a similar study by INSPEC. Both studies and the recommendations of the INSPEC study are reported uncritically by Lancaster. However, the INSPEC study, like the Cranfield study and many others of its kind, proves nothing with respect to the principle of request-oriented indexing set forth above. This is so because in these studies the controlled vocabulary was *not* used to give the indexers a frame of reference for analyzing documents. To the contrary, the indexers assigned free terms suggested by the documents (document-oriented indexing), and these terms were then translated into a controlled vocabulary. At least in the Cranfield text the controlled vocabulary was built *after* the indexing of the test documents was completed. This is to say that these studies do not give a reliable assessment of the difference in performance between systems using controlled and uncontrolled

https://doi.org/10.5771/0943-7444-1975-1-51 https://www.inlibra.com/de/page/agb - Open Access - (cc) BY

vocabularies, respectively. It is *not* to say that a controlled vocabulary is preferable in all situations.

The recommendations of the INSPEC investigators in favor of an uncontrolled vocabulary in their system are all the less understandable in that they mention among the disadvantages of a controlled language: "It requires the setting up and continuous development of a thesaurus for use in indexing; the updated thesaurus must be available to each indexer." (quoted in Lancaster, p. 149). However, the use of natural language as the indexing language does not obviate the need for a thesaurus. On the contrary, as Lancaster states, "the searcher who most needs a thesaurus is the searcher in a natural language system". Whereas the reader of Chapter 16 is left with the impression that Lancaster recommends natural language searching, possibly amended by the use of a very broad controlled vocabulary, Chapter 24 on cost-effectiveness brings the matter into a somewhat different perspective. There an important practical consideration, mentioned only in passing in Chapter 16, is fully brought out: Natural language searching places an immense burden on the searcher, especially in on-line systems. Particularly in on-line systems it also requires larger files and more computer time in processing search requests. This reviewer would surmise that in a cost-benefit analysis of a heavily used reference storage and retrieval system, a controlled vocabulary would come out ahead for these reasons alone.

Part 2 – *Detailed comments*

The following more detailed account of the contents of the book brings together chapters on the same topic area rather than following the sequence of chapters of the book which, for good reasons, is different.

Chapters 1 to 5 provide an *introduction* on a very general level. Chapter 1, an "Introduction to the Introduction" discusses the why of vocabulary control (p. 1–5). All the topics mentioned in it are taken up in more detail later on in the book. Chapter 2 gives an excellent discussion of pre-coordination and post-coordination and of enumeration and synthesis (p. 5–7). The discussion makes very clear that these are two different dimensions in the analysis of the structure of vocabularies, whereas frequently the two are assumed to be the same or at least correlated in the sense that pre-coordinate systems are enumerative and post-coordinate systems synthetic. Chapters 3–5 discuss Classification scheme, Subject heading list, and Thesaurus as three different concepts (p. 8–26). However, the underlying commonality is recognized implicitly. For example: in Chapter 3 on classification it is mentioned that the concepts listed in a faceted classification are "highly suitable for use in a post-coordinate system" (p. 12), e. g., in a peek-a-boo system. (Thus one should either eliminate the definitional link between "thesaurus" and post-coordinate systems, as this reviewer prefers, or else call a faceted scheme a thesaurus if it is used in a post-coordinateway.) On the difference between subject heading lists and thesauri, Lancaster has this to say: "the principal difference between the two tools lies in their mode of application. The subject heading stands alone in the alphabetical subject catalog, whereas the descriptor is used in conjunction with other descriptors even though in itself it may already be highly pre-coordinated . . .". This reviewer does not believe that usage should enter into the definition of various types of subject access vocabularies, especially since many vocabularies are used in both ways (the National Library of Medicine's Medical Subject Headings is a prime example). Lancaster goes on to say "However, the thesaurus will include terms that one would never find in a list of subject headings. These terms are somewhat meaningless on their own and would never stand alone, but they are useful when used in conjunction with other descriptors to narrow the scope of a class. Examples are general property terms such as FINE or FINENESS, general characteristic terms such as EFFICIENCY and general process terms such as TESTING or REFINING." Of course, many such terms occur in lists of common subheadings that logically form part of subject heading lists.

Conceptual Structure and Presentation/Display of Vocabularies are treated in the following chapters:

Chapter 7 *Organizing and displaying the vocabulary* (with emphasis on overall conceptual structure and display of that structure through a linear sequence with indentions expressing hierarchy or by graphical means) (p. 38–65).

Chapter 10 *The Reference structure of the thesaurus* (with emphasis on presenting the conceptual structure through crossreferences given for each term in the main part of the thesaurus) (p. 77–89).

Chapter 14 and 15 *Characteristics and Components of an index language* (p. 115–134). This is divided into a discussion of the vocabulary and of so-called "auxiliary devices" which include synonym control, hierarchical structure, links, and role indicators, among others. These two chapters are intended to give a more formal definition of an index language and its structure, but they do not quite achieve this aim.

Chapter 18 *Compatibility and convertibility of vocabularies* (p. 161–176) (in part).

Chapter 21 *Supplementary vocabulary tools* (p. 191–203). This rather mistitled chapter deals with a variety of topics as follows:

1. Display of small vocabularies à la Mooers
2. Document analysis sheets on which all or part of the index language is preprinted to save clerical work in indexing.
3. In-house tools such as the integrated authority file or special documents containing indexing instructions for specific topical areas maintained in the National Library of Medicine.
4. MEDLARS "hedges", which, as Lancaster mentions, are nothing else but general concepts created for searching.
5. The Medical Subject Headings tree structures.

Most of these topics would fit better in other places in the book, especially in Chapter 7. Most of the tools discussed are central rather than supplementary tools. Furthermore, a comment on the MSH Integrated Authority File and the "Hedges" is in order. Lancaster makes the point that these tools "are intended primarily to aid the indexer but they should also be of value

https://doi.org/10.5771/0943-7444-1975-1-51 https://www.inlibra.com/de/page/agb - Open Access - (cc) BY

in the conduct of searches". This reviewer would like to make the point more forcefully. These tools are just as important, or perhaps even more so, for searching as for indexing. To the extent possible, they should therefore be integrated into the version of the vocabulary available to the public; for example, "hedges" could simply be added as new concepts to the vocabulary. Of course, the appropriate hierarchical relationships would have to be added as well. Such an integration would lead to a much more orderly display of the total vocabulary used in indexing and searching operations. This reviewer feels that in a book on vocabulary control, it would be appropriate to discuss such desiderata rather than merely describe the status quo.

The following chapters are devoted entirely to examples illustrating the conceptual structure and the representation/display of vocabularies:

Chapter 8 – *The Thesauro-facet* (p. 66–69).
Chapter 9 – *Some thesaural rules and conventions* (p. 70–76).
Chapter 19 – *Some further controlled vocabularies* (p. 177–184)

(The appendix lists further controlled vocabularies for study and examination.)

The main chapter on *thesaurus building* is Chapter 6 *Generating the controlled vocabulary* (p. 27–37). The process of thesaurus construction is illustrated through a number of examples. This reviewer does not agree with the sequence of major steps given at the beginning of Chapter 6. This sequence places selection of terms and deciding on the exact form of the terms before developing the conceptual structure and displaying that structure. However, meaningful selection decisions can only be made in the framework of an overall structure which must be shown in a preliminary display. Deciding upon the exact form of a term (e. g., singular or plural) is a more or less clerical matter and should be performed towards the end of the process when most other problems are solved. Further remarks on thesaurus construction can be found in Chapter 11 *Computer manipulation of thesaurus data* (p. 90–91); Chapter 17 *Creating index languages automatically* (153–160); and Chapter 23 *Vocabulary in the on-line retrieval situation* (211–217). Each of these three Chapters also deals with problems of thesaurus use. Chapter 11, for example, discusses the use of a machine-stored thesaurus for validating descriptors used in indexing, Chapter 17 deals also with automatic indexing, and Chapter 23 deals also with the use of a vocabulary in an on-line retrieval system. This organization tends to confuse two problems, namely that of construction of an index language or thesaurus and that of application and use of that index language or thesaurus. It is not at all necessary and not always advisable to use on-line methods for constructing a thesaurus just because this thesaurus will be used in an on-line retrieval system. Nor is it necessary or always advisable to use machine methods in manipulating thesaurus data just because the thesaurus will be used in a mechanized information storage and retrieval system. It would be much clearer to base the organization of material into Chapters on the functions, namely, thesaurus construction versus thesaurus application and use, and then within each function discuss the various techniques, such as batch processing or on-line computer use. Of course, there are interrelationships, such as on-line updating of a thesaurus by a searcher while he is using this thesaurus for retrieving in an on-line mode, but these relationships could be pointed out at appropriate places. It is also true that different retrieval techniques may pose different requirements for a thesaurus. But agian, these requirements do not by themselves prescribe the methods by which the thesaurus could be constructed.

Thesaurus updating, an extension of thesaurus building over time, is dealt with in Chapter 12 *Vocabulary growth and updating* (p. 98–106). However, this chapter concentrates on presenting statistics on the growth of various vocabularies and says very little on updating procedures.

The following chapters deal with the *Application and use of indexing languages and thesauri* in reference storage and retrieval systems and the effect of the vocabulary on systems performance: Chapter 13 *The influence of system vocabulary on the performance of a retrieval system* (p. 107–114). Figure 51 of this chapter gives the common and dangerous oversimplification: specific vocabulary – high precision, low recall, and non-specific vocabulary – high recall, low precision. This is all the less understandable as the text points out that in a system using a specific vocabulary high recall can always be attained through proper search requests formulation. Whereas it is true that "we can not compensate for lack of specificity in order to improve precision" (p. 112) if the search request is specific, in a system where most search requests are fairly general a non-specific vocabulary will not lead to low precision. This reviewer also wants to take exception to the following statement: "The size of the classes defined by a vocabulary is much more important than the arrangement of the classes." (p. 114) Whereas it is obviously true that the arrangement of the descriptors in the vocabulary does not affect retrieval once the indexing is done and the search requests are formulated, the arrangement of the descriptors might well be a very important factor in determining the quality of indexing and of search request formulation and thus in overall systems performance.

Chapter 16 *Searching natural language data base* (p. 135–152) (see the discussion in part 1).

Chapter 20 *The role of the controlled vocabulary in indexing and searching operations* (p. 185–190).

Chapter 22 *Vocabulary use and dynamics in a very large information system* (p. 204–210). This is a good summary of the report "Structure and uses of vocabulary in MEDLARS II by Ahrley, A. J. and Lancaster, F. W., Silver Spring, Maryland; Computer Science Corporation, 1969.

Chapter 24 *Some cost effectiveness aspects of vocabulary control* (p. 218–222). This is one of the best chapters in the book. In particular, it brings the considerations of natural language searching in proper perspective as was discussed in part 1. As mentioned before, Chapters 11, 17 and 23 deal in part with vocabulary application.

Chapter 18 *Compatibility and convertibility of vocabularies* (p. 161–176) (in part).

Chapter 25 *Synopsis* (p. 223–225) gives, on little more

https://doi.org/10.5771/0943-7444-1975-1-51 https://www.inlibra.com/de/page/agb - Open Access - (cc) BY

than two pages, a very useful list of the "most salient points on vocabulary control that this book has attempted to illustrate" (p. 123), cross-referred to the appropriate chapter(s).

The critical comments on individual points should not detract from the intrinsic value of the book. Based on his vast experience the author provides a useful overview of the structure and display of vocabularies, of methods for their construction, and of experiments that were intended to clarify the role of the vocabulary in the performance of reference storage and retrieval systems. Although this review has disputed a number of points, the author presents tenets that are widely accepted. In this sense the critical comments in this review reflect really a controversy in the field. In summary, the book is a singificant contribution to the literature of information science. *Dagobert Soergel*

SOERGEL, Dagobert: *Indexing Languages and Thesauri: Construction & Maintenance.* Los Angeles: Melville 1974. XXXIX, 632 p. ISBN 0-471-81047-9, A Wiley-Becker & Hayes Series Book.

This volume deals with the characteristics and construction of controlled vocabularies. It is very complete and, by and large, extremely accurate. The contents are divided into four major areas: the structure of indexing languages, methods by which such vocabularies are arranged and presented, methods by which they may be constructed and maintained, and the use of thesauri as the basis of cooperation among information services. A novel feature of this book is that it presents the material at various clearly defined levels. A reader who wants only a general understanding of indexing langnages need read only designated sections of the work. Other sections are marked as "technical", "special" or "advanced". These need be read only by those who want a deeper understanding of the subject or who wish to extract information relating to a special problem area. The way the volume is structured, then, makes it more suitable for use as a handbook – a volume to consult when we need to find out about a particular aspect of vocabulary control – than as a textbook or as a series of chapters to be read consecutively. Soergel, however, would like to think of it as both a handbook and a textbook.

Viewed as a handbook, the work is excellent. I find myself in complete agreement with much that the author says. There is a great deal of common sense here, and the author strips away the unnecessary mystique that surrounds much of the other writing in this area. He is insistent, and rightly so, that an effective controlled vocabulary must be built around the special needs of the user group it is to serve. Consequently, the maker of a controlled vocabulary must learn as much as he can about the characteristics of this user group, especially the types of requests they are likely to make to the system. I support these sentiments fully. Elsewhere I have said that "user warrant" is even more important than "bibliographic warrant" in the construction of an indexing language.

Soergel is a careful writer. In particular, he is careful to define all the terms that he uses. Some may consider him too careful, that there is too much definition, and that some of this is hair splitting. I do feel that terms should be carefully defined but I also feel that the author goes overboard on occasions. Sometimes I find myself thinking "all this precise definition is fine, but let's get on with the discussion". He also introduces new terms for familiar concepts when he feels that the "old" terms are inappropriate. For example, he prefers the terms "precombination" and "postcombination" to "precoordinate" and "postcoordinate". Again, I find myself mostly in agreement with his terminology, but I am not always certain that the new terms he introduces are an improvement on the ones they replace. Sometimes his choice of terminology is unfortunate I feel (e. g., "quasi-synonym" used for "near synonym").

Soergel's book is mostly well arranged, although on occasions he introduces terms that he has not yet defined. For example, on page 6 he introduces the terms "precombination" and "postcombination" long before these terms have been explained. It is, of course, difficult to maintain an optimum sequence in a work of this type and minor blemishes of this kind can be forgiven. More annoying is the fact that the proofreading of the text leaves a lot to be desired. For example, on page 20 the word "in" appears twice in place of the correct "ion" and on page 22 "lightning" is listed as a synonym of "illumation". While such errors should be obvious to the reader, it is unfortunate that an author who is so careful in his definitions should allow typographical errors of this kind to creep in.

There are some other defects that I would like to point out. One of these is the tendency of the author to make sweeping, authoritative assertions without in any way justifying them. For example, he says categorically that "the higher the degree of mechanization of an ISAR system, the greater the need for a good thesaurus that indicates conceptual relationships". I am not at all sure that this is true. At least, I cannot accept a statement of this kind without some justification being given. But such justification is lacking in the text. Let me quote one more example. The author states that, in determining the appropriate level of exhaustivity of indexing, and specificity of vocabulary, important factors to be considered include amount of time available to do a search and expected frequency of search requests. Why are these important? It is not at all obvious, at least to me.

Another criticism I have relates to the incomplete treatment accorded to certain topics. On page 9, for example, Soergel lists three "criteria for the evaluation of a thesaurus", namely degree of conceptual completeness, degree of terminological completeness, and quality of the display. Although he is well aware of the importance of specificity of the vocabulary, Soergel makes no mention here of this evaluation criterion, which is an extremely important one. Such omissions are dangerous in a handbook that is not necessarily read in toto.

Very occasionally the text is inaccurate. His statement on page 56, for instance, that roles and links cannot be used with peek-a-boo cards is just not true. It is difficult but it can be done.

I have deliberately looked for defects in this book and I have pointed these out when I have found them. They are, however, minor blemishes in what is otherwise an

https://doi.org/10.5771/0943-7444-1975-1-51 https://www.inlibra.com/de/page/agb - Open Access -