Book Reviews

LANCASTER, F.W.: **Vocabulary Control for Information Retrieval.** 2nd ed. Arlington, VA: Information Resources Press 1986. 270p., 57 figs., ISBN 0-87815-053-6

Since the appearance of the first edition in 1972, the subject of Lancaster's book has lost none of its relevance. Lancaster's survey goes far back into the history of the various kinds of information systems, some of which stem from the pioneer days of information retrieval. The emphasis lies on the thesaurus as a means of vocabulary control, perhaps, too, because of the particularly great interest that this device has found in practical information supply.

Right at the beginning of the book, the process of and the necessity for vocabulary control is presented in great detail and illustrated in clear, well-chosen examples. For those concerned with the structuring and application of thesauri, this book offers a valuable store of information, provided that the author's opinions presented in other parts of the book do not deter them from further pursuing this course; for, in addition, Lancaster's book not only gives a good impression of the consequences of the lack of theory, but also of the numerous controversies and inconsistencies widely met in the literature on information systems.

The attentive reader will soon discover that this book falls into two distinctive parts. In the first part the necessity for having any form of vocabulary control is postulated: "In information systems, it is usually *necessary to control the vocabulary* used to describe the subject matter being dealt with" (p.1).

On p.5 the necessity of vocabulary control is explained in the sentence: "One can get some idea of what might occur if the system operated without control by examining the list of terms in Exhibit 2". There then follows a list of fifty words, expressly described as a *selection* which, in the case of uncontrolled input, must all be taken into consideration as alternative search parameters when a search for literature on the concept of "joining" is to be phrased. (Even more convincing would be an example of a search for insects or arthropods in some context. In this case, hundreds of thousands of names would have to be collected and worked off by the program.)

The role of vocabulary control and indexing is taken up again and defined more precisely on p.7: "To promote the consistent representation of the subject matter... the control (merging) of synonymous and nearly synonymous expressions ...distinguishing among homographs...to facilitate the conduct of a comprehensive search on some topic by linking together terms whose meanings are related...".

The usefulness of descriptor assignment is also mentioned in another example (aerodynamics, p.160). Thus an essential task of vocabulary control is, at the beginning, still seen as facilitating, if not enabling, the formulation of a productive search. In the case of uncontrolled input, words of related meaning would simply be widely scattered in an alphabetic list (p.6). The surmounting of this "serious problem" is justifiably seen in vocabulary control.

In clear contrast to this convincing argumentation another opinion is presented in the second part of the book, which seems to begin with chapter 17, "Controlled Vocabularies vs. Natural Language". Concerning the future of vocabulary control, Lancaster states: "It seems certain that natural language will become the norm in information retrieval and that the use of conventional controlled vocabularies will decline" (p.173; the use of natural language terminology in a thesaurus does not, in Lancaster's view, fall under "natural language").

According to what importance is given to the one or the other part of the book, one can come to diametrically opposed conclusions. However, the reasons presented against vocabulary control are weak as compared to those as set down in the first part of the book. For the most part they consist of a mere rejection of the aforementioned arguments. For example, the need for distinguishing homographs is rejected: "The homograph problem is the most trivial; it is more theoretical than actual" (p.162), and the reasons given for this statement are hardly convincing.

The same kind of reversal is to be found in an example on p.164, which must be seen in relation to that of a search for the concept of "joining" in the first part of the book. Now, absolutely no difficulty is seen in collecting natural language terms on a given subject. In a search for literature on "What people eat", two, and only two terms are named as natural language search alternatives, namely "diet" and "nutrition". Here, Lancaster disregards that this is only a very small selection of search words. Roget's thesaurus, for example, would suggest hundreds of natural language terms for this topic which should all be taken into consideration as search alternatives (milk, cheese, bread, sugar, vitamin A, B, C...). To these could be added thousands of other terms from the food sciences, and their number is likely to increase steadily in the course of time. Lancaster's example with only two natural language search parameters wrongly suggests that a searcher can be expected to collect all the alternatives necessary for an adequately phrased search. It is merely admitted that some ingenuity and more effort is required than in using a file based on controlled vocabulary input. But, what in effect is required from the searcher here is not only knowledge of terminology or of given subjects and intelligence (and a steady increase in time and concentration expenditure with a correspondingly high percentage of failures), but also a clairvoyant sense. This will be all the more true when the files have increased in volume and when the language diversity has correspondingly grown.

The argumentation in favour of uncontrolled storing neglects the very core of the problem which was seen so clearly at the beginning of the book, namely that the number of conceivable expressions and, hence, of possible and necessary search alternatives can be extremely large and, what is more, constantly grows and that such a collection of search terms can hardly ever be complete. It is inevitable that the results of correspondingly defectively phrased searches will be incomplete. Therefore, merely experience and subject

https://doi.org/10.5771/0943-7444-1987-3-164 https://www.inlibra.com/de/page/agb - Open Access - (cc) BY

knowledge can in no way compensate for the renunciation of vocabulary control. This holds true at least as far as searches for generic concepts are concerned (see below).

In another example Lancaster maintains that in an information system without vocabulary control all terms concerning the concept "South America" can be collected and used to phrase a search (p.165). He continues that the search possibilities in the uncontrolled file are not only just as good as those in the indexed file, but are even better. But this only seems to be so because in his example only the names of countries have been collected as natural language search alternatives. Again it has been overlooked that "South America" can be expressed and implied in many different ways in a text, namely in the names of mountains, rivers, landscapes, ethnic groups, regional dress and food, regional personalities and animals, and so on and so forth. All these would have to be collected to achieve a result equal to that derived from an indexed file in which the indexers, in accordance with their sets of rules, had assigned the term "South America" to those texts.

On several other occasions, too, Lancaster states the superiority of the natural language method. In most cases, however, he has arrived at such a conclusion because the bases for his arguments are inadequate thesauri. It is true that a weakness of many indexing languages lies in the fact that they do not represent the full specificity of the original. However, this can be countered by taking appropriate measures (more specific descriptors *and* a good arrangement in the thesaurus *and* a good index language syntax). None of the thesauri used by Lancaster in comparison has these qualities. It is therefore questionable to conclude in such a situation that all work with a thesaurus or any other kind of controlled vocabulary is inferior as compared to work without such facilities.

In Lancaster' s book there is hardly any mention at all of those problems with which we are confronted in uncontrolled input in connexion with concepts for which neither colloquial nor technical language has found any concise mode of expression in word-form and which therefore always occur in the form of paraphrases or definitions. Only once, on p.166, there is a slight indication of this problem: "A single word may be equivalent to a phrase". It would have been better to say "... equivalent to thousands of conceivable, different phrases"! The number of such conceivable non-lexical expressions is practically unlimited, and it is almost impossible to collect them even approximately completely as search alternatives and have them worked off from the program.

Such difficulties even occur when such a word exists but is not used consistently by all writers. In practice, the reasons for this are often compelling. In all these cases we are confronted with a practically boundless diversity of natural language modes of expression impossible to master without any control unless one finds oneself in the artificial atmosphere of a research laboratory where one can spend hours and hours on the formulation of a single question.

In the view expressed in the second part of the book, Lancaster makes no difference between generic concepts on the one hand and non-generic concepts on the other. In another place (p.121), however, he does see the necessity of handling these two kinds of concepts in different ways. Non-generic terms in particular are open to a great deal of simplifications, and the use of a non-controlled natural language mode of expression is for them (but only for them) quite promising.

Post-controlled vocabularies, for example, which Lancaster considers an effective substitute for indexing and conventional vocabulary control, can only function satisfactorily with non-generic concepts. Only these are almost always expressed lexically, that is not as paraphrases or definitions, and only with them is the meaning of their names relatively independent of the context, which is a precondition for effective post-control.

In his assessment, Lancaster also makes no difference between the various subject fields and the purposes of information systems. If, for example, it suffices to order and search for texts merely according to their receivers and senders or their authors, then any form of vocabulary control becomes superfluous as a matter of course. This also applies to generic concepts as long as the searcher can associate the subject of his question to the names of persons, corporations, etc. with a sufficient degree of reliability. But an information system which works on this basis only functions at the initial state, which is always a deceptive one, not only in this particular regard.

Lancaster also sees an inferiority of controlled vocabularies concerning the compatibility of files. This conclusion was reached by comparing the compatibility of two different groups of files. One group comprises files made up *without indexing in one and the same language*. The other consists of files which were *indexed* manually, and that *in different indexing languages.* Inevitably, under such unequal conditions the manually indexed files come off worse than the natural language ones. - A similar one-sidedness is to be seen in the fact that Lancaster bases his arguments on the interpretation of the Cranfield Experiments as given by the experimenters themselves and, among others, by Lancaster himself (p.170). The vast amount of literature proving how ill-designed many of these experiments were and how slight their evidential value has been given no room.

As far as the chemical sector in particular is concerned, the views expressed by Lancaster in the second part of the book are void of any validity. If we were to content ourselves solely with the text-words written down by the authors of chemical publications for chemical compounds, then it would only be possible to search for individual compounds (and that only with a great deal of steadily increasing time expenditure) and hardly for generical chemical concepts such as substance classes or reaction types. Firstly, such names are only rarely used by the authors. The structural formula diagram is mainly used instead. Secondly, most chemical reactions and a lot of generic concept terms have no names at all. In such cases, we are completely dependent on the indexers' transferring such non-lexical, mostly graphic modes of description into another, mostly topological form of representation with a high degree of representational fidelity and which is easily predictable at the time of a search.

When Lancaster sees a general trend away from

https://doi.org/10.5771/0943-7444-1987-3-164 https://www.inlibra.com/de/page/agb - Open Access - (cc) BY

manual indexing, then, at least as far as chemistry is concerned, this must be expressly denied. In this field, due to the increased use and the rapid growth of files, the need for more efficient indexing methods has risen even though they may be more expensive. The transition to the topological storing of chemical structural formulae in the Chemical Abstracts Service and the widespread development work on such an expensive indexing speak clearly for themselves. The field of chemistry deserves consideration, for, apart from medicine, the largest and most intensively used information systems are surely to be found in this area.

It is indeed true that the "escalating costs of human intellectual processing..." are an essential planning factor in the field of information supply (p.173). But we must not forget that, though this may not be so obvious, an information deficit and inadequately selective information systems, that is systems which have been established with an emphasis on *input* parsimonity, also cause an increase in costs. It is this consideration that has led to the perfecting and expansion of indexing in many places.

At this point, reservations must be made concerning the standpoint which is held in the second part of the book and here in such a high degree of generalization. Lancaster's, for the most part unjustified, criticism of manual indexing could lead into the transition to (or into the immediate use of) only deceptively and only initially adequate, more primitive information systems. Thus, the book might endanger the existence of an operational information system without offering a workable, alternative solution.

Lancaster states that the future prospects of "hybrid systems" are good, that is, systems which work partly with and partly without vocabulary control. The existing results based on experience with such systems confirm this opinion as in these systems the respective specific weaknesses in controlled vocabularies on the one hand and in non-controlled input on the other can be overcome. Lancaster, though, prefers those systems which work with a minimum of vocabulary control. But nervertheless it is difficult to fathom how all this can be reconciled with doing away with vocabulary control as is, at least implicitly, recommended in the second part of the book. If this were to happen, then hybrid systems could not exist either.

It is difficult to make an overall assessment of this book as it supports two conflicting standpoints on the central issue of the usefulness and economics of vocabulary control. The collection of facts is instructive and well worth reading, many conclusions and recommendations which Lancaster has drawn from these facts, however, cannot be supported in the generalised form in which they are presented.

Robert Fugmann

Dr.R.Fugmann, Alte Poststr. 13, D-6270 Idstein

SOUTH, Mary L. (Ed.): **Dewey Decimal Classification for School Libraries.** British and international edition. Albany, N.Y.: Forest Press 1986. IX,179p. ISBN 0-910 608-35-0

Translated into at least 12 languages and used in some 135 countries of the world, the Dewey Decimal Classification continues to be a widely used scheme. Its popularity outside the United States, its home country, has always moved along a spiky graph. So far, about 40% of the 47 000 sets of its current 19th edition have been sold outside the USA. Besides its two principally known editions, the unabridged and the abridged version, now in their 19th and 11th editions respectively (both published in 1979), there are numerous officially sponsored as well as unauthorised home-made adaptations available to meet the needs of libraries in various cultures and nations. Its non-literal notation of Indo-Arabic numerals promotes its use in all linguistic regions. The hierarchical nature of this notation makes the scheme amenable to use in all sizes of libraries by permitting the notational string to be cut at any desired point from the right end. But use of the DDC is not as mechanical as that. It is a judicious process involving judgement and knowledge. To help libraries to truncate the number at a suitable point, the DDC numbers on LC printed cards and MARC tapes are since January 1967 being presented in two or three segments indicated by prime marks. If they so wish, libraries can mechanically delete any full segment from the right end. However, for various reasons such devices are not available to all small libraries, nor do pertinent services cover all publications or libraries. Therefore, although edited and promoted with a view to international acceptance and usage, the DDC fails to fully meet the needs of its varied and large body of users.

To help such small libraries, smaller versions are available. Historically speaking, in 1894 a first brief outline of the scheme was issued which became the harbinger of the now well-established Abridged Edition first published in 1921. The current Unabridged (19th) and Abridged (11th) Editions list 29 528 and 2 516 classes respectively. The abridged edition is meant for libraries comprising some 20 000 books. This still is too large a version for small and school libraries. Therefore, to meet the classificatory needs of school libraries in the UK, Forest Press and the School Library Association of the UK co-published in 1961 a first school version based on the 8th Abridged Edition (2). Its success prompted a second (1968) and a third (1977) edition (3-4). Since then the work has secured a safe niche in the DDC house and history.

In the de facto 4th edition of the book reviewed here, the two short forewords describing its brief history are followed by a detailed Introduction (p.1-22), which, although quite useful, may not make easy reading for those it is meant for - the staff of small and school libraries. Proper comprehension of this Introduction may be a bit hard for users in African and Asian countries where English is not the language of the many. A simplified introduction should have been a key concern. The Introduction is followed by four tables - showing considerable trimming - of auxiliary notations, namely: Standard Subdivisions (Table 1), Areas (Table 2), Subdivisions for Individual Literatures (Table 3), and

https://doi.org/10.5771/0943-7444-1987-3-164 https://www.inlibra.com/de/page/agb - Open Access - (cc) BY