NATURAL LANGUAGE VS. CONTROLLED VOCABULARY

(LANGAGE NATUREL VS. VOCABULAIRE CONTRÔLÉ)

Elaine Svenonius
School of Library and Information Science
The University of Western Ontario
London, Ontario
N6A 5B9

ABSTRACT

The purpose of this paper is first to explicate the meaning of controlled vocabulary and then to argue for the use of controlled vocabularies in retrieval systems. A vocabulary is regarded as controlled to the degree it permits the classification of terms. Bibliographical control cannot exist without vocabulary control.(Cet exposé a pour but de définir le concept de "vocabulaire contrôlé" et de démontrer la nécessité de vocabulaires contrôlés dans les systèmes de repérage de l'information. Un vocabulaire est "contrôlé" dans la mesure où il permet une classification des termes. Aucun contrôle bibliographique n'est possible sans un contrôle de vocabulaire.)

CONTROLLED VOCABULARY

Numerous articles in the recent information retrieval literature extoll the virtues of natural language indexing and searching. But not much has been written in argument for controlled vocabularies. The purpose of this paper is first, to explicate the meanings of the expressions "natural language" and "controlled vocabulary" and then to argue for the use of controlled vocabularies in retrieval systems.

A linguist would frown upon the use of "natural language" to mean the opposite of "controlled vocabulary". The implication here is that "natural language" is synonomous to "uncontrolled vocabulary". This is unduly restrictive. By "natural language" is understood more than just a vocabulary; there are also rules of syntax and rules of pragmatics. However, since the information retrieval literature uses "natural language" to mean "uncontrolled vocabulary," it is this sense that will be assumed in the present paper. The question then becomes what is meant by a "controlled vocabulary"?

In his book Vocabulary Control for Information Retrieval (1972) Lancaster uses the expression "controlled vocabulary" as roughly synonomous with "authority list" (p. 1). Generally it is assumed that vocabulary control exists when a user of a retrieval system is denied freedom of expression and is constrained to select search terms from a specialized vocabulary. Examples of controlled vocabularies are the Library of Congress subject headings and the various select list of descriptors used as aids in searching computer data bases.

According to this view a controlled vocabulary is simply a restricted vocabulary. But perhaps a more useful meaning of control is possible? For instance, it would be useful if an operational definition of vocabulary control could be developed, a measure whereby the concept might be quantified and treated as a variable which could be used in experimental or theoretical research. This paper suggests, but does not develop in any detail, one approach to such a measure. Basically the approach is to regard a vocabulary as controlled insofar as it permits the classification of terms.

Even the simplest vocabulary control involves classification. The first step towards controlling a vocabulary might be the classing together of grammatical variants of the same term--for instance, the singular and plural variants of a noun or grammatical variants representing different inflections of a verb. For the purpose of retrieval, variants such as these might well be treated as indiscernible. A second, larger step toward vocabulary control might be to class together all words used to describe the same concept or similar concepts; that is, synonyms or words which are equivalent in meaning. What is meant by "the same concept" or "equivalence in meaning" is philosophically problematical but may be less difficult to explicate in the context of information retrieval. Currently work is being done on this problem by Gillian Michell (see her paper at this conference); the problem includes also defining what is meant by quasi-synonomy or nearness in meaning. The classification of terms on the basis of similarity relations such as "being a grammatical variant of" or "being equivalent in meaning to" is the primary function of see references in subject heading lists and use references in thesauri.

There are other relations which might be useful for the purpose of controlling vocabulary. An interesting type are those which are definable in terms of automatic processes. For instance, all terms which begin with the same letters (up to a specified number) may be classed together. This is the function of the right truncation option used in some on-line retrieval systems. Orthographic similarity need not be restricted to the first letters of terms--one might be interested in terms with similar suffixes or infixes--but this is probably the most useful sort of orthographic likeness that can be exploited for retrieval purposes. Another mechanical means of achieving vocabulary control is to form classes or clusters of terms which are related by virtue of their co-occurrence in the same units of text within a collection of documents, queries or search strategies. A term is said to belong to a given class or cluster if it co-occurs (according to some measure and with respect to certain texts) above a threshold number of times with other terms in the cluster.

Classes formed on the basis of certain similarity relations such as "occurring in similar contexts,""having similar spelling" or "having similar meaning" are conceptually easy to understand and, it is to be hoped, are even capable of precise definition. However, there may be other relations of similarity which are useful in retrieval but where rigorous definition may be neither possible nor helpful. Examples are the thesaurus groups used in the SMART system or some of the classes brought together by the EXPLODE command in the MEDLINE system, terms associated by the related-term references used in most thesauri and terms belonging to the same facet or subsumed under the same heading in the UNISIST Broad System of Ordering.

Recognition of hierarchical relations is a function of most controlled vocabularies. Mathematically the hierarchical relation is like the similarity relation, differing only in the respect that it is antisymmetric rather than symmetric. Terms related hierarchically may be said to belong to the same class but to differ in specificity; for instance, animal and man. A more cogent definition of hierarchy would depend on an explication of specificity and cannot be pursued here (Svenonius, 1971). The point to note is that the hierarchical relation, like the similarity relation, is instrumental in the formation of classes. In this case it is special kinds of classes, viz. subclasses.

Summarizing the above few paragraphs, vocabulary control is achieved in an index language by virtue of its classificatory or syndetic structure. The degree of control achieved is a function of the number of kinds of similarity and hierarchical relations which are made explicit in the language. Making explicit a syndetic structure is characteristic of controlled vocabularies. As was mentioned, another characteristic of controlled vocabularies is that they represent a restricted subset of all possible natural language terms. According to the view of vocabulary control advocated here, the first characteristic is essential and the second incidental. From this point of view the dichotomy between natural language and a controlled vocabulary is untenable and may be even insidious as it impedes clear thinking. One can speak of natural language terms without any control or these terms subjected to minimal or more than minimal control. Thus, "controlled vocabulary" and "natural language" might be regarded as scalar antonyms; that is, like "big" and "small" they are not complementary but refer, albeit vaguely, to degrees along a continuum.

An ideal thesaurus would include every natural language term as part of its entry vocabulary. In other words, the entry vocabulary of this thesaurus would demonstrate a hospitality so complete that the user would find in it any term he could think of. This is not to say that he could search on any term. Most thesauri convert a good portion of entry terms to search terms (sometimes called descriptors) along the lines of the classificatory techniques suggested above. While natural language allows the user to address the system with any term he might think of, the classificatory structure of the index language may help him to find terms he could not imagine. What should be the reduction ratio of entry terms to descriptor terms, for different languages and different retrieval purposes, is a theoretically interesting question. It is, perhaps the real question in the natural language _vs_. controlled vocabulary controversy.

The misleading use of the expression "natural language" has been noted. Some further terminological confusion might be mentioned. The expression "post-controlled" is sometimes used to refer to the situation where a user addresses a system using "natural language" and yet is permitted to exploit certain thesauric defined relationships. Before and after what? Also misleading are the expressions "uncontrolled" and "controlled" as they are used to refer to the entry part and the descriptor part of an indexing language. In one sense, at least, entry terms are controlled and that is because they are related by _use_ references.

Of course a user should be permitted to approach a data base or book collection on his own terms and using his own language. This has been a first principle of classification and subject indexing since the time of Cutter. Another first principle has been that works on the same subject must be brought together. It is this latter principle which is questioned by some advocates of natural language indexing and searching. The case for natural language is stated succinctly by Lancaster:". . . the arguments for natural language searching include the great specificity that is possible in systems of this type, the fact that natural language is a "user-oriented" language, the fact that several respectable research projects have shown that natural language systems can produce results at least as good as if not better than controlled vocabulary systems, and the fact that many machine-readable data bases (e.g., of searchable abstracts) are now available as by-products of other operations, notably publishing activities". (Lancaster, 1975) To these arguments can be added Michael Keen's observation that "the all-round acceptability of the Uncontrolled Language is considerably enhanced by considering the lack of extensive intellectural effort that such languages require at the input and index language construction stage". (Keen, 1973)

Lancaster regards the second of his arguments as the most compelling. This is the argument that natural language as opposed to an artificial or restricted language, is user-oriented. It is unrealistic he argues for users to master the nuances of an artifical language such as MeSH. This goes without saying. But it is not an argument for unlicensed natural language. It is to be hoped that the user would not have to memorize the MeSH subject headings or second-guess them. But the fact that he must guess is a consequence not of the control that is introduced but of the control that is not there. The MeSH vocabulary is insufficiently controlled in the sense that its syndetic structure is incomplete. It is not an ideal thesaurus in that not all possible entry terms are recognized and converted to the restricted set of descriptor terms. Complete control carries with it the implication that the entry vocabulary accommodates the language of all users, i.e. natural language. Only when control is complete in this sense is the user free to use his own language in designing a search strategy, confident that his search will not end in frustration.

Lancaster contends that "one of the significant advantages of on-line systems is that they may be used directly by a scientist . . . who has some information need to satisfy (i.e. they may be used in a nondelegated mode)". (Lancaster, 1975) On-line systems must therefore employ natural language. The implication here is that a controlled vocabulary, since it is not user-oriented, is incompatible with a nondelegated search mode. This is surprising. Reports from users of natural language on-line systems seem to indicate that a nondelegated search mode is not only frustrating but expensive. Usually an intermediary is employed to negotiate the user's query and to design a search strategy--in other words, to mediate the language of the user and the language of the retrieval system. Especially when the two languages are both "natural" would it seem that an intermediary is needed to achieve compatibility. For the most part humans have been used for delegated searching, but it is conceivable that a user might negotiate with a mini-computer equipped with a well-controlled vocabulary and the capability of presenting terminological displays on-line. Once a search strategy has been developed the mini-computer could translate it into the command language appropriate to whatever data bases are to be searched.

Delegated searching, which makes use of the term associations in a well-trained human mind or the syndetic structure of an on-line thesaurus, is surely cheaper than browsing a large document file. Michael Keen's argument cited above overlooks the phenomenon of suboptimization. True, intellectual effort is saved at the indexing stage when natural language is used. But what happens at the searching stage? How can the intellectual effort be avoided? A user is not given freedom of expression when he is permitted to guess at all the ways in his subject may be represented in the literature. Under the guise of liberal policy (natural language) the indexer's burden is transferred to the user who may be paying for his possibly amateur thinking at the rate of &100. or more an hour. Surely the producers of on-line systems have something to gain in advocating uncontrolled vocabularies?

The cost factor is especially serious when the data bases to be searched are very large. One measure of large is defined with respect to the number of searchable terms in the data base. (Giering, 1975) With the use of a controlled vocabulary the number of such terms could be kept small, say around 1000. If instead natural language titles, abstracts or full text were searched, then every word of text might figure as a searchable term. Conceivably there could be 100,000 or more such terms. If a serial search were required of this many terms, the difference between large and small would be a matter of considerable significance and cost. Even if inverted files or more sophisticated file structures were used, the difference in cost could still be significant.

It is argued that "respectable" retrieval experiments give evidence to support the belief that natural language indexing is superior to that using a controlled vocabulary. (Lancaster, 1975) The most well known of these experiments is probably the Cranfield II experiment conducted by Cyril Cleverdon. (Cleverdon, 1962) Cleverdon, in his discussion of the test results, finds "that it is difficult to believe that a controlled vocabulary should be less efficient than natural language, even though evidence of the test points to such a conclusion." (Vol. 2, p. 263b) An alternative to rejecting what is difficult to believe is to question the evidence. The evidence brought by any retrieval experiment is eminently questionable. The sampling and definitional problems which beset retrieval experiments are far from being solved. (Svenonius, 1975) It would be premature to suppose that they have reached respectability. A brief example will illustrate.

A particular finding of Cranfield II is that as increasing control is introduced, beginning with natural language terms and then controlling for synonyms, quasi-synonyms and word forms, recall improves and precision becomes worse. (See Table 1)

Table 1.

Relative Effectiveness of Different Degrees of Vocabulary Control as Determined by the Cranfield II Experiment

Degree of Control	Test Collection*					
	A		B		C	
	Precision	Recall	Precision	Recall	Precision	Recall
Natural Language	4.1%	59.5%	5.2%	54.7%	3.2%	66.7%
Natural Language and Synonym	4.0%	61.7%	5.1%	57.2%	2.8%	67.7%
Natural Language and Synonym and Quasi	2.6%	70.5%	2.5%	67.6%	1.5%	72.7%
Natural Language and Synonym and Quasi and Word Forms	2.5%	73.3%	2.4%	70.4%	1.4%	76.3%

*These results are for single term languages, the 1400 document collection at coordinate level 3. Test Collection A is the 221 question set, B the 35 question set and C the 42 question set. The information is taken from pages 87 to 95 of (Cleverdon, 1962).

Given the initial low precision figures, could any impairment in precision be serious? This we do not know since significance tests were not carried out. Neither do we know whether recall improved significantly. Are we to conclude that term control at the synonym and quasi-synonym level is a waste of intellectual energy? The answer depends on so much! Primarily, it depends on what the user wants. He may be willing to sacrifice a great deal of precision in the interest of good recall. The answer depends also on the determination of statistical significance and on how the various control measures (synonomy, quasi-synonomy, words forms) have been defined. There are then the usual problems of relevance and whether precision and recall are useful as measures of retrieval effectiveness. Finally, and perhaps most perplexing, is the problem of generalization. Should we transfer a finding, and possibly an insignificant one, from an experiment using a specialized collection to what would be the case for any indexing of any collection?

One of the criticisms of a controlled vocabulary is that it lacks specificity (Lancaster 1975). Insofar as the classification of terms on the basis of similarity relations is a recall device it increases the number of documents retrieved. It would appear that specificity is lost, and the more so as one proceeds to introduce more controls from synonyms to quasi-synonyms, to word forms and so on. But the loss may be only apparent. There are two arguments here. The first is that for most retrieval purposes certain terms are rather obviously indistinguisable. Examples are the singular and plural variants of a term and the twenty-one trade names for aspirin. It is highly probable that regarding these as different terms would lead to the loss of relevant documents. It is well-recognized that the inevitable tradeoff between precision and recall oversimplifies the facts. Another related but less well-recognized oversimplification is that the smaller the document classes retrieved in response to a request the greater precision is likely to result. There can be too much specificity. -- The second argument is that the control introduced by syndetic recall devices need be neither mandatory nor automatic. Such devices can be used according to the user's discretion. Some user may really want to distinguish twenty-one kinds of aspirin and he should be given this option in negotiating a search strategy.

Classing like things together is fundamental to bibliographical control. In the area of cataloging the literary unit principle has the function of classing together all works of a given author -- for which reason there are authority files. In subject indexing there is a similar principle, but it is not so explicitly stated. Subject headings bring together all works on a given subject. Cutter developed subject headings in reaction to title term indexing, a nineteenth century precursor of keyword indexing. The objections to title term indexing were 1) the title of a book might be fanciful in which case it would not lend itself to retrieval by title terms and 2) books on the same subject might be separated because of the use of synonyms in their titles. Bibliographical control in the past has been predicated upon the need to normalize names of people and names of subjects. Can we say now that there is more literature than ever to control that this is too costly or that it is useless? The argument that a manual system has a tendency to breakdown if completely uncontrolled natural language is used but that a computer-based system might still operate effectively (Lancaster 1975) seems to attribute magical powers to the computer.

Bibliographical control cannot exist without vocabulary control. If a user needs only one or a few documents on a subject, the probability could be high that the language he chooses to search with will find a hit somewhere. Since he is not interested in having everything on his subject it will not matter that he misses some possibly relevant documents; indeed he might be grateful. Nevertheless we cannot depend upon random selection measures to do as well as artfully constructed ones. -- And, again, not all users are the same. There are users for whom it is imperative to know the state of the art of a subject. Historians are an obvious example. But also scientists have a requirement for comprehensiveness. (It cannot be seriously argued that scientific rediscovery is cheaper than vocabulary control?) So long as we are interested in the comprehensive coverage of some domain of knowledge vocabulary control is a *sine qua non*. The truth of this statement is apparent on rational grounds and needs no validation by experiment.

REFERENCES

CLEVERDON, C. and Keen, M. 1962 Report on the Testing and Analysis of an Investigation in the Comparative Efficiency of Indexing Systems. ASLIB Cranfield Research Project. Cranfield, College of Aeronautics.

GIERING, R.H. 1975 Search Strategies and User Interface. Journal of Chemical Information and Computer Sciences, 25: 6-11.

KEEN, M. 1973 The Aberystwyth Index Languages Test. Journal of Documentation, 29: 1-32.

LANCASTER, F.W. 1972 Vocabulary Control for Information Retrieval. Washington, D.C., Information Resources Press.

LANCASTER, F.W. 1975 Vocabulary Control for On-Line Interactive Retrieval Systems: Requirements and Possible Approaches. Third International Study Conference on Classification Research, Bombay, January 1975. To be published.

CONTROLLED VOCABULARY

SVENONIUS, E. 1971 The Effect of Indexing Specificity of Retrieval Performance. Ph.D. Dissertation. Chicago, University of Chicago, Graduate Library School.

SVENONIUS, E. 1975 Good Indexing: A Question of Evidence. Library Science with a Slant to Documentation, 12: 33-39.