THESAURIS IN A FULL-TEXT WORLD

Jessica L. Milstead

ABSTRACT

Despite early claims to the contrary, thesauri continue to find use as access tools for information in the full-text environment. Their mode of use is changing, but this change actually represents an expansion rather than a contraction of their utility. Thesauri and similar vocabulary tools can complement full-text access by aiding users in focusing their searches, by supplementing the linguistic analysis of the text search engine, and even by serving as one of the tools used by the linguistic engine for its analysis. While human indexing continues to be used for many databases, the trend is to increase the use of machine aids for this purpose. All machine-aided indexing (MAI) systems rely on thesauri as the basis for term selection. In the twenty-first century, the balance of effort between human and machine will change at both input and output, but thesauri will continue to play an important role for the foreseeable future.

INTRODUCTION

With the dramatic increase in availability of searchable full text—and the increasing availability of powerful engines for searching the text—it is reasonable to ask if there is any place left for thesauri in this new information retrieval scenario. It is my thesis that there is a place for thesauri—or something like them—but that they must change in order to continue to be of value, and it is hard to predict just what the changes will be.

First, it is important to define what is meant by the word “thesaurus” in this paper. Simple equivalence lists, the kind of “thesaurus” most often supported by text retrieval packages, are much too limited to be considered. Certainly equivalence lists are vital to effective information retrieval,
but these are not enough. They can only suggest other ways of expressing an idea which is already in the user's mind; they do not remind the user of related ideas that might be valuable in searching.

A true thesaurus has equivalence relationships, but it also supports other kinds of relationships—such as genus-species—and provides navigation assistance by means of scope notes and other aids. In other words, a thesaurus is a tool designed to aid users in finding their way around a vocabulary database. In addition to its primary use as an authority for the terms used in indexing the database, it offers reminders of terms the user might not even have considered.

The ANSI/NISO standard for thesauri (NISO, 1994) provides the best available information on what thesauri should do and how they should be built, but it predates the explosion of full text and powerful search engines that we have recently seen, and it is not an adequate guide to future needs and potential.

In order to set present-day thesauri in context, it is useful to look briefly at their history. The first thesauri were actually produced before electronic searching was widely available, but their full development coincided with the growth of online bibliographic databases. The unitary terms of a thesaurus provide much greater flexibility in searching than a subject heading list, with its complexities of subdivision and inversion. Consider a subject heading:

Automobile engines—Manufacturing

Now consider one of the ways in which this complex concept might be indexed with a thesaurus:

Automobiles
   Engines
   Manufacturing

The specifics depend on the design of the particular thesaurus and, in particular, the extent to which it precoordinates the elements of a complex concept. Regardless, the thesaurus indexing offers far greater searching flexibility, though with a possible penalty in false retrievals. Whether based on an ANSI/NISO standard thesaurus or not, most databases today are indexed with thesaurus-type terms. The exception is some of the databases designed primarily for schools and public libraries, which use more complex terms, and typically based on Library of Congress subject headings.

The earliest electronic files consisted only of titles, bibliographic descriptions, and indexing; if you were lucky there were abstracts, but this was by no means to be taken for granted in the days when storage space
was a very precious commodity and acquiring anything in electronic form generally meant rekeying it. In this environment, indexing had to be of high quality if information was to be retrieved at all, hence the obvious need for thesauri.

Today abstracts are practically universal, and it is beginning to seem as if all information is available in full text. However, this is not true, nor will it be true in the immediate future. (Retrieval of graphic images is not considered here, because image searching still relies so heavily on text captions or descriptions.) Vast numbers of legacy documents remain, and converting these to searchable text is an expensive long-term proposition. Furthermore, many documents are still being produced in printed form only.

Therefore, thesauri and indexing will continue to have a place—at least for awhile—in facilitating access to documents for which electronic text is not available. Their long-run value, however, depends on integration with full-text search.

**THESAURI AND SEARCH ENGINES**

Thesauri actually have a place at both ends of the information access process—i.e., at storage and at eventual retrieval. The universe of electronically accessible full text is so immense, and is growing so fast, that users need all the help they can get in accessing it. The explosive growth of Web search engines, with their rather primitive algorithms, has had some rather unfortunate effects, to my mind. Some of these engines appear to have been developed by people who saw a need, but who had not the vaguest idea that there was already a history of development of tools to fulfill similar needs. There is little evidence that these developers had ever used either Dialog or a library catalog.

Not long ago, in a meeting of a national information society, a speaker gave an example of natural language retrieval of 92 citations from his database on the effect of alcohol on heart disease. A representative of a Web search engine countered with a report of carrying out a search using his engine on the Web and retrieving over 600,000 items. This speaker actually saw this 600,000 as better than 92. True, the 600,000 items were ranked (but so were the 92), but the speaker did not go on to show the relevance of the top ranked items to the query, or how many good items might actually have ranked so low that the user would never have looked at them. In fact, the audience was told nothing at all about how these 600,000 citations were presented to the user. It was almost as if the number itself were intoxicating to the speaker.

A distinction should be made among kinds of tools for facilitating access to full text on the basis of the attention they give to semantics.
Older, exact-match (Boolean) systems give no attention to semantics. Furthermore, they retrieve purely on the basis of the occurrence of the search word or phrase in the document. This means that search terms must appear in the text for the document to be retrieved—if a term appears in the text at all the document will be retrieved regardless of whether the term is important to the meaning of the document or not.

Another approach relies on statistical information—co-occurrence of words in the document, frequency, etc. Natural language parsing may be included as well, but there is no concern with the meaning of the words. The fact that two words co-occur in a document means only that; it does not imply that there is any relationship between their meanings.

Boolean and statistically based systems have been found to have comparable retrieval performance, but to produce very different retrieval sets. That is, searches of the same database using a Boolean engine and a statistically based one often produce about the same number of relevant hits—but there may be little overlap between the two sets of hits.

Intelligent retrieval systems integrate statistical and semantic information—as well as a full battery of linguistic techniques—to retrieve more useful results. Such a system may contain an extensive lexicon, not just of word meanings and equivalents but of word types and relationships. Text is parsed—to a greater or lesser extent depending on the system—and there are often tools for disambiguation of terms. Phrases rather than just single words can also be handled. The most powerful systems actually can determine syntactic or structural meaning, permitting them to retrieve a concept expressed in different words that are not actually in the lexicon. One of these systems is DR-LINK, discussed elsewhere in this volume.

Any of these types of system can produce better results if controlled-vocabulary indexing is present. The index terms can be weighted more heavily than the running text in either statistical or intelligent systems, causing documents which have been predetermined by human (or automatic) analysis to be relevant to the query to rank more highly. In a Boolean system, the chances of retrieving relevant documents that do not happen to contain the words of the search query are improved, though precision is not helped unless the search is specifically limited to controlled vocabulary terms.

Searchers consistently state that they need indexed, searchable full text (Pritchard-Schoch, 1993). For some kinds of queries, statistical techniques applied to the full text have been satisfactory, while others just cannot be answered satisfactorily without indexing. In general, searchers have not had much access to intelligent systems; when they do, it seems most likely that the presence of indexing will continue to improve the retrieval.
USE OF THESAURUS IN SEARCHING

In the traditional scenario, an indexer uses the thesaurus to select index terms for inclusion in the document record. Then the searcher, hopefully referring to the same thesaurus, selects terms which seem likely to produce relevant results and searches the indexing, retrieving on the basis of exact match. Even if the searcher has not referred to the thesaurus, she or he is aided by the indexing because, if the query words appear in the indexing, then all documents indexed with those words will be retrieved, whether the words happen to appear in the text or not.

The basic design of thesauri to date, then, has been as indexing aids, with the expectation that searchers would be able to use these aids as a guide to searching. The notation used in term relationships is abstruse; the fact that "BT" and "NT" mean that two terms are related hierarchically is obvious only to specialists. Furthermore, database producers frequently do not mount their thesauri on search systems. And if the thesaurus is mounted, the search system may not support the full range of navigational information. In other words, the thesaurus is an indexing aid which we hope can also be used for searching, but we frequently haven't put much effort into making this use possible, let alone easy.

It is easy to find evidence in the literature that thesauri are underused by searchers; this is probably due at least partly to the fact that the thesaurus for a database is unlikely to be readily available to searchers. Even without significant changes in the nature of the thesaurus itself, provision of a tool such as the IODyne thesaurus navigator, described in another paper in this volume, should increase searcher use substantially. Permitting the searcher to switch seamlessly between navigating the thesaurus and searching the database can only improve access.

An obvious way in which a thesaurus can be applied directly in retrieval is to use the relationships as a means of expanding the search. Research, however, has shown that these relationships must be used with caution. In general, expanding a search to include the narrower terms tends to improve recall without great sacrifice in precision. Expanding to include broader or related terms, while it does improve recall, typically has a significant negative impact on precision.

In pragmatic terms, the purpose of distinguishing hierarchical relationships is to indicate to users everything that is a "kind" or "part" of the broader term, in order to facilitate making searches more inclusive. However, it is not known for certain that this is what users need or want—or to what extent they need it. We do not know how far up or down hierarchies it would be useful to go in expansion of searches; if a hierarchy is nine levels deep, would a user starting at the top really want to broaden the search all the way down or would stopping at an intermediate level be
preferable? On the other hand, limitation to a single level of expansion is probably not adequate.

**MAKING THESAURUS MORE ACCESSIBLE TO SEARCHERS**

Over the years there have been proposals for end-user thesauri designed specifically to facilitate searching. Bates (1986) and Anderson and Rowley (1992) have both made interesting proposals for development of such thesauri. The end-user thesaurus differs from a conventional thesaurus in two primary ways—its term inclusion and organization and its displays. It is designed to reflect and organize the total specialized vocabulary of users in a field rather than to provide a limited list of authorized terms. It provides more information about the scope of terms, and its displays are designed around the way in which users approach information. For instance, one design of Bates's used term clusters as a device to aid users in enriching their searches. These clusters were like the sublanguages of different specialist groups. Anderson proposed collecting words and phrases from full text and organizing them to build the end-user thesaurus.

The idea of end-user thesauri has not been widely accepted, probably for a number of reasons. Conventional thesauri are costly to develop and maintain; the additional access in an end-user thesaurus would be even more costly. Simultaneously, until recently there seems not to have been a real understanding on the part of system designers that simply making full text available—even with a powerful search engine—is not adequate. The more full text there is, the more help users need in navigating it.

At the same time, users have certainly not been demanding richer thesauri, though I am aware of more than one instance where a major database producer was motivated by user demand to develop a conventional thesaurus. If end users—particularly the more sophisticated ones—were aware of the aid that better semantic tools could provide, they would demand them. Unfortunately, people generally do not miss what they have never had.

**CHANGES IN INDEXING**

Meanwhile, indexing is changing in a way that makes even greater demands on the thesaurus. As stated above, the traditional scenario is one in which the indexer consults a thesaurus as a source for terms to use in indexing. This work is repetitive, labor-intensive, and inherently inconsistent. It places heavy demands on the indexer, who must remember all indexing rules and policies; when indexers must work outside their specialized area, they are handicapped by their suboptimal knowledge of the thesaurus in the new area. Few organizations to date have found a way to provide more than clerical aids to indexers.
For many years there have been a few systems using machine-aided indexing (MAI). In these older MAI systems, the text of titles and abstracts is run against a rule base; when a rule is matched, the applicable thesaurus term is assigned to the document. The indexer reviews these candidate index terms, adding and deleting as appropriate. While their users have found that the systems increase indexer productivity significantly, there has been no great move to MAI by other database producers in the twenty or more years that these systems have been in use.

This lack of growth in use is probably due to the immense up-front cost of developing a rule-based MAI system. First, the system depends on availability of a well-developed thesaurus. Then it is necessary to develop rules for matching sequences of characters in text to produce indexing with a high degree of reliability. This rule base must continually be refined and updated if it is to remain useful. While no published data exist, the rule base probably costs at least as much to develop and maintain as the thesaurus itself.

Within the past few years, one MAI shell system has become commercially available (Hlava & Hainebach, 1996), but it is still necessary to develop the actual rule base. As an aside, it is worth noting that this system is actually multilingual—an aspect of indexing which may be expected to increase in importance in the future.

The availability of powerful text analysis software is changing this scenario dramatically. The same analysis used to provide good relevance-ranked search results can be used to suggest candidate terms for indexing without manual development of a rule base. Instead, a substantial number of already-indexed documents is used to train the text analysis software, which then assigns candidate index terms to the documents for indexer review. Without human review, of course, the same scenario produces automatic indexing.

MAI assumes a developed thesaurus, and ongoing maintenance and refinement of the term assignment criteria. It shifts much of the analysis effort away from review of individual documents to maintenance of the vocabulary and retraining of the system. Indexer productivity can be increased significantly; it is known to have increased in the older rule-based machine-aided indexing systems. The shift from rule base development to more automatic training of the system will also make the process of MAI system development and maintenance less labor-intensive. In fact, none of the changes described have reduced the need for a thesaurus; if anything, they have increased the demands made on these tools and, as a result, are bringing more of their limitations to light.

**PROBLEMS OF THESAURUS DESIGN**

There are fundamental problems in the basic design of thesauri that
make them less than optimally useful for more powerful retrieval scenarios. There is no reason to expect that a tool designed for Boolean search on index terms will be optimized when full text is searched by a powerful engine. Unfortunately, the ways in which thesauri could be redesigned to be more useful are not immediately obvious.

The number of kinds of relationships in the present design is limited—and yet even this specification of types is probably only of marginal direct value to users. As indicated earlier, users do not necessarily recognize that “BT” and “NT” mean a relationship is hierarchical, and “Use” and “UF” mean the terms are equivalent, while “RT” means the relationship is something else—that something being unspecified.

For a thesaurus developer, even deciding when a relationship is hierarchical or part/whole can be difficult. The determination is fairly easy when concrete objects (e.g., truck/motor vehicle) are the issue. However, in a world where the same thing may be a “particle” (i.e., concrete) or a “wave” (not concrete), depending on how the observer happens to be looking at the thing at the moment, deciding whether something is a “thing” or a “process” may not only be difficult, it is likely to be futile. As an example, recently I encountered in building a thesaurus the problem of how to relate “Codons,” the basic units of genetic information, and “Codon usage.” This certainly sounded like a clear case of thing/process—i.e., RT—but it turned out that “Codon usage” was used in the field not for a process, but for studies of the types of codons being used. Thesaurus practice does not offer a good way to distinguish dictionary meanings from actual use of terms in the literature.

If the distinction between hierarchical and other relationships is that porous in fact, of how much value is it to users? The distinction is probably of significant value when it is clearly “things” or fairly concrete entities that are being related, but of much less value when the entities being related are less concrete. Yet, even for very abstract entities—processes and the like—we find ourselves wanting to say “these terms are very closely related, while these others, though less related, might still be useful for you.” Obviously, weighting is involved here, but there is no way to build weighting into a standard thesaurus.

At the same time, text analysis software theoretically can make use of much richer semantic analysis, not only of the relationships between terms, but of the kind of term—e.g., a process, thing, or property. Historically, this kind of analysis has been even more labor-intensive than that required to develop the relationships in a standard thesaurus. For instance, efforts such as the Cyc project have involved manual development of a knowledge base that would permit automatic analysis. On the near horizon, though, are systems which will automate development of abstractions—such as relationships among concepts.
Equivalence relationships grow out of the print paradigm, where everything had to be entered in a single place—i.e., it was not feasible to place a copy of the record under all equivalents of each term. If terms are truly equivalent, perhaps we should treat them as an "equivalence cluster," so that including one of the terms in a query retrieves them all, either automatically or at the user’s option.

Displaying the relationships of a thesaurus in print has always involved compromises. For instance, the typical alphabetical display can only show a single level of upward and downward hierarchical relationships. Thesauri which include the full hierarchy of terms in the alphabetical display become much more voluminous. If the full hierarchical display is relegated to a separate listing, it can be difficult in the alphabetical display to show where to enter the hierarchical listing to see the full hierarchy of the term.

While electronic display of a thesaurus can ameliorate some of the limitations of the print display, making it possible, for instance, to switch back and forth between alphabetical and hierarchical display, the limitations of the screen are substituted for the limitations of the printed page. The screen display does offer possibilities of flexibility and customization that simply are not possible in print, and it is to be hoped that the IODyne browser will turn out to be only the first of a new generation of tools which supports end-user thesaurus access in a friendly and powerful way. More and richer connections between thesaurus and text may be expected as the thesaurus becomes a resource for detecting relationships and refining searches.

**FUTURE OF THESAURI**

These tools, originally designed to facilitate consistent analysis of documents at input to an information retrieval system, are already well on their way to becoming vital retrieval tools as well. In fact, I anticipate that, in the near future, thesauri will be used more at retrieval than at input. They may work behind the scenes much of the time. While users should certainly have access to any available vocabulary aids if they want them, we need to design our interfaces so that users need not interact directly with the thesaurus to any greater extent than they wish or need to.

Given all the problems and limitations indicated, how is it possible to remain positive about the need for continued use of thesauri? There are two fundamental reasons, one philosophical and one pragmatic:

- **Philosophically,** just as thesauri built on subject heading lists, providing more structured relationships and terms better fitted to the cur-
rent searching environment, thesauri can be built on to develop vocabulary tools that meet the needs of users in the search environment of the near future.

- Pragmatically, there is increasing evidence of a realization on the part of text analysis system developers of the need to include a semantic component in their software. Whether this semantic component is a formal ANSI/NISO standard thesaurus is not as important as the fact that a rich semantic tool—not just an equivalence list—is embedded in the system.

A thesaurus can become the basis of a more extensive semantic network, providing information, not just on what terms are used in indexing, but on how they are used within the system. Most often a semantic network includes richer relationships than a thesaurus, but there is no reason not to build the less sophisticated system, using it as a resource when it becomes feasible to develop the more powerful system.

On the retrieval side, the advent of intelligent information retrieval systems, like those discussed in these proceedings, changes the picture of indexing and therefore of thesauri. The question then arises: Which kinds of retrieval can best be left to the intelligent system, and which will be facilitated by indexing—and therefore by use of a thesaurus? This is a very important question, but I know of no attempts to answer it. The concentration has been on refining intelligent retrieval systems and demonstrating their value. Yet, if we knew which kinds of information or queries could be well served by a text retrieval system without human input or refinement, we would be free to improve productivity by concentrating human effort on the types of retrieval needs where it could really add value.

Thesauri and intelligent retrieval systems can be complementary in another way: The thesaurus shows a variety of relationships among terms; these relationships can be used by the system to supplement its statistical and linguistic analyses. Conversely, by flagging phrases which do not match any of its existing criteria, the intelligent retrieval system can assist in thesaurus updating.

SUMMARY

Thesauri were developed to meet the needs of a different kind of retrieval system than the full-text systems which are available today. However, the basic concept of the thesaurus remains useful; the problems encountered have more to do with the implementation than with the concept itself. More work is needed to assure that thesauri built in the future are optimally suited to the needs of full-text systems.
ACKNOWLEDGMENT

Susan Feldman reviewed a draft of this paper and provided extremely useful advice. Any errors of fact or interpretation, however, remain the responsibility of the author.

REFERENCES


