Winfried Schmitz-Esser
Information Systems Consultancy, Hamburg

New Approaches
in Thesaurus Application

Schmitz-Esser, W.: New approaches in thesaurus application.
Int.Classif.18(1991)No.3, p. 143-147, 9 refs.
To show the difference and explain the move to a new kind of
thesauri in the information science area, some of the main
characteristics of conventional thesauri are pointed out as well
as their side-effects. The new approaches for thesauri applica-
tion are seen to exist in (1) expert systems, (2) interface
systems, (3) object-oriented design and programming, (4)
hypertext systems, (5) machine translation, and (6) machine
abstracting. These areas are shortly described including also
the new problems which they might create. A discussion of the
limitations of the new thesaurus application areas finishes the
article which challenges, finally, an awareness to meet the new
possibilities of a thesaurus revival.   (I.C.)

1. Thesauri, a Standard Tool in Information Retrieval

In a way, the application of thesauri in information
retrieval has come of age. Construction and maintenan-
ce of thesauri is an established, well-proven technology.
Everyday application of thesauri has become the back-
bone of hundreds of information systems, many of them
of substantial size. By the dozens count systems in which
multilingual thesauri are used as a basis for a high degree
of international concerted action and co-operation in
indexing and information retrieval.

It is safe to say, that this state of the art and
application is not the end of the story. There is more to
come. Thesauri are about to stride over the rather
narrow boundaries of the library and archival fields, and
to penetrate a much larger area - that of Language and
Knowledge Engineering. And in doing so, they are
changing face, behavior, and, sometimes, even their
name. They feature new types of relations, it is more and
more the machine rather than a human indexer or
searcher who uses them, and they come along, or are
intertwined, with all sorts of other pieces of new techno-
logies.

To show the difference and explain the move, let me
point out some of the main characteristics of the thesauri
as we all know them.

2. Conventional Thesaurus Qualifications and Desired Characteristics

Structures represented in thesauri are relations
that are found to exist between the single terms of a
chosen vocabulary. Such relations may be equivalence,
hierarchy, and some other more or less defined rela-
tionships, mostly categorized in the thesauri as "Related
Terms (RT)". But there may be also more specific, or
subtle relations, such as partitive, constitutive, consecu-
tive, etc.

You may consider a thesaurus as an attempt of
semantic mapping of terms, or of establishing a semantic
network between them.

But however well defined in their structure, thesau-
ri give little or no answer at all with respect to the
semantics of a single term. Some semantic information
may come along with the definition of a structure - e.g.
in the case "Gauge - Metering equipment", or in: "Ferry
- Naval Transport + Shuttle System". Some further
semantic information may be given in scope notes, as can
be found in many thesauri. But apart from this, it can be
stated, that thesauri do not provide semantic definitions
of single terms, and therefore are not semantic tools in
themselves, but rather functional tools in so far as their
main object is to establish a defined order between the
different terms of a vocabulary the semantics of its single
elements being regarded a matter of fact, or, more
precisely: a matter of dominant paradigm, sanctioned by
general or specific acceptance and use of these words in
a given society at a certain period of time.

This is a shortcoming and an outstanding capabili-
ty of an instrument of ordering at the same time. With a
thesaurus of that type, one can establish any order that
may be desired, or felt useful, or seems obvious, or is the
result of algorithmic processing. There is no need of
giving proof, e.g. that the term "Subway station" is a
narrower term of "Urban transport system", and whe-
ther this is a true statement in terms of logic or not. It
may be useful to have the relation between these two
terms defined this way, or in another commonly intelli-
gible way (which can be an important factor in practical
handling), but it is not stringent. If the subway track is
closed and the station is used as a fleamarket, it may be
more adequate to range it under "Urban public market
places", and if it is an architectural masterpiece by Otto
Wagner, or Alfred Grenander, relations could as well be
defined in the direction of "Industrial architecture",
"Style", etc.

It is the focus of interest of the user of the system - or the intended user - which largely influences, or even governs what are valid relations between the terms of a thesaurus. There is a theory saying that the better the users' special points of interest are reflected in term choice and relations, the better the efficiency of the thesaurus as a tool for information retrieval will be.

A second statement refers to the number of documents, and/or physical form and mass of the media carrying the knowledge that is up for ordering by means of a thesaurus. These may be books, abstracts, newspaper articles, picture and film descriptions, etc. Nobody would think of applying a 20,000-term thesaurus on nuclear energy, or medicine, or electricity, to a press clipping library in which 20, 60, or 5 well-chosen thesaurus terms, respectively, would perfectly do to cope with all documents dealing with the subjects related to these fields of knowledge.

A thesaurus must not put more intellectual strain on indexers or searchers than is justified by the content of the documents which are up for order. All what is desired from a thesaurus to bring about is a fair distribution of, and fair possibility of discrimination among, the different terms (elements) of the collection. A thesaurus can be constructed along those lines, and, here as well, there is a theory saying that the better the terminology of the thesaurus reflects the content and level of physical compression of the collection, the more effective its application will be.

Likewise, a thesaurus may be created to correspond to other, more specific needs, like positioning of terms for facts and concepts in time or space, or reflections of such terms in the different media, like items of an exposition (e.g. in a museum), audio-visual media, pieces of art, items of special collections, citations, etc.

All these properties qualify the thesaurus as a most versatile and flexible instrument for the purpose of materials ordering and knowledge organization, and it is not by chance that this has been, up to now, it's prime field of application.

3. Side-effects of Thesauri

If there is a feeling that thesauri generally lack flexibility, and are cumbersome instruments in most cases, this may be, among others, a side-effect of one property which, as a rule, is most welcome - stability of an ordering system in the course of the time (diachronic stability), and for which thesauri originally are chosen. They are meant to serve as a diachronically stable basis for term-based ordering processes.

Those side-effects were difficult to get under control in the past, mostly for reasons of unsufficient trakking and analysis of the use of the outside-world language, and for lack of computing facilities. The problem: Whereas thesauri were kept as systems of diachronical stability, the real-world terminology was undergoing rapid change, so that thesauri ended up looking like

saurians yet before the Information Age had really started.

This handicap is now coming more and more under control. To-day, we know much better about the interest of the users and their behavior in the informations gathering process. By means of machine-aided processing, we can better relate the size, structure, and content of the collections to the structure, level of abstraction, and extension of the thesauri. Also, we know better about the relationship between free text (i.e. the real-world terminology) and controlled (and pre-structured) vocabulary. We know better about the structure, and use, of the language, as a representation of thought and concepts. And although the problem of the meaning has not yet been solved, one can hardly deny that quite some advancements in Language Technology have been made in the meantime.

4. New Approaches for Thesauri

This is where the thesaurus comes in again.

The following main approaches are currently visible, being tried out, or under development:

1) An *expert system* is designed to keep track of the real-world terminology and to relate it to the terms of the pre-established structure of a thesaurus.

This is carried out by means of machine-assisted, intellectual analysis of relevant samples of texts. The expert system may then be used as

- an aid for machine-assisted, interactive human indexing or abstracting, or for all sorts of conditioning of free texts to "enhance" their quality in search procedures.

- an aid for searching free text, along the lines of a thesaurus structure which serves as orientation, or serendipity machine

- an aid to combine terminology requirements in mixcd vocabulary collections.

2) An *interface system* is designed to accept free language query statements from the users and to convert them into query statements as required by the controlled or mixed vocabulary used in the information system. The language the user is allowed to use would be conditioned by a syntax in some way, but would broadly correspond to a language typically used in query situations. Command of real-world search terms would come from an expert which may be the one described above. The result of the interfacing procedure would be a query statement more or less "as if" produced by a human intermediary.

3) Much work is going on in *object-oriented design and programming*, with an aim to get a more clear-cut definition of the objects -, what they are, their procedural aspects, and how they are interlinked, or communicate with each other. Such objects may be terms, or statements. Here, as well, expert systems play a part, and generators of such systems are being developed to serve the needs.

Structures of objects and their behavior created in AI resemble very much thesauri; they may be regarded as highly improved thesauri, or thesauri of the next generation. In a more modest line of approach, sets of objects (terms) are linked by relations such as used, or discussed as possibilities, in traditional thesauri.

4) Preferential terms of a thesaurus could be made to mark nodes in *hypertext systems*, from which on search into more specific subjects/statements would be offered to readers/users. An expert system of the type explained in my first example would serve to establish the relations between the free text terminology and the controlled vocabulary of the thesaurus. As an advantage over normal Hypertext it is expected that the reader/user would obtain the so-much needed "navigational help" in travelling through the documents. This guidance would follow the established logical structure of the relations between the different terms, and it would overcome the difficulties caused by the rather stochastic occurance of words in the texts.

5) Another new field of thesaurus application is *machine translation*. Thesauri, with their knowledge of term strukture, can help the machine to brush up in the appropriate sub-dictionaries, or to disambiguate and find out the intended meanings, both, in case-frame grammars as well as in probabilistic procedures. This will be of utmost importance in the field of speech recognition, where the brush-up has to be extremely fast as to enable real-time machine processing.

6) Thesauri will also be helpful in *machine abstracting*, or, what seems more realistic, machine-aided, human abstracting. With an adequate thesaurus structure at hand, different levels of abstraction could be indicated to the abstracter in an interactive process. This would certainly result in more homogeneous abstracts, broader overview on larger fields of knowledge, and improved properties of the abstracts in search processes.

5. New Problems from New Requirements

This revival of the thesaurus, however, comes along with new requirements and poses some serious problems, some of which I am going to mention now.

When I said in the beginning that a thesaurus was an ideal instrument as to its semantically open nature, flexibility and usefulness in machine processing of information, this has to be related to the question: What sort of machine processing, and to what ends?

At this point, at the latest, it becomes obvious that the traditional thesaurus concept does not suffice for use in the new technologies outlined above. It's because of its weak semantics.

In a thesaurus of the old type, e.g., we may encounter a hierarchical relation between two terms which, in one way or in another, corresponds to the dominant paradigm of the meaning of these terms. Let us also assume that it is evident to a human intellect why there is a hierarchy (and not a related term case, or another

relation). But the definition of what, in terms of semantics, should be represented by a hierarchical relation in our thesaurus is missing. What is it meant to represent? When we said that "Subway Station" should be a NT to "Urban Transport System", what did we mean?

- Is it part of the Urban Transport System? (A statement as to the real world)

- If this is so, are the two descriptors meant as a class which directly represents the station and the system?

- And if so, does it represent all such stations in all such systems? And if yes, what about single ones?

- Or shall we just consider such stations as part of such systems? (A statement as to the perception of the real world and its normal course)

- Shall we do so only as long as the station is actually a part of the UTS (i.e. as long as the trains stop there)?

- etc. etc.

The classic approach - analyzing the single criteria or properties of what in their entity should constitute the broader term (class) - reveals in this case, that evidently it was omitted to define the semantics and conceptual basis that should govern the hierarchical distinction(s).

When it comes to what we normally enter as Related Terms (RT), the situation becomes all but clearer.

Looking at close range, all the well-known relationships are fuzzy in most thesauri. We could afford to allow them to be fuzzy as long as their only purpose was to achieve the desired degree of order in our documents, which is a modest requirement compared with what we need for Language and Knowledge Engineering.

As an example: Isolate a thesaurus from its information collection and the indexers and intermediaries, its users - what then is it good for? Separated from its collection, the choice of terms as well as the term structure no longer will have its justification - much of what was stipulated in the thesaurus will be meaningless.

One might argue that it could be more meaningful if our thesaurus was more detailed, universal, comprehensive, and if it was open for much more than one purpose. Then, however, it could scarcely be economically applied, at least in the traditional way, i.e. for use by humans. Indexers as well as searchers would be lost between unneeded specificity on the one hand, and an almost open space of abstraction on the other.

Also, the construction of such a thesaurus would be very difficult, since it would have to be a result of an amalgamation of a number of more specific thesauri which cannot be expected to be homogeneous in concept or detail. Its maintenance would also cause tremendous problems. In the extreme it would be a monster outdated the day it is accomplished. Almost certainly, it would be impossible to keep track with the day-to-day development of relations between the terms and what they are standing for in the real world.

Nevertheless, it must be stated, that since years quite some effort is being devoted to the question of metathesauri, superthesauri, roof thesauri, etc., mostly in connection with the creation of improved interfaces for the users of online databank systems, - admittedly, however with mixed success.

6. Limitations to New Thesaurus Applications

But I wanted to tell you about new thesaurus applications, especially those which are designed to serve as machine tools or inference machines in language and knowledge technology. The scientific community would certainly be happy if there was a universal, multi-purpose, if possible: multi-lingual, thesaurus which can be used along with lexicons of the same qualifications. But it is an open question whether such a tool will ever become a reality. Looking at the almost unimaginable richness and variety of relations that can exist between terms or objects I am doubtful. All I can see are massive restrictions.

In the new applications, as well as in the old, thesauri will have to be *tailored to special requirements* - they will only function within the limits of special fields of application, well-defined purposes, and levels of abstraction. Here, they can do a good job, and even become indispensable in machine-aided solutions. In such worlds, we can design knowledge ordering-systems with thesaurus generating components (as is shown in the case of the German TEGEN project).

So, returning from wishful thinking, we find ourselves back in the rather small and scattered worlds of real applications. It is a familiar picture, by the way, to all concerned with linguistics and AI. Why should this be different with thesauri?

A *second restriction* is posed by the fact that transition from natural language (free text) to controlled vocabulary is not just a matter of terms, or nominal phrases. We have to consider syntax and some further conditions apart from term semantics, a problem very close, or equal, to what is encountered in machine translation. Thesauri can be helpful for both, text generation as well as text understanding, and in text understanding, the question is how much closer we can get to the meaning of a text by the help of a thesaurus, and by which instruments this can be accomplished.

Definitions as used in object-oriented programming show a way how this can be done, but still work with sets of rather modest suppositions. They will have to be more elaborated and tuned to the special needs of Language Technology. They will have to be intertwined with semantic data modeling.

Low-key solutions may be at hand earlier, e.g. in such cases where terms of a controlled vocabulary are offered instead of free text terms as a suggestion to abstracters, indexers or translators who have to make their intellectual choice(s) in interactive processes.

But it will be extremely demanding once the thesaurus term is meant to be a term requiring the application of some syntax, alone or along with other terms of free language and/or special languages, like a query language - e.g. when we are dealing with syntactic indexing, or machine-aided abstracting. This requires a higher degree of understanding natural language text.

The applicability of thesauri in such machine-aided processes will thus depend on the progress of the linguists in natural language understanding, and their progress will depend, at least in part, from the progress and applicability of special thesauri designed for their solutions in language understanding.

A *third limitation* may lie in the level of abstraction applied in the respective thesauri. Seen against a limited number of, say, 100,000 abstracts on art and history of art, it may be useful to have a descriptor like "Graphic art and society", which would come along with many other, more specific descriptors on graphic art and on society. It is obvious that, from a point of view of vocabulary control, the descriptor "Graphic art and society" escapes term control, since it is impossible to identify all imaginable terms and term combinations, and to determine the inter-relations that may concern, in one way or in another, "Graphic art and society". An expert system would scarcely be better off in learning about such terms and their relations, and in detecting them in the texts. From this we can conclude: There is a higher probability that automated or machine-aided solutions can be found on lower levels of abstraction.

There are quite some more restrictions that should be mentioned in this context. Let me just touch one of them which is of major importance: It is the *lack of standardization*.

Standards for thesauri - monolingual and multilingual ones - exist, but have been developed with a view of application in information retrieval. As a rule, they are not apt to solve problems in other fields of application. Above all, they do not cover the use of thesauri for inference purposes as needed in Language and Knowledge Technology.

Given the tremendous diversity of formal and structural requirements, and the multitude of desirable applications, it is obvious that it will be rather difficult to get a common understanding of how a thesaurus fitting these new needs should look like and how it can be constructed.

In fact, nobody has come forth yet with a proposal for such a new standard thesaurus. Some think it should be integrated in a new type of dictionaries needed in Language Technology (which have to be standardized as well), some refrain from any encyclopedic approach.

I am not referring here on agreements regarding minimal formal requirements, e.g. such of the Edifact type; this is also necessary, and better than nothing, but wouldn't solve the basic problem, i.e. to assure the rather reticent appliers of Language Technology, AI, Information Retrieval, and related technologies of the availabi-

lity of reliable standard thesaurus tools (e.g. on CD-ROM's), on which market products, like MT or MA program packages can be based and built.

Interest from the industry appears scattered and highly fragmented, and therefore can hardly be expected to build up sizeable pressure in favour of normalization. It is the scientists themselves, and their associations and special interest groups who could do a useful job in this. Everybody concerned or interested is invited to co-operate in the venture.

7. What Are the New thesauri?

Summing up, it can be stated: There is a revival of the thesaurus idea. Thesauri are badly needed for solutions in Language Technology, AI and other related technologies used for Knowledge Organization. More and more, the classic realm of information retrieval becomes a matter of these new technologies. Thesauri have to be made fit to prompt the new needs. They will change their face, and probably show up under a number of new names, and along with all sorts of other technologies. Relations featured, and above all, the semantics of such relations, have to be defined/re-defined in a much more elaborate way, to fit the new needs. As far as possible, they have to be standardized, and this can only be done in co-operation with the new appliers in the respective disciplines.

First solutions are showing up in smaller, well defined areas of application. Whether those new thesauri (or however they may be called) can be made to serve multi-purpose applications, and later become universal, is still an open question. The answer will very much depend on the quality of the definitions stipulated.

References

(1) Buchan, R.L.: Intertwining thesauri and dictionaries. Information Services and Use 9(1989)p.171-175

(2) Hjerppe, R.: The Role of Classification in Hypertext. Issues in implementing Roget's thesaurus as a Hypertext. In: Tools for Knowledge Organization and the Human Interface, First Int.ISKO Conf., Darmsstadt, Aug.1990. Frankfurt:Indeks 1990. p.206-215

(3) Lee, J., Malone,T.W.: How can groups communicate when they use different languages? Translating between partially shared type hierarchies. In: Proc. Conf.Office Inform. Systems, 1988, Palo Alto, California

(4) Poulsen, C.: Subject Access to New Subjects, Specific Paradigms and Surveys: PARADOKS-registration. Libri 40(1990)No.3, p.179-202

(5) Rada, R.: Maintaining thesauri and metathesauri. Int.Classification 17(1990)No.3/4, p.158-164

(6) Ruge, G., Schwartz, C.: Linguistically based term associations: A new semantic component for a hypertext system. In: Tools for Knowledge Organization and the Human Interface. Frankfurt: Indeks Verlag 1990. p.88-95

(7) Sarre, F., Mittermaier, J. et al.: Learning behaviour and user acceptance of TEGEN, a thesaurus-generating information retrieval system. In: Czap, H., Nedobity,W. (Eds.): TKE '90, Terminology and Knowledge Engineering. Frankfurt: Indeks Verlag 1990. p.382-394

(8) Tjoa, A.M., Wagner, R.R.: Basic conceptual elements of knowledge-based systems. A unified view and terminology of semantic data models. In: Czap, H., Nedobity, W.(Eds.): TKE '90, Terminology and Knowledge Engineering. Frankfurt: Indeks Verlag 1990. p.6-14

(9) Wallmannsberger, J.: Language limits and world limits in the age of AI. In: Retti/Leidlmair (Eds.): Proc.5th Austrian AI Meeting, Vienna, 1989.

Address: Dr.Winfried Schmitz-Esser, Information Systems Consultancy, Ballindamm 6, D-2000 Hamburg 1

Conference Announcement
SIGIR '91

14th INTERNATIONAL CONFERENCE

on

RESEARCH AND DEVELOPMENT

in

INFORMATION RETRIEVAL

Location:	Chicago, Illinois
Dates:	October 13–16, 1991
Sponsored by:	ACM SIGIR
In co-operation with:	AICA - GLIR (Italy)
	BCS - IRSG (UK)
	GI (Germany)
	INRIA (France)

TOPICS FOR SIGIR '91:	Information Retrieval Theory	Interface Issues
	Artificial Intelligence Applications	Hypertext and Multimedia Systems
	Natural Language Processing	Implementation Issues

| FOR PROGRAM AND APPLICATION FORMS, CONTACT: | Prof. Abraham Bookstein Conference Chair 1100 E. 57th, CILS University of Chicago Chicago, IL 60637, USA | Email: bkst@tira.uchicago.edu Telephone: (312) 702-8268 FAX: (312) 702-0775 |