Winfried Schmitz-Esser
Information Systems Consultancy, Hamburg

Thesauri Facing New Challenges

First Generation of Software Finally Hit the Market, Second Needs Specification
– 10 questions, 10 answers to a topic of renewed interest –

Schmitz-Esser, W: **Thesauri facing new challenges.**
Int. Classif. 17 (1990) No. 3/4, p. 129 – 132.

The chairman of the Thesaurus Software Seminar held on August 14, 1990 in Darmstadt, introduces into the topic by asking the following 10 questions and by providing his answers to them: 1) What is new in the view? 2) What is the real point of attraction? 3) Cannot Information Retrieval (IR) profit from machine-processing of language? 4) Can we do better now? 5) How can we do better? 6) When does fully automatic IR arrive? 7) Thesauri for machine-aided IR – how do we get there? 8) Which is the right way, which is the model, what to standardize? 9) Can IR people do it alone? 10) Are there advanced information services with a truly human interface?
(I.C.)

After a period of rather slackening attention, thesauri as used in Information Retrieval (IR) are back in the focus of interest again. The Seminar we are going to stage today intends to pay tribute to that finding.

Much has changed since the early days of computer application in the field of information when the Thesaurus appeared as a key problem solver in IR, and was regarded - and usually kept - as some sort of Holy Grale in the respective systems. These thesauri typically dealt with Scientific and Technical Information (STI), were constructed and maintained by manual procedures, printed on paper, and at best could be integrated in the retrieval systems proper by customized, non-portable software.

The fact that today a seminar on thesaurus software is taking place shows at least that such software must be available now. Ten or fifteen years ago we could at best have had a seminar on thesaurus construction and maintenance. In the middle of the seventies, the only special software program on sale for such purposes was an IBM program called TLS (Thesaurus and Linguistic System), a rather heavy and - at that time - horribly expensive system which could only be run on IBM mainframe machines.

Question No. 1: *What is new in the view? It is not just the Micro.*

It would be too simple to say that, at a later stage, it was the advent of the micro-computer and its stunningly growing capabilities which changed this all.

Of course, it is the micro to which we owe the first bunch of cheap, readily available thesaurus software packages. It is the micro which opened a hitherto unknown perspective of a thesaurus software which could be used as a tool for IR in the same way as text processing packages are used as tools for computer editing. And it is the micro which, by its universal use and ubiquitious availability, underscores the need for application in much broader fields than just STI.

Some of the main lines of micro-based systems are being reflected in the few samples of exhibits on show. On a whole, it can be stated, that, as a result of the micro, thesaurus software has become as ubiquitious as the micro itself, and that prices of software are low.

Micro-based thesaurus programs, as well as some systems demanding a more powerful IR environment will be demonstrated and discussed in this meeting. Later on in the Congress, some more special questions will be discussed, like superthesauri and cognitive aspects.

Nevertheless, all the world-wide, ever-growing activities together, based on such systems, do not explain the current thesaurus renaissance. There is more about it.

Question No. 2: *What is the real point of attraction? It is Conceptual Structuring. This is needed for AI, and Language Understanding in particular.*

Principles applied in thesauri increasingly are becoming of interest to solve problems in computer linguistics, or, more precisely, Linguistic Engineering (LE), which is a special branch of applied Artificial Intelligence (AI). LE is the now fashionable name for machine or machine-aided language processing.

In this context, IR represents just one (and possibly not the most important one) out of a series of target areas of LE R&D, such as Natural Language Interfacing, Machine Translation (MT), Machine Abstracting, Speech Recognition, and Language Understanding.

This does not mean that existing thesauri as of today can be used right-a-way in a single one of these new areas of machine or machine-aided, language-related problem solving.

Yet, in traditional thesauri applications for IR it could be observed that thesauri tuned to the needs of very special domains of knowledge, and very special IR tasks, performed better than those in broader domains and general-purpose environments.

The real point of the newly-risen interest in thesauri are some of their main functions, such as:

- Mapping of a term's meaning (semantics)
- Mapping, and selective definition, of semantic and other relations between different terms of a natural language
  (Both is needed to establish a conceptually defined space in terminology-based systems - a prerequisite to language understanding and ordering systems)
- Establishment of Preferential Terms, and, along with it:
  1) Standardization of term use in closed language systems
  2) Improvement of predictability of term use in such systems
  3) Definition of semantic and other relations between preferential terms (system language) and terms of the natural language
  4) Definition of semantic and other relations between preferential terms (system language) and terms or values used in artificial languages, like query languages, or ordering systems, such as notations and classifications, statistically or otherwise defined terms and expressions.
- These functions with regard to terms of more than one single natural language.

In a rather rough statement one could say that it is now the linguists, who on the one hand, need thesauri, or thesaurus-like conceptual structures, to solve their problem of meaning, i.e. the problem of language understanding, whereas the IR systems designers, in turn, finally came to know that basic LE is required in their systems to come up with more efficient, intelligent, machine-aided IR systems.

Moreover, thesauri up to now have been designed for use by human indexers and data bank searchers. This concept is being challenged by the machine since some time.

Question No. 3: *If LE can profit from conceptual structures as are thesauri, can't IR profit from machine-processing of the language? The answer is: Yes.*

Computers have become capable of handling much larger and much more complex relational systems, and therefore can be expected to brush up themselves with good or fair chances of reasonable results. This suggests that it may now be possible to leave to a human specialist's intervention only the more tricky cases, - interventions like disambiguation, word selection, etc., which then would have to be done by interaction.

So, the question being posed today is this: How must a machine-operated thesaurus look like, and how can it be built and maintained, if it is to meet the needs of such machine or machine-aided natural language processing - among others: IR?

Before turning to my 4th question, the rather rough sketch of the situation outlined above needs some more detail and refinement.

When I referred to "a term's meaning", this may have been typical of an earlier state of the art, suggesting that it is viable that a given term can be defined as the bearer of all its possible meanings, and that these meanings can be adequately expressed by other terms of the natural language.

After more than two decades of discussions on "uniterms", composite terms, noun phrases, prepositional and other logics, as well as all sorts of term frequencies in texts and text collections, it has become clear, that properties of terms and term use in more complex statements like phrases must be given much more attention, and that considerably higher grades of definition must be applied, if the goal of machine-operability of natural language text in IR is to be achieved.

The requirement then draws very close to what is needed in MT, and consequently, a thesaurus then probably is about to look very much like a machine-operable dictionary. At any rate it is clear that the old type of thesaurus would not fit for the purpose.

It should not be forgotten that IR since its very beginning (and this remains as a heritage from the pre-computer era), has always been working with the assumption that an abridged terminology only was needed to "just find the document", and that it was definitely not anything comparable to the terminology of the much more refined one-by-one-procedures to be applied in other fields of computer linguistics, like MT.

It went without saying that indexing then also should be less complicated, and cheaper, than lingustic one-by-one-processing.

As an example, classic thesauri do not make a distinction between plurals and singulars, they just know the singular form (with some few exceptions, as a rule). Verbs, the very lifeblood of natural language systems, don't occur in them. Among the many semantic relations that exist between single terms and the different ways of expression, just some of the more important ones are considered, like hierarchy, partitive, resemblance. Only very few among the thesauri applied in practice feature syntactic rules in a proper sense, the majority of them being geared to the rather poor conditions of coordinate indexing.

It is true, that, for the sake of better predictability - on this we will have to speak later on in this Congress - with all this stripping in our thesauri, we effectively brought down the articulative power of the artificial IR languages to almost zero.

Of course, it is understandable, why we took this approach. It earned us much criticism, and left Free Text procedures appear more attractive. However, and this is:

Question No.4: *Can we do better now? The answer is: Yes, we can.*

In the light of the new requirements and systems possibilities, this approach must and can be corrected. It can give us a better grip on information and warrant an

improved organization of knowledge contained in our documents.

Almost twenty years after its introduction to the information world, it has become obvious that the early, naive "Full Text" approach has not kept its promise. Nevertheless we have to state that it has made heavy inroads. Highly specialized, ever more expensive labor needed for tiring indexing tasks had come under heavy pressure from ever cheaper, fully automated text indexing procedures, with electronic typesetting as the real trigger. The result is known. The tacid reserve, if not open rejection of the electronic "Full Text" collections by their prospective users is a matter of fact. Garbage databanks widely have discredited the young information services market.

Question No.5: *How can we do better? We must machine-process the natural language text to condition it for IR.*

Indexing has to be reconsidered. The integral text in a machine-readable form is there. The computer power to swiftly process even very large corpora of text is there. Since expensive, highly-trained intellectual labor remains scarce, it is a must that it be concentrated on true problem-solving in machine-aided, interactive procedures. No longer can it be wasted for routine jobs and repetitive work. The machine is powerful enough to take over these functions. In the now-emerging post-"Full Text" era, the integral text, or parts of it, remain the basis of processing in so far as it is natural language text that is treated, but this natural language text must be processed by methods of true LE. Playing around at random with single words the candid way should stop. The language engineered programs must get the appropriate tools to do the basic indexing in the best possible way. The tools in question, whether integrated in other dictionaries or not, are the thesauri, and these will be the *Thesauri of the Second Generation.* In them, the IR language will be much more refined.

Such thesauri will also be useful to solve other essential functions in IR, like machine-aided query understanding and formulation, machine-aided abstracting, machine-aided text evaluation and selection.

Question No.6: *When does fully automatic IR arrive? Probably never.*

Above all, we have to give in to the evidence that the best we can achieve in this effort is machine-aided IR. We should write off all hope to see "Fully Automatic" IR systems successfully at work, at least for the next ten years or so...

So, thesauri must be made fit for that purpose.

Question No.7: *Thesauri for machine-aided IR - how do we get there from here?*

There are several different ways to achieve this goal. Either (1) the thesaurus with all necessary definitions and relation work is made part of a larger lexicon which serves the broader and more basic LE tasks, like parsing, or MT. In this case, it must be defined which entries are needed for which purposes of IR, e.g., indexing, query formulation, abstracting, text or statement selection, etc. An interface must be agreed upon to enable the IR machine to successfully brush up in the thesaurus entries of the dictionary. Both, the IR machine as well as the natural language text processing machine, would refer to the same dictionary, whereas the IR machine would start working on the basis of the text analysis results rendered by the natural language text processing machine. The IR machine then would control the different interactive processes necessary to solve the IR task.

Or, (2), an interface to the LE machine is agreed upon to condition both, the handling of the thesaurus entries by that machine, as well as the processing needed for the different functions of the IR system, in conjunction with other dictionaries needed by the LE machine.

Another (3) way may consist of an integration of a thesaurus in other more special text processing systems, like Hypertext. The thesaurus then would guide the system in choosing or proposing the appropriate terms for referral to the source texts. This also raises the question of where to instal the interface, and which properties are needed.

That (4) a thesaurus structure can be formulated in a frame system as used in a more general AI environment will be demonstrated later in this Seminar.

Question No.8: *Which is the right way? Which is the model? What to standardize? If we only had the answer to this!*

Most of the thesaurus software on sale over here in Europe conforms with the traditional pattern of thesauri as outlined in the classic works by Soergel, Wersig, Aitchison/Gilchrist as national and international standards, which, by the way, were only recommendations. They were available at a relatively early stage of computerization in the information sector.

It was this Thesaurus Committee, the sponsor of this meeting, which - as early as 1965 - started out with defining a thesaurus and drafting a guide for the construction of thesauri which later on was compiled by D.SOERGEL in his first book on thesauri (1969, in German) which successfully laid the ground for today's standardization.

After some earlier work, the German pre-standard DIN 1463 was open for public discussion in 1972 and emerged as a full-fledged standard in 1976. "Concepts and Terms, General Principles" (DIN 2330) were presented in 1974, and "Systems of Concepts and their Presentation" (DIN 2331) in 1976. All major standards, including AFNOR Z47-100, BS5723: 1979, ISO 2788, all on monolingual thesauri, had been issued up to the second half of the seventies. The discussion on multilingual thesauri was on in the middle of the seventies (ISO/TC46/WG5), and "Rules for Building Multilingual Thesauri" (ISO 5964) was out in 1977.

Nevertheless, it took some years for the first packages of isolated software to appear on the market. Admittedly, the overall development in computer hardware had some influence in this, but the process shows something

about how much time it takes for a standard to come to application in readily available market products, and it also testifies the overall importance of standardization in this very special field.

A new round of standardization is now required, if second generation, machine-aided IR one day is to be triggered off. The process of elaboration of the new thesaurus standard could follow the lines exposed above. Since the IR procedures to be affected by such standards will be highly automated, and will cover large areas of application, substantial investment is at stake. The money would not flow before the standards are there.

Question No.9: *Can the IR people do it alone? The answer is: No. Teaming up with specialists from other disciplines is necessary.*

But before working on standards, we must know what we are going to standardize. The Thesaurus Committee wants to discuss this with all interested and potentially important partners. Somebody would have to come forth with a model, or a set of possible models, from which practical work can start. We know that the computer linguists are faced with a similar problem in their search for machine-operable, standard lexica. We suggest teaming up with them to discuss both our matters. We are eager to present our case and not to let that chance pass by.

Early in the seventies, in our footnote on the validity of our Thesaurus Guidelines, we wrote: "This standard is not valid for the construction of thesauri in the sense of Linguistic Science (e.g. synonym dictionaries)".

We are now discussing the new challenges which lie ahead, and we are prepared and willing to help elaborate and promote the models and standards needed for IR in more complex LE environments, being fully aware of the fact that all this, of course, is linked to, or part of, the Linguistic Sciences (as well as it is part of other disciplines, like logic).

The Committee will be trying to establish contacts to all appropriate groups or bodies dealing with the question of how to normalize dictionaries, terminology formats, etc., and it intends to discuss with them what thesaurus theory and practice can offer to machine-aided IR in LE environments.

It is obvious that thesauri of that kind will also be useful for other tasks of LE, such as speech recognition, and text understanding. With our knowledge of what can be achieved by means of thesaurus systems, we may be in a position to stretch out a helpful hand to what until now was at best a neighbouring discipline, - in exchange for other basic language technology which we feel is badly needed on our side, and of which we know it is obviously at hand.

Question No.10: *Advanced information services with a truly human interface? Not without thesauri.*

Most certainly, machine-aided LE will play a major role in overcoming the language barriers in our future, European Single Market, which, by the addition of the Eastern countries, will appear even more Babylonic today. The availability of native, natural language information services in the different member countries, including also the smaller ones, will become a vital issue in this context, and it is safe that this cannot occur without the availability of appropriate multi-lingual thesauri of the second generation.

As the present chairman of the German Committee for Classification and Thesaurus Research I am particularly glad you all came here to participate in this Seminar, and in the name of all our members I give you a warm welcome to this meeting. We will be most pleased to hear your comments and suggestions, and we are hopeful that at the end of the day we will know better how and how best to reach our common goal.

This common goal, needless to say it, is: Adequate tools for improved, intelligent, machine-aided, monolingual and multilingual IR.

Dr.Winfried SCHMITZ-ESSER
Oderfelder Str. 13, 2000 Hamburg 13

International Conference on Symbolic - Numeric Data Analysis and Learning

From 17-20 Sept. 1191 The Institut National de Recherche en Informatique et en Automatique (INRIA) will hold its next conference at the Université Paris Dauphine with English and French as conference languages. Proposals for papers (in four copies) of 12 pages max. should be submitted by Nov.30, 1990. The topics include: Data Analysis, Machine Learning and Modelling; Clustering ordering, distances; Representation, analysis and synthesis of symbolic and numeric knowledge (structured, noisy, uncertain, etc.); Symbolic - numeric induction, knowledge acquisition from data; Formation and recognition of conceptual structures: discovery of laws, rules, inheritance trees, decision graphs, lattices; Neural aspects; Coherency, stability, and validation of results; Software, and Applications. For further information contact: INRIA. Service des Relations Exterieures. Domaine de Voluceau - BP 105 - Rocquencourt, F-78153 Lc Chesnay Cedex, France.

Cluster Analysis in Chemistry

The 1990 meeting of the British Classification Society will be held at the AFRC Institute of Feed Research, Shinfield, Reading, on Oct.23, 1990. It is a joint meeting with the UK Chemometrics Discussion Group and the Multivariate Study Group of the Royal Statistical Society and covers the topic "Cluster Analysis in Chemistry". The following five papers will be presented: Nick BRATCHELL: Review/Tutorial. - Simon PACK: Applications of cluster analysis. - Mandy PARSELL, Steve ELMORE: Cluster analysis for sample selection. - Mike ADAMS: Application of cluster analysis to infra-red. - Dave LIVINGSTONE: Applications of cluster analysis in QSAR and molecular modelling. - For further information contact: Dr.S.E. Hitchcock, Secretary, Brit.Classif. Soc., The Open University, Faculty of Mathematics, Walton Hall, Milton Keynes MK7 6AA, England.