Thesaurus and Beyond: An Advanced Formula for Linguistic Engineering and Information Retrieval

Winfried Schmitz-Esser

University of Applied Sciences, Hamburg, Germany

Dr. rer. pol. Winfried Schmitz-Esser, Diplom-Volkswirt, is a professor at the University of Applied Sciences in Hamburg, Germany, in the chair of „Mediendokumentation", a new discipline covering Media, Archives and Libraries. His early career was that of journalist and editor. A pioneer and adviser in IR applications, he is chairman of the German Thesaurus Committee, the birthplace of the present paper. He is a founding member of ISKO and serves as chairman of ISKO's Scientific Advisory Council and as the Council's representative on the ISKO Executive Board.

Schmitz-Esser, W. (1999): **Thesaurus and Beyond: An Advanced Formula for Linguistic Engineering and Information Retrieval.** *Knowledge Organization, 26(1).* 10-22. 2 refs.

ABSTRACT: This paper describes a proposal for a new approach to thesaurus design and construction that could have significant implications for change in the way multilingual thesauri are handled and integrated with each other. The formula presented here has its origin in the work of the German Thesaurus Committee and has had input from a number of scientists and practitioners in the field. The emphasis is on the various types of relationships found among concepts, notions and universals in languages. These relationships are analysed and refined beyond the approach taken in existing thesauri. This proposal is very much at the discussion stage and the author invites the assistance of interested readers through criticisms, discussion and dialogue. Applications of the proposed thesaurus are included and the major goal of this proposal is to provide the basis for improved design and integration of multilingual thesauri.

Preamble

Knowledge spaces as tackled in modern linguistic engineering, artificial intelligence work and the like usually reflect mini-worlds, whereas, where larger worlds have been successfully defined, they are typically found to be highly specialized. In theory there are many universal approaches, but looking at the practical side, partial, purpose-bound solutions have cropped up at best. The inter-language complex is mostly neglected and all of these approaches are lacking general acceptance.

It is evident that such a dilemma is due to unsolved problems in achieving valid definitions of relationships – those between language and thought on the one hand, and those between thought and instances (i.e. anything that has a name) on the other. Work on this frontier could be dramatically eased if we had terminological lexicons with such definitions. Then we could use them as knowledge banks in many applications. They would have to be encyclopaedic and universal, and, given the immensity of the task, would have to be constructed and continually updated by a multitude of contributors (i.e. in a distributed way).

But what should the structure for such a *machine-readable, linguistic, plurilingual, lexicographic, universal and domain-independent thesaurus* look like? This was an open question a decade ago. Since that time, the German Committee for Classification and Thesaurus Research (KTF) of the Deutsche Gesellschaft für Dokumentation (DGD), Frankfurt, including a number of scientists and practitioners from ISKO, has been discussing this question in a vivid, and sometimes even controversial, way. The following paper owes much to these discussions. Credit, however, must also be given to many other colleagues and friends with whom I have had the privilege of exchanging ideas on the subject.

In quite some depth, this paper is an outline for the structure of a new type of thesaurus which I think could be useful in some classic areas of linguistic engineering, such as machine-assisted translation, abstracting, and information retrieval. It is felt that it could also open up new frontiers towards new more animative, sensitive, surprising, fun and playful approaches to information. That this side of information science has been largely neglected so far is well known.

This paper is presented here to the readers of *Knowledge Organization* so that they may consider its ideas and respond to them. In doing so, they are cordially invited to put forth their comments, suggestions and criticisms. All contributions are very much welcome.

It is hoped that, after due discussion on an international level, this paper might lead to a text which could be made as a recommendation for the construction and maintenance of such thesauri. It is important for the reader to understand that the text presented here is not yet that final text. Rather, this presentation is meant to be a first description of the model with the aim of explaining and encouraging such a solution. The paper also describes the basic mechanisms by which a thesaurus of this type could be produced in a distributed way.

On a practical level, advantage could be taken of an opportunity to put some basic elements contained in this paper to a larger field test. This was possible along with the preparations for the EXPO2000 World Exhibition to be held in Hanover. Reports on this project were given by the author in two separate papers. One paper[1] (describing the goals and intentions) of the project was presented at the Conference of the French ISKO Chapter in 1997 in Lille, and the other[2] (results) at the ISKO International Conference, one year later, also in Lille.

On a scientific level, the paper reflects the dispute between two lines of thought: the computational approach to the problems of intra-language, interlanguage expression and thought on the one side, and a more phenomenologic, constructivistic and dynamic approach on the other side.

The thesaurus model outlined here is clearly onomasiologic. It gives way to a definition of the phenomena of the world (fundamentals) in such a way as the one who defines them sees them. There is no limit as to what subject can be defined. Thus, the model is open to new definitions and the work of updating by anybody at any time, while allowing its use without domain restraint under closed world assumptions (CWA). Entries in this thesaurus do not impede more detailed work on an algebraic or linguistic level. The model also tolerates different views and even contradictions. All of these are properties not much different from those offered by language where the power to express thought is only limited by what the paradigm allows to be expressed in words and phrases. The concept also removes the idea that systems must be complete in order to work with them.

The price for this, evidently, is limited power of articulation, and the question now is whether the level envisaged is considered to be useful and valid for a substantial number of different real world applications. This can only be determined through discussion among scientists and all persons otherwise concerned, and on an international level. All criticism and suggestions, therefore, are warmly welcomed. Please send your message to the author's address (at the end of this article).

The Formula

As indicated in the above preamble, this paper outlines the elements and structure of a novel, advanced thesaurus format by which it is hoped that some barriers encountered so far in artificial intelligence, linguistic engineering and information retrieval could be overcome. It is suggested that a corresponding recommendation should follow the ensuing lines:

1. Two different classes of relations: Concept-term, and concept-concept

Looking at relations as they exist between different objects of thought (concepts, notions), and given the fact that in most human communication such objects are expressed by means of some terms in a natural language, a problem arises in addressing a given concept in different languages, but can be overcome. Even if a concept is unique and therefore can best be expressed in a distinct language, for example "fado" in Portuguese, or "Parteienfilz" in German, some possibility for clearly expressing that concept in other evolved languages should always exist, no matter how many words may be necessary for this. Then, if this is true, it must be viable that, in a multilingual thesaurus all concepts addressed can be made subject to the stipulation of conceptual interrelations according to one and the same set of different types of relations. Of course these must be "universals", well defined and, as much as is possible in practical work, be free of overlap and intersection.

In this thesaurus format, relations pertaining to this set, or class, irrespective of a particular language, shall be called "Concept Relations Proper" (CRP) or, for reasons to be explained later, Class II relations.

There is another class of relationships which differs from these Relations Proper insofar as the term-concept relation involved may only be found in one individual language. In a multilingual thesaurus, this class of relationships therefore must be stipulated language by language. These are the Class I relations. There are two types of these:
Synonyms
Polysemes

In the case of synonyms, two or more expressions of an individual language have the same meaning. In the case of polysemes, one expression of that language has more than one meaning.

Table 1: Class I Relations valid for each language separately

<table>
<tr><td><b>Equivalence</b></td><td>US
use preferential
Synonym

<i>Reforestation</i>
<i>US afforestation</i></td><td colspan="2">UF
preferential Synonym
used for

<i>afforestation</i>
<i>UF reforestation</i></td></tr>
<tr><td rowspan="2"><b>Polyseme</b></td><td rowspan="2">UD
use Descriptor

<i>bank</i>
<i>UD credit institute</i>
<i>UD riverside</i></td><td colspan="2">UF
Descriptor used for
(among others)</td></tr>
<tr><td><i>credit institute</i>
<i>UFbank</i></td><td><i>Riverside</i>
<i>UF bank</i></td></tr>
</table>

2. Descriptors for ambiguity control

For reasons of practical manageability and systems control of an already complex thesaurus system, it is a prerequisite that a single given concept in the thesaurus be addressed by a term, or a sequence of terms, in such a way that this term, or sequence of terms, is unique in the system. Also it must clearly express the meaning of the object of thought (concept, notion, universal), and by no means any other object of thought, in that language. Such a term – unique in the term system – is called a Descriptor (D). In a chain of expressions of equal meaning, this may be a preferential expression as known from traditional thesauri. Perhaps, the only difference is that what is considered "equal in meaning" should be focussed more precisely under the new format, and that "quasi-equivalencies" should not be admitted.

All other terms, or sequences of terms of equal meaning are "Additional Access Expressions" (AAE). It should be a policy in the practical construction of a thesaurus of the proposed format, that as many AAEs are considered as are known from common communication in the respective languages. Descriptors as well as AAEs are altogether "Access Expressions" (AE). There is no term length restriction in AEs. In the stipulation of a descriptor, the prime requirement of both a clear denomination of the object of thought and non-ambiguity in the respective language must be the absolute priority over shortness.

3. What about the Intermediate Language?

As to the Class II relations (Concept Relations Proper), in any multilingual thesaurus there would be no need to provide the same definitions in each of the individual languages considered. It suffices to do it in one language, which then works as a middle language (intermediate, or source language). The other languages would be formally dealt with as target languages. However, distributed work on such a thesaurus that would put one language in the middle would be affected by some obstacles of political correctness. As seen from an aspect of logic, however, it is not necessary to have a natural language as a middle language at all. A numbering system would do perfectly. So, it is proposed here that a Meta Language Identification Number (MLIN) be assigned to each Equivalence Chain of Descriptors (ECD), and that the Concept Relations Proper (CRP) be formally applied among MLINs.

To give an example for this: If the chain of descriptors would read:

	English	French	Spanish	German
ECD:	*airplane*	*avion*	*avión*	*Flugzeug*

then to this entire chain would be attributed a distinct MLIN, a number which, of course, must be unique as an identifier in the system, and the relationships to other objects of thought (ECDs) represented by their respective MLINs would be stipulated on the basis of their ECDs.

(Figure 1: The Relations Proper). Any relationship of the Class II type stipulated, on the basis of any of the languages considered in a given thesaurus, would then be available to all the other respective language entries, in as much as the respective chain is complete.

Individual thesauri following this formula could be integrated provided they are at least bilingual in such a sense that at least one of the two languages of that thesaurus matches with the language pattern of the receiving, multilingual, thesaurus. This, then, would lead to provisionally incomplete ECDs, the missing parts of which could be taken up and subsequently filled and completed by other partners contributing to the implementation of the thesaurus.

Monolingual thesauri, otherwise following this format, could also be used to enhance the multilingual thesaurus in such a way that those of the Additional Access Expressions (AAEs) which have not been considered yet are imported to the receiving thesaurus. Then the aim is to enrich the overall number of AAEs in that particular language.

The thesaurus software system must be constructed in such a way that a given AAE of a particular language admits an entry to be left open for later definition and stipulation as a polyseme. Once the polysemes are sufficiently defined, they can be used in the respective IR systems to guide users interactively during their searches. Or, context related procedures might be applied in indexing or search operations for machine-aided disambiguation.

4. The Expressions

Expressions entered in this thesaurus format may have any length provided they do not contain subordinate clauses. They must be part of the respective natural language and should reflect this language's paradigm as used in general communication. Expert vocabulary should be integrated, and insofar as this fulfils the prime condition of clear, unambiguous denomination of a concept, this may prove especially welcome to serve as a descriptor.

Where taxonomies exist (e.g. for animals or plants) they should be checked as to whether they fully comply with the Abstract/Generic conditions set out in this paper. Then they may be especially recommended as descriptors. So, in the example given for "birds" (further down) it might prove advisable to replace it with the scientific term "aves", and take "aves" as a descriptor for those languages where this term is unequivocal. There is no need to choose a more frequent word as a descriptor just for the sake of higher frequency in common use. Such a word as "birds" would be guaranteed as an AAE in any case, be it in its quality as a synonym, or be it as a polyseme. This, by the way is the case here: "birds" may also mean "sounds of disapproval". Special but uncommon terms chosen for clarity could be earmarked in such a way that in modes of presentation to a general public the user is offered the most common expression, as in the above-mentioned case of "birds".

No rule can be given as to the recommended composition of an Access Expression (AE). This may be a single term like "aves", with the quality of a monem (one term, one semantic root). It might be a synthem (one term, more than one semantic root), or a pair of synthems or monems; or an even higher synthemised expression, which mostly will come as a noun phrase, thus indicating a whole theme. One reason for this is obvious. It is known that what appears in one language as a more than one word expression may be expressed in one single word in another language. An example of this would be the French "pain cuit au four de bois", which in German has its equivalent as "Holzofenbrot". Since, in a way, a high degree of synthemisation reflects high validity and applicability of a thesaurus (provided the constituent synthems and monems are equally open for search), such specific expressions are highly desirable in a term collection of the proposed type.

Also, in field application, it was found that more and better defined types of relationships between concepts can best be applied, and exploited, with a terminology which is highly synthemised. This means that terminology describing a complex concept tends to comply better with the requirements for a plurilingual set of descriptors than with lower synthemised terms, not to speak of monems, which often display characteristics of polysemes.

5. Names and Acronyms

Names and acronyms are determiners for phenomena unique in space and time, i.e. instances. These instances may be living beings (persons, animals), or institutions or groups of persons by force of law or otherwise sanctioned by society. They are included in this thesaurus format and formally dealt with in the same way as descriptors, but marked for entry/use in column 2 (name) or 3 (event) of what is called here the *Basic Semantic Reference Structure (BSRS)* (Table 3 below). That means, above all, that the full range of relations among them is applicable with one exception – the Abstract/Generic relationship. Acronyms are treated as AAEs.

6. Earmarking of Descriptors and Names

To achieve full semantic control in a plurilingual environment it is not only necessary to define the meaning of an expression, but also to define the type of use which one is going to make of the expression, since this also affects the expression's meaning. This is why each Equivalence Chain of Descriptors or Names must be earmarked in such a way that its intended/possible use becomes clear. The respective entries refer to one or more classes of elements (columns) of the *Basic Semantic Reference Structure (BSRS).* As depicted by the columns in Table 3, Applications Case A, this reference structure features a total of eight different types of use of Descriptors' Names and other determiners.

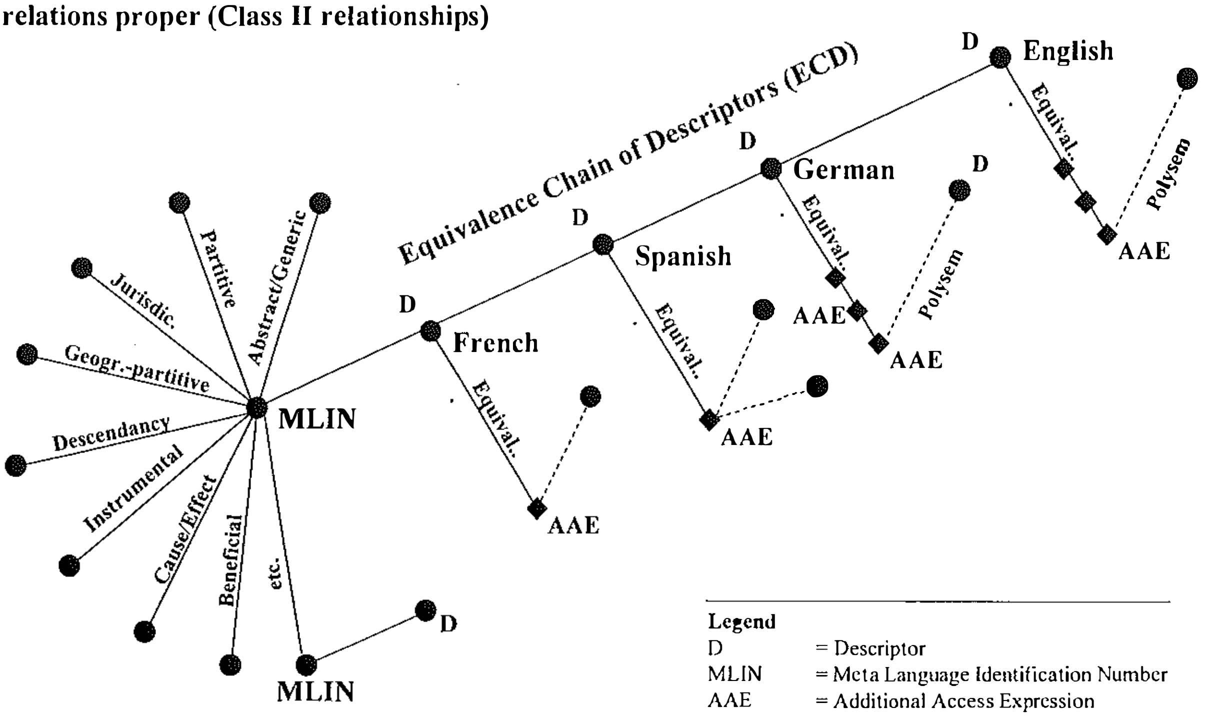
Descriptors entered in the concept column (1) of the *BSRS* generally appear in their plural form, so as to best express their nature as object classes. Entries to be posted in columns 2 and 3 (instances) may take any form provided it is encountered in reality and in the respective language. This will be mostly singular. Entries to be posted in columns 4 and 5 (location) are singular, except for words which only exist as a plural form, like "Azores". Aspects (column 6) are presented in their canonic form as proposed in the table. The two final columns (7 and 8) are foreseen as determiners of time.

7. The Data Model, Principles of Thesaurus Construction and Interchange

The data model for the representation of all thirteen relations is the same:

A relationship exists between two given objects whereby the relation is determined by two elements: (1) the type of the relation, and (2) its direction.

Figure 1:
The relations proper (Class II relationships)



Each object is represented by a Meta Language Identification Number (MLIN) attributed to an Equivalence Chain of Descriptors (ECD). The ECD has a natural language expression in each language considered. In this language, it denominates the object represented by the MLIN in a clear and unambiguous way.

The meaning underlying the direction is determined by the respective relational definition given in this format in Table 2. In such a way, it is possible to express, for example, what is part of what, what is detrimental for what, or what are specific items of a broader concept. The known generic hierarchy is thus seen as just one case out of the thirteen other relations of this proposal, of which some may still be regarded as hierarchies, some others, in contrast, in a totally different way.

8. The Class II Relationships

Responding to the obvious need for more, and better, refined relations, the idea of (unspecified) "Related Terms", one of the basic pillars of traditional IR thesauri, is dropped. One other critical point about traditional thesauri is their poorly defined hierarchies that allow for different types of relationships to be accommodated under one hierarchical roof. As a rule, then, the generic relationship, and various forms of partitive relationships encountered are found mixed up.

The thesaurus proposed here distinguishes thirteen different relationships proper, or class II relations. Other, more specific relationships may be added as standardised options at a later date, as need for them appears. Each of them comes with a definition as well as some explanation clarifying the boundaries between them and indicating some rules of applicability. Not every application in IR or linguistic engineering will call for a full set of these relationships to work with. This is why they should be given a code number allowing classification of an individual type of thesaurus. In a way, this would indicate how powerful a given thesaurus is as a terminological tool. But likewise such coding would be useful in case of interchange or amalgamation with other thesauri following this format.

Table 2: Class II Relations valid irrespective of an individual languages of the thesaurus

1. **Abstract/Generic**	BC broader concept *coconuts* *BC vegetal products*	NC narrower concept *vegetal products* *NC coconuts*
2. **Partitive** *Physical and theoretical*	PO part of *private sector* *PO economy*	HE has elements *economy* *HE private sector* *HE public sector*
3. **Part/Whole** Law & Jurisdiction	PT pertains to *France* *PT European Union*	CA constituents are *European Union* *CA France* *CA The Netherlands*
4. **Geographic-partitive** *Geographical, topographical,*	SO is space of *South America* *SO Latin America*	CS consists of spaces *Latin America* *CS South America* *CS Central America*
5. **Descendancy**	DF descends from *father* *DF grandfather*	PC is precedent of *grandfather* *PC father*

6. **Instrumental**	IF
	is instrumental for
	torch
	IF welding
		BI
		by instruments
		welding
		BI torch

7. **Cause/effect**	CE causes effect of *wood-slashing* *CE desertification*	CB caused by *desertification* *CB wood-slashing*
8. **Beneficial**	BF beneficial for *tree planting* *BF water balance regulation*	PF profits from *water balance regulation* *PF tree planting*
9. **Detrimental**	DT is detrimental to *overfertilization* *DT biotopes*	HB harmed by *biotopes* *HB overfertilization*
10. **Matter**	MO is matter of *iron* *MO Earth core*	CO consists of *Earth core* *CO iron*

11. Form and appearance	AA appears as *portal* *AA Roman arc*	SA shapes the appearance of *Roman arc* *SA portal*
12. Process	**PU** process applied in *progressive assembly line* *PU production*	**IP** involves process *production* *IP progressive assembly line*
13. State	FE is form of existence of *ice* *FE water*	EF exists in the form of *water* *EF ice*

9. Definitions

9.1. Abstract/Generic

The abstract or generic relationship applies to a constellation where concept A is a broader concept of concept B. Both concepts are conceived as classes, each member of the class featuring the same constituent criteria of the respective class. Two classic conditions must be fulfilled:

First condition:	B *is* A *(in an abstract or generic sense)*
Second condition:	B is a specific kind of A *This means that there is a „differentia specifica" (at least one) giving rise to a meaningful differentiation between the two. As a rule, other concepts like C, D, E, etc. exist, all complying with the first condition, but each with a specific „differentia" or more.*

Examples:

A	birds
B	birds of prey, whereby
C	may be water birds,
D	may be podicipediformes, etc.

The direction points to that which is considered to denominate, or give a name to, the broader of the two concepts in question.

9.2 Partitive

The partitive relationship is a part/whole relation existing and/or discernible between two given concepts. It is applicable to:

- all physical entities and their constituent parts, e.g. "book"/"pages of a book", or "automobile"/"radiator cover";

but also to

- objects of thought, e.g. "national economy"/"private sector", "public sector", or "industries"/ "chemical industry".

Since borderlines are fluent between reality as it exists and reality as it is individually or socially perceived, a distinction between physical whole/part relationships and relationships which are a product of thought seem misplaced. Nevertheless, care must be taken to distinguish this part/whole relation from the two special relationships in 9.3 and 9.4 below.

If one has to express an aggregate concept like „social and economic development", here is the solution. The direction points to what is considered the entity between the two concepts in question.

9.3 Part/Whole relations determined by Law and Jurisdiction

An object of thought B is a part of A as a consequence of private, public or international law or jurisdiction. This is the relation fitting for such a case. It will apply mainly to cases such as the following: "Bezirk Eimsbüttel" is part of "Freie und Hansestadt Hamburg", or: "Spain" is part of (a member of) "NATO", or, to give a third example, the "BBC World Service" is part of (a division of) the "British Broadcasting Corporation".

Objects to be linked by this type of part/whole relationship will necessarily be actors with a defined status as a "subject" of private, public or international law. In the *BSRS*, these actors appear as instances grouped in element 2 of this scheme. The direction of this relationship is toward the term or concept that is considered the entity between the two instances in question (entity is A).

9.4 Geographic-partitive

This is a special relationship dealing with geographic wholes and parts. It is designed to serve as a normalised specifier in elements 4 and 5 of the *BSRS*. Seen under this viewpoint, B is a geographical part of the whole A, as in the case of "Massif Central"/ "France", or "São Miguel"/"Azores". If the geographical part stretches over two or more other regions which are also parts of the same whole, as in the case of "Alps" (Europe), the nearest whole (Europe) covering the parts would have to be stipulated as a whole. Except for rare and special cases, states down to countries (to be dealt with under the Part/Whole relation as defined by Law) are normally considered as countries, i.e. in the meaning of their geographical dimension. This is why a parallel entry of an object pair may be fully justified. This entry, by the way, normally will turn out to be much shorter, as is exemplified in "Eimsbüttel"/"Hamburg" (see above), or "Hamburg"/ "Germany". Again, the direction points at what is considered the entity.

9.5. Descendancy

This relation expresses the relation in which concept B descends from concept A. The descendance may mean:

a) a genetic predecessor relationship (e.g. "Son"/ "Father") or
b) a state before/state after relationship in the sense that state B is derived from A in a process (e.g. "film paper copy"/"negative film");
c) any other entity/predecessor relationship (e.g. "OECD"/"OECE", or "butane"/"crude oil").

The relation points to the respective predecessor.

9.6 Instrumental

This relation expresses the fact that concept B is instrumental to achieve, as a result, concept A. For instance: "screw"/"assembling"; OR "torch"/"welding". It is important to note that the instrument considered may be one applied by a living being (man, animal), or a machine, or a system. The sense of the relation points to the result achieved/aimed at by use of the instrument.

9.7 Cause-effect

This is the case where concept B causes concept A to happen, as, for example, in a reaction, "explosion"/ "abrupt generation of energy", or "wood-slashing"/ "desertification".
The direction points to the effect produced by the cause.

9.8 Beneficial

This is a relationship that indicates that concept B is beneficial, or useful to concept A. The underlying values of what can be considered as "beneficial" in a universally accepted sense should be linked to the results of the international discussion on values. One sizeable result of such world-wide discussion is the Agenda 2000 of the Rio Earth Summit of 1992, which was approved by more than 130 governments. It posted as a focal target the value of "Sustainability". Sustainability as it is reflected in this Conference, and which is still being reflected in the consecutive world summits and meetings, could well serve as a guideline for the Beneficial Relationship presented here.

The validity of such an interpretation is obvious. Thus, a relationship such as "fish staircases"/"protection of living marine resources" may precisely indicate the desired intrinsic goal of what fish staircases are good for. As a rule, however, recurrence to such high-level and world-wide accepted standards of values wouldn't be necessary in many normal, everyday cases, such as in the statement "alphabetisation" is useful to "higher quality of life".

Any involvement of particular, vested or aspired, interests, however, or the consideration of short-lived effects, must be avoided in applying this relationship. Otherwise, one would end up with statements such as "overfishing" is beneficial to "fishery industries", which obviously is not true when seen under the aspect of sustainability. In the object space, this relationship points in the direction of the term standing for the desired, positive effect.

9.9 Detrimental

From a point of logic, the introduction of a special Detrimental relationship may appear superfluous at first glance when "Beneficial" exists and the thesaurus constructor is free to formulate a descriptor with a built-in negative value. Experience, however, showed that it could be most useful in openly addressing negative effects. That "overfishing", for example, has a detrimental bearing on the "diversity of marine life", is not disputed. With this relationship one can demonstrate whole chains of adverse causes and effects, from, for example, "overfertilization" to "dying biotopes" to "dearth of fish" to "endangered biodiversity". An extended interpretation of this relation which may turn out as even more fruitful in practice could be a general "adversity".

In accordance with the scheme applied above, the arrow points to the end of what is being endangered.

9.10 Matter

Then there is a special relationship indicating that concept A consists of concept B when seen under an aspect of matter or material used. So, a "conveyor belt" may consist of "rubber", or one may have a reason to state "Earth core"/"iron".

If it has to be expressed that the Earth core consists of still other matter, this can be made the point of a parallel entry. There is no possibility of indicating the percentages of each constituent matter in a whole. So, when one reads "photographic film"/"silver" this does not mean that such a film consists mainly of silver, but that a certain amount of silver, in fact, is a constituent. "Matter" considered in this relationship may be chemical elements, compounds or any kind of material, be it natural or synthetic. Excluded are constructions and artefacts.

In this object space representation, the direction of the relationship points to the item considered, not to the matter.

9.11 Form and appearance

This is a relationship looking at concept A as it is recognisable to the human visual sense: form, shape, or colour. This also applies to phenomena of the microcosm as far as they can be visualised in scanners, microscopes, etc. Examples of this relationship are: "double helix"/"chromosome thread", and "Roman arc"/"portal".

Concept A is the considered form or appearance, B indicates what has this form or appearance. So the direction of this relationship goes to the considered form or appearance.

9.12 Process

This is a case where concept B indicates a process involved in the concept of A, such as in "production"/ "progressive assembly line", or "embellishment of house walls"/"canned liquid colour spraying", and where it would be misplaced to apply Descendancy.

This relationship points in the direction of what is considered (A) under an aspect of the process which then appears as B.

9.13 State

A concept A may be looked at under the aspect of its state (B). So, a "photographic film" may have status as a "positive", or a "negative"; similar examples are "water" as "ice", "carbonic acid"/"solution in liquid", etc.

The relation of state points to the object considered, which is A.

The need for more, and other conceptual relationships which also should be included, for standardisation may occur in the course of the time. They may be added to this format scheme as possible options.

10. A Semantic Reference Structure - Why?

Better definition of the relationships, however, is not sufficient. This new thesaurus format started out with the idea that the following ambitious goals are highly desirable, and therefore must be achieved:

a) high articulating power on the descriptive level while maintaining terminological control;
b) consistency in i) synthesis of conceptual aggregates and ii) in the analysis of their constituting parts;
c) definition of a common reference platform to allow distributed construction and maintenance of linguistic thesauri as well as thesaurus data interchange.

This goal can only be achieved by means of an underlying, standardised *Basic Semantic Reference Structure (BSRS).* Therefore this structure forms an integrated part of this thesaurus.

The *BSRS* features a total of eight different potential semantic properties universally basic for the interpretation of monems, synthems and higher synthemised verbal expressions of a natural language. They certainly do not cover the whole wealth of semantic properties normally offered by most modern natural languages, but they seem to be sufficiently explicit and exhaustive as to enable the definitional goals pursued in this thesaurus which aims at practical use.

Each semantic property of use would be given a separate column in the *BSRS* and each descriptor or name shall be earmarked for use in one or more columns. So, each individual entry in the thesaurus will be attributed the number(s) of the column into which it fits. Entries requiring definition by semantic elements listed in more than one column would list these properties as elements of a tuple, as indicated below.

11. The Basic Semantic Reference Structure (BSRS)

The following formula for the *BSRS* is proposed:

An expression can be semantically determined by up to eight different conceptual elements. For examples see Table 3 below. They are represented:

a) by a preferential term (descriptor) of the thesaurus

and, in two special fields,

b) by the indication of a date in an abbreviated, standardised format.

Formally speaking, these entries are treated as distinct elements of a chain, or tuple. To interpret them, they are dealt with all on an equal footing, some of them forming attributes to distinguish others, some fully applicable alone or in combination with others. The underlying meaning of an ECD is co-determined by other entries according to the syntax of the tuple. The *BSRS* gives the general syntactic rule for reading such a chain, and the entry in a particular place of the *BSRS* conditions the meaning of the ECD entry.

Element 1

This is a descriptor for a general concept, seen as a class, or for classes of an existing taxonomy, such as are common for animals and plants. It is not used to determine a location as elements 4 or 5 are, nor as an aspect as element no. 6 is. Its definition excludes a descriptor in this element field that is used to describe such objects which are instances, i.e. objects unique and singular in space and time. Generally speaking, element no. 1 answers the question: *"What is it?"*

Element 2

This element is the denomination of such objects singular in space and time which are living beings and as such appear as actual or potential actors. These denominations are names. They may be the name of:

a) an individual (e.g. "Thomas Mann", "Knautschke") ;
b) an entity stipulated by law (e.g. company, foundation, government, church); or
c) group of persons acting under a distinct banner (e.g. "Franciscans", "Gruppe 47")

Element 2 answers the question: "Who is it?" "What legal or otherwise social body is it?"

Element 3

This element is the entry of a name for other phenomena, products, brands, events, happenings that are singular in space and time. Such entries may also apply to artistic styles (e.g. "art déco"), schools of thought (e.g. "Manchester capitalism"), reigns (e.g. "Louis Quinze") and regimes (e.g. "Stalinism"). Element no. 3 answers the question: *"What occurrence, material or immaterial process is it?"*

Elements 4 and 5

Element 4 is the item to fix the location or space in which element 1 occurs. This is not necessarily the same as would be needed for elements 2 and 3. More than one entry might be needed to locate element 1, as in the case of monetary union (1), France (4), Italy (4), Germany (4) etc., which then would form what, in the meantime has emerged as "Euroland", an entry which would have to be posted as element 3.

When used to express an aspect of spacial extension from element 4, this location should be entered as element 5. To give an example of this: the entry "narrow gauge lines" (1) can be seen under the aspect of "between" or "from-to", when the entry in element 4 is "Brazzaville", and 5 is "Pointe Noire". This, then, would read: narrow gauge line between Brazzaville and Pointe Noire.

This element/these elements answer the questions: *"Where is what it is?" "What are the extensions of what it is?"* or to put it in a simpler way: "Between which locations is it? From where to where?"

Element 6

In this element, a special viewpoint can be expressed under which element 1 (in conjunction with elements 4, 7 and 8) e.g. is seen as a "technical aspect", or "vision". The number and the definition of aspects admitted is limited. Aspects, with all their conceptual relationships, are, of course, listed and interrelated in the thesaurus, but especially marked for exclusive use in element 6, since they are not meant to be used in their potential as classes (which would be column 1).

Figure 2: List of admitted aspects proposed

Definition
Technical
Figures
Macro-figures
Process description
Design
Construction
Vision
Impact desired

Habit
Legal issue
Object of Art
Lifespan
Duration

(open to further entries)

This element is one of the most important elements, since it answers the question: *"Under which viewpoint is the descriptor in element 1 to be looked at?"* It must not be applied to living actors (element 2) or occurrences (element 3).

Element 7

This element contains whole year numbers. Use of months, days and smaller lapses of time may be optional. Open at both ends (>, <), and proximity (~).

Element 8

This element specifies any extension of time, including an open future (>2000).

12. Application of the BSRS

a) As a general semantic reference tool

To safely determine a meaning, it is not sufficient to say that a term or a noun phrase reads like this or like that. To achieve an unequivocal identification of a word's meaning it is necessary to say something about the semantic quality in which the word or expression is/shall be used, whether the expression means the object as a class, as an individual (name), or whether it is meant as a specifier for space and time, etc.

Table 3: Basic Semantic Reference Structure (BSRS) – Examples
Application case A: As a general semantic reference tool

What is it?	Who is it?	What an event is it?	Where is it?	Local extension	As seen from?	When is it?	Extension in time
Concept 1	Name 2	Event 3	Location 1 4	Location 2 5	Aspect 6	Time 1 7	Time 2 8
An expression for a concept seen as a class e.g. use of solar energy e.g. myopia e.g. poverty alleviation	A name or acronym of an instance seen as an actor. Class A: Living beings e.g. Yves Montand e.g. Micky Mouse Class B: Corporate beings e.g. Deutsche Shell AG e.g. The Beatles	The name for all other phenomena, products, brands, events, happenings, singular in space and time e.g. Stalinism e.g. art déco e.g. Uhu e.g. Viagra e.g. World War I	The name of a geographical location, area, or space e.g. Buenos Aires e.g. The Alps e.g. Rhône River e.g. Uranus	like 4	One of the elements as listed in canon above e.g. vision e.g. process description	Full year >, <, ~ e.g. > 1998 e.g. < 1815 ~ -200,000	like 7

With few exceptions, all entries in columns 1 – 6, on top of their semantic quality of use as reflected in these columns, do benefit from the semantic links with the other entries according to the relationships described above. This is to ensure the greatest possible number of entry paths to the user of a system based on this thesaurus format, e.g. from Latin America-->South America-->Argentina-->Buenos Aires, as defined in relation no. 4 (geographic-partitive).

b) As an explanatory instrument for instances

The *Basic Semantic Reference Structure* also serves as an instrument to normalise the nature and sequence of term elements needed to explain instances, i.e. phenomena unique and singular in space and in time, and to express them in such a way that full semantic and terminological control is maintained. If this is valid for all the languages of the thesaurus, this must be dealt with on the level of the middle language, or MLIN; if in a monolingual thesaurus – otherwise following this format – only one language is concerned, this must be done on the level of this single language.

In such a way, for example, it can be expressed as follows:

i) that "Baroque" (3) is by definition (6) an "event of style of life and art" (1) in "Western Countries" (4), or
ii) that "Sioux" (2) are by definition (6) a group of "indigenous peoples" (1) in "North America" (4), or
iii) that 'Deutscher Bundestag' (2) is the "first chamber of parliament" (1) in "Germany" (4), valid for the timeline 1948 → 2000, as opposed to former events in Germany with the name of Bundestag.

Table 4: Application case B: Explaining the instances

What is it?	Who is it?	What an event is it?	Where is it?	Local extension	As seen from?	When is it?	Extension in time
Concept 1	**Name** 2	**Event** 3	**Location 1** 4	**Location 2** 5	**Aspect** 6	**Time 1** 7	**Time 2** 8
Statesman	Thomas Moore		England		Definition		
Indigenous peoples	Sioux		North America		Definition		
Styles of life and art		Baroque	Western countries		Definition		
Civil wars		Spanish Civil War	Spain		Definition		
First chamber of Parliament	Deutscher Bundestag		Germany		Definition	1948	>2000

c) As a general, normalised indexing scheme

To serve specific IR applications, the *Basic Semantic Reference Structure (BSRS)* can be used, and is especially recommended, as a general, semantic/syntactic indexing scheme. In indexing, entry lines of a *BSRS* could be constructed in such a way as to best reflect the meaning of the content to be indexed. Such a line of entries could be dealt with as a tuple in a relational data bank. A document up for indexing would be represented in the database by a number of tuples, and the defined, semantic interrelations desired could be exploited in an easy way in searches. Tuples could have contradictory content. All this would enable search procedures to be much more efficient than those applied today (search procedures mainly based on term occurrence in chaotic texts, and Boolean algebra). In such a way, a more general thesaurus entry like the above-mentioned "Baroque" (3), could be supplemented by a separate entry tuple as follows:

Baroque (3), Germany (4), lifespan (6), ~1600 (7), ~1800 (8)

Such a tuple could serve as a conceptual bridge to the main bibliographic classification systems and other order systems in the World. Equivalent entries from the authority files to UDC, DDC, etc. would then be added as columns 9 and following (Table 3).

Table 5: Application case C: As a general indexing scheme

In excess of its function to normalise expressions for concepts contained in the thesaurus, the common *BSRS* can be used as a means of further syntactic indexing. Examples for this, are the following:

What is it?	Who is it?	What an event is it?	Where is it?	Local extension	As seen from?	When is it?	Extension in time
Concept 1	**Name** 2	**Event** 3	**Location 1** 4	**Location 2** 5	**Aspect** 6	**Time 1** 7	**Time 2** 8
Tree planting			Peru		Technical aspect	1991	>2000
Canal profiles			France		Design	1600	1700
Portable heaters			World		Design	1960	>2000
Narrow gauge lines		Congo-Océan	Brazzaville	Pointe Noire	Construction	1921	1934
Civil wars		Spanish Civil War	Spain		Process description	1936	1939

https://doi.org/10.5771/0943-7444-1999-1-10 https://www.inlibra.com/de/agb - Open Access -

d) As a construction principle for Encyclopaedias

Lexicographers would get a chance to supplement entries in the multilingual thesaurus. Thesauri could be developed into universally applicable, machine-readable, multilingual Encyclopaedias. Uncertainties of structure and functionalities, and economic risks of the implementation of such Encyclopaedias, would be removed in this way. They remain barriers to the realisation of such works to the present day,

Table 6: Application case D: To supplement standard lexical entries, like

	Thomas Moore		England		Lifespan	~ 1478	1535
		Baroque	Germany		Duration	~ 1600	~ 1800

This type of application opens a way to a defined, normalised construction of universal encyclopaedias which can then be used as knowledge bases for inference processes as needed in linguistic engineering and work on Artificial Intelligence matters.

12 Conclusion

A thesaurus along the lines of this formula could bring information retrieval (IR), and some other beginnings in linguistic engineering, a good step further into the future. One has to see the thesaurus applied to collections of text in which discourse follows the declarative mode. This is the case in most structured or unstructured text collections of our time, be they small or large. The thesaurus then can serve as a bridge of at least some robustness between language and thought, whereby the use of terms is tamed by syntactic rules given in the *Basic Semantic Reference Structure (BSRS)*. Both are constituent parts of the new formula.

- Firstly, such a thesaurus could be used as a machine-readable encyclopaedia explaining the phenomena of the world. Multilingual as it is, it would also serve as an inter-language lexicon.
- Secondly, of course, it could be used as an advanced, largely machine-aided, indexing tool.
- Thirdly, however, it would probably open up a whole new world in information gathering when directly applied as a tool in IR operations. This is because each single relationship between two universals, when stipulated in a valid way, means that a corresponding question can automatically be answered by a system fed from such a knowledge base. Such knowledge can then be offered to interested users in many ways, from the traditional question-and-answer type to the most advanced forms featuring animation, play, surprise, entertainment and adventure. Its wealth of knowledge would also be most welcome in modern education.

Such a machine-readable, domain independent, multilingual, encyclopaedic thesaurus will only be feasible on the condition that all interested parties are offered equal opportunity to contribute to the whole. This is why a clear formula is of utmost importance. Enforcement of the rules would be in the interest of those who apply the thesaurus, and are left free to use the powers of its interaction, and whose Meta Language Term Numbering system, after the umpteenth exchange of data, would finally prevail..

This is why the outline above also suggests the way in which such a joint effort could be brought about and carried out to an end, a wish which is not out of the question.

References:

1 Schmitz-Esser, W. (1999). Modélisation, au moyen d'un thésaurus encyclopédique et plurilingue, des connaissances présentées au cours de l'Exposition Mondiale de l'an 2000. In: J. Maniez, W. Mustafa el Hadi (Eds.) *Organisation des connaissances en vue de leur intégration dans les systèmes de représentation et de recherche d'information.* Villeneuve d'Ascq (Nord): Université Charles de Gaulle-Lille 3. 57-69.

2 Schmitz-Esser, W. (1998). Defining the conceptual space for a World Exhibition – first experiences. In: W. Mustafa el Hadi, J. Maniez and S.A. Pollitt (Eds.). *Structures and Relations in Knowledge Organization: Proceedings of the 5th International ISKO Conference, 25-29 August 1998, Lille, France.* Würzburg: ERGON Verlag. 146-152.

Prof. Dr. Winfried Schmitz-Esser,
Rothenbaumchaussee 3, D-20148 Hamburg,
Tel::++49/40/450 38 604, Fax: ++49/40/450 38 606, e-mail: Schmitz_Esser@csi.com