Multilingual thesauri and ontologies in cross-language retrieval

Dagobert Soergel
College of Library and Information Services
University of Maryland
College Park, MD 20742
ds52@umail.umd.edu

Presented at the AAAI Spring Symposium on Cross-Language Text and Speech Retrieval
Stanford University, March 24-26, 1997

Abstract

This paper sets forth a framework for the use of thesauri and ontologies as knowledge bases in cross-language retrieval. It provides a general introduction to thesaurus functions, structure, and construction with particular attention to the problems of multilingual thesauri. It proposes the creation of environments for distributed collaborative knowledge base development as a way to make large-scale knowledge-based systems more affordable.

Outline

1 Introduction
2 Thesaurus functions
  2.1 User-centered /request-oriented indexing
3 Thesaurus structure
  3.1 Brief review of thesaurus structure principles
  3.2 Special issues in multilingual thesauri
4 Implementing thesaurus functions in retrieval
  4.1 Controlled vocabulary
  4.2 Free-text searching
  4.3 Knowledge-based support of searching
5 Thesaurus construction
6 Affordable implementation of knowledge-based approaches

1 Introduction

A thesaurus is a structure that manages the complexities of terminology in language and provides conceptual relationships, ideally through an embedded classification/ontology. This paper will

- give a tutorial on thesaurus functions and structure;
- present a concept retrieval perspective - concepts bridge languages in retrieval;
- argue the perspective that cross-language retrieval is also cross-cultural retrieval.

The paper covers thesaurus functions beyond retrieval; in retrieval, it considers any kind of object, not just text.

We start out with some definitions (Figures 1a and 1b) and then give examples illustrating thesaurus problems and thesaurus structure (Figures 2 and 3).

Cross-language retrieval is the retrieval of any type of object (texts, images, products, etc.) composed or indexed in one language (the target language) with a query formulated in another language (the source language). There may be any number of source languages and any number of target languages. Queries can be written or spoken or constructed by selections from a menu presented in the source language.

Text retrieval (broadly defined) is the retrieval of text, written or spoken. (While *text retrieval* has come to mean retrieval of written text, and *speech retrieval* retrieval of spoken text, the broad meaning of *text* used here follows usage in linguistics).

Free-text retrieval is text retrieval based on the text itself without index terms or other cues.

Fig. 1a. Definitions: Retrieval

A **dictionary** is a listing of words and phrases giving information such as spelling, morphology and part of speech, senses, definitions, usage, origin, and equivalents in other languages (bi- or multilingual dictionary).

A **thesaurus** is a structure that manages the complexities of terminology in language and provides conceptual relationships, ideally through a classification/ontology.

A thesaurus may specify descriptors authorized for indexing and searching. These descriptors then form a **controlled vocabulary** (**authority list**, **index language**).

A **monolingual thesaurus** has terms from one language, a **multilingual thesaurus** from two or more languages.

A **classification** is a structure that organizes concepts into a hierarchy, possibly in a scheme of facets. The term **ontology** is often used for a shallow classification of basic categories or a classification used in linguistics, data element definition, or knowledge management.

Fig. 1b. Definitions: Thesaurus, etc.

English	German
simian Affe	
monkey	*niederer Affe*
ape	Menschenaffe
timepiece	Uhr
clock	*Wanduhr, Standuhr, Turmuhr*
wall clock	Wanduhr
standing clock	Standuhr
tower clock	Turmuhr
watch	*Taschenuhr, Armbanduhr*
pocket watch	Taschenuhr
wrist watch	Armbanduhr
alarm clock	Wecker
blanket, rug, carpet	Teppich
blanket	Betteppich
rug, carpet	Bodenteppich
rug (or carpet)	*loser Bodenteppich*
long, narrow rug	Läufer
wall-to-wall carpet	Teppichfußboden
hanging rug	Wandteppich

Italics denotes terms created to express a concept not lexicalized in English or German, respectively.

Note that most English-German dictionaries would have you believe that the German equivalent for "monkey" is "Affe", but that equivalence holds only in some contexts.

Another difficulty arises when two terms mean almost the same thing but differ slightly in meaning or connotation, such as *alcoholism* in English and *alcoholisme* in French, or *vegetable* in English (which includes potatoes) and *Gemüse* in German, which does not. If the difference is big enough, one needs to introduce two separate concepts under a broader term; otherwise a scope note needs to clearly instruct indexers in all languages how the term is to be used so that the indexing stays, as far as possible, free from cultural bias or reflects multiple biases by assigning several descriptors.

Fig. 2. Multilingual thesaurus problems

The example in Figure 2 illustrates the conceptual and terminological problems in aligning the vocabularies of two languages. Concepts lexicalized in one language may not be lexicalized in the other and vice versa, creating significant problems for translation. The complexities of term correspondence are best managed with a conceptual approach, establishing a concept interlingua, so to speak.

Figures 3a and 3b give a first illustration of thesaurus structure, to be discussed more fully in Section 3. The emphasis is on illustrating the concept-based approach to vocabulary management and retrieval, so the examples are drawn from a thesaurus that epitomizes that approach, even though it is monolingual. The reader may want to look at the sample pages from several multilingual thesauri given in the appendix.

Figure 3a presents an excerpt from a thesaurus hierarchy. Through identifying the relevant facets (facet headings EF2, EF4, and EF6) and arranging the concepts within each facet in meaningful order displaying the concept relationships, the hierarchy elucidates the conceptual structure of the domain: The *route of administration* of drugs can be described by giving the intended *scope of drug action*, the *method of administration*, and the *body site* where the drug is administered. The last two facets have been combined because they are strongly intertwined. The hierarchy gives a logical arrangement for concepts within a facet, allowing the reader to form a clear mental image of methods available for administering drugs. The hierarchy also allows for hierarchic query expansion (whether searching with a controlled vocabulary or free-text).

Figure 3b gives examples of full thesaurus entries. The entry for EF gives many synonyms that can be used for synonym expansion of query terms. The RT cross-references suggest further descriptors that might be useful for searching. The entries for EF2 give scope notes that carefully define each concept. Thus, the thesaurus can serve as a reference. Juxtaposing the scope notes for hierarchical neighbor concepts allows the indexer or searcher to pick the right concept at the right hierarchical level.

Having established a general understanding of thesaurus structure, we can now deal with the functions, structure, and construction of thesauri in more detail.

EF	route of administration
EF2	**— by scope of drug action**
EF2.2	. topical and local administration
EF2.2.2	. . topical administration
EF2.2.4	. . local drug administration
EF2.4	. systemic administration
EF4	**— by method or body site**
EF4.2	. enteral administration
EF4.2.2	. . oral enteral administration
EF4.2.4	. . rectal enteral administration
EF4.4	. mucosal administration
EF4.4.2	. . transdermal administration
EF4.4.4	. . inhalation, smoking, sniffing
EF4.4.4.2	. . . smoking
EF4.4.4.2.2	. . . . smoking w/out inhalation
EF4.4.4.2.4	. . . . smoking with inhalation
EF4.4.4.4	. . . nasal administration
EF4.4.4.6	. . . pulmonary administration
EF4.4.6	. . oral mucosal administration
EF4.4.6.2	. . . buccal administration
EF4.4.6.4	. . . sublingual administration
EF4.4.8	. . rectal mucosal administration
EF4.6	. parenteral administration
EF4.6.2	. . intravenous injection
EF4.6.2.2	. . . intravenous infusion
EF4.6.4	. . intra-arterial injection
EF4.6.6	. . intraperitoneal administration
EF4.6.8	. . intracutaneous injection
EF4.6.10	. . admin. through skin implant
EF4.6.12	. . subcutaneous injection
EF4.6.14	. . intramuscular injection
EF4.6.16	. . CNS injection
EF4.6.16.2	. . . intrathecal injection
EF4.8	. skin administration (The full entry shows Narrower Term cross-references to the more specific methods involving the skin: EF4.4.2, EF4.6.8, EF4.6.10, and EF4.6.12)
EF4.10	. oral administration (NT to EF4.2.2, EF4.4.4.2, and EF4.4.6)
EF4.10	. rectal administration (NT to EF4.2.4 and EF4.4.8)
EF6	**drug administration by self vs. others**
EF6.2	. self administration of drugs
EF6.4	. drug administration by others

Fig. 3a. Excerpt from a thesaurus hierarchy

EF	route of administration
	ST *medication route* ST *method of delivery of drugs or food* ST *mode of substance administration* ST *route of drug application* ST *route of drug entry* ST *route of exposure* BT +EE12 pharmacokinetics RT +AA2 AOD use RT +BS AOD substance by route of admin. RT EE12.2e drug absorption RT +EE14.4.8 drug effect by location RT +HR drug therapy RT MD2.2.2.2 drug paraphernalia
EF2	**route of admin. by scope of drug action** SN Use one of these descriptors in combination with a descriptor from **+EF4 route of admin. by method or body site**.
EF2.2	. topical and local administration SN The application of a substance to a localized area, chiefly for local effects at this site. NT HU4.2 local anesthesia RT GH10.2 chemical injury
EF2.2.2	. . topical administration SN The application of a substance on the surface of the skin or on a mucous membrane (incl. the gastrointestinal membrane) so that the substance will take effect on the surface or on a localized layer under the surface. For example, for the administration of a decongestant spray, use **EF2.2.2 topical administration** combined with **EF4.4.4.4 nasal administration**. ST *topical application*
EF2.2.4	. . local drug administration SN The introduction of a substance into a localized area of the skin or other tissue, as through injection. NT EF4.6.4 intra-arterial injection NT EF4.6.8 intracutaneous injection NT +EF4.6.16 CNS injection
EF2.4	. systemic administration SN The introduction of a substance into systemic circulation so that it is carried to the site of effect. NT +EF4.6.2e intravenous injection NT EF4.6.10 admin. through skin implant NT HU4.4 general anesthesia RT +GH10.4 chemical poisoning

Fig. 3b. Examples of full thesaurus entries

2 Thesaurus functions

A thesaurus with its embedded classification/ontology serves many functions, all of which are significantly affected by multilinguality. Our emphasis will be on thesaurus functions in retrieval (Figure 4a), but from a broader perspective one must not lose sight of the many other functions a thesaurus and its embedded classification/ontology can serve (Figure 4b).

In information retrieval a thesaurus or a classification/ ontology without the surrounding terminological structure can be used in two scenarios:

(1) knowledge-based support of free-text searching (applicable only to written or spoken text, although the text could point to another object, e.g., retrieving images through a free-text search of image captions or through a search of the text portion of a movie);

(2) controlled vocabulary indexing and searching (applicable to any kind of retrieval).

The first two functions apply to either scenario. A user can always profit from looking at a conceptual framework of the domain to clarify the search topic, and the thesaurus then further assists in finding good search terms for the concepts identified, Synonym expansion includes mapping to terms from a different language.

Using a thesaurus as an indexing tool applies only to controlled vocabulary indexing. Indexing, the assignment of a set of descriptors to a document or other object, can be manual or automated. Of particular importance, so often overlooked, is an approach to indexing that places the users, their problems and questions squarely at the center of attention: user-centered or request-oriented (or problem-oriented) indexing. The few empirical studies evaluating user-centered (as opposed to the commonly used document-centered) indexing show a positive effect on retrieval performance (Pejtersen 1983). This approach is central in information retrieval; it will be discussed in Section 2.1.

To look at thesaurus functions more generally, we first observe that a thesaurus is a knowledge base of concepts and terminology; other such knowledge bases are dictionaries and ontologies developed for AI applications, linguistic systems, or data element definition. Since these different types of knowledge bases — though developed for different purposes — overlap greatly, it would be best to integrate them through a common access system (Soergel 1996). The functions to be served by such a virtual integrated knowledge base of concepts and terminology are listed in Figure 4b.

Knowledge-based support of searching
(explicit assistance to the user or behind the scenes)

Menu trees

Guided conceptual analysis of a search topic

Browsing a hierarchy to identify search concepts

Mapping from query terms to descriptors used in one or more databases or synonym expansion of query terms for free-text searching

Hierarchical expansion of query terms

Meaningful arrangement of search results

Tool for indexing

vocabulary control
user-centered indexing

Fig. 4a. Thesaurus functions in information retrieval

2.1 User-centered /request-oriented indexing

As summarized in Figures 5a and 5b, user-centered indexing involves analyzing actual and anticipated user queries and interests and constructing a framework, a hierarchically structured controlled vocabulary, that includes the concepts of interest to the users and thus communicates these interests to the indexers or an expert system that can infer user-relevant concepts from text. The indexers then become the "eyes and ears" of the users and index materials from the users' perspective. The indexer uses the structured list of user-relevant concepts as a checklist, applying her understanding of a document (or other object) to judge its relevance to any of these concepts. This process ensures that users will find the documents that they themselves would judge relevant upon examination.

Request-oriented indexing contrasts with document-oriented indexing, where the indexer simply expresses what the document is about or where simply the terms in the text are used. But a document can be **relevant** for a concept without being **about** the concept: a document titled *The percentage of children of blue-collar workers going to college* is not necessarily about *intergenerational social mobility*, but a researcher interested in that topic would surely like to find it, so it is relevant. Another example: Since users are interested in the *biochemical basis of behavior* and also in *longitudinal studies*, these descriptors are in the thesaurus. The indexer examines the document *CSF studies on alcoholism and related behaviors* and finds that it is relevant to both descriptors. *Longitudinal* is not mentioned in the document, but careful examination of the methods section reveals the concept.

- **Provide a semantic road map to individual fields and the relationships among fields; relate concepts to terms, and provide definitions**, thus providing orientation and serving as a reference tool.
- **Improve communication and learning generally**:
  - Assist writers: suggest from a semantic field the term that best conveys the intended meaning and connotation.
  - Assist readers in ascertaining the proper meaning of a term and placing it in context.
  - Support learning through conceptual frameworks.
  - Support language learning and the development of instructional materials.
- Provide the **conceptual basis for the design of good research and practice**.
  - Assist researchers and practitioners in exploring the conceptual context of a research project, policy, plan, or implementation project and in **structuring the problem**.
  - Assist in the consistent definition of variables and measures for more comparable and cumulative research and evaluation results. Especially important for cross-national comparisons.
- **Provide classification for action**:
  - a classification of diseases for diagnosis,
  - of medical procedures for insurance billing,
  - of commodities for customs.
- **Knowledge base to support information retrieval** (Fig. 4a)
- **Ontology for data element definition**. Data element dictionary. Consider data processing systems in a multinational corporation
- **Conceptual basis for knowledge-based systems**.
- **Do all this across multiple languages**
- **Mono-, bi-, or multilingual dictionary for human use**. **Dictionary/knowledge base for automated language processing** - machine translation and natural language understanding (data extraction, automatic abstracting/indexing).

Fig. 4b. **Broader functions of a knowledge base of concepts and terminology**

This kind of indexing is expensive, unless it can (to a degree) be automated through a knowledge-based system for automated indexing. Is it worthwhile? The worth derived from improved performance depends on the use of the retrieval results.

Construct a classification/ontology (embedded in a thesaurus) based on actual and anticipated user queries and interests.

Thus provide a conceptual framework that organizes user interests and communicates them to indexers.

Index materials from users' perspective:
Add need-based retrieval clues beyond those available in the document. Increase probability that a retrieval clue corresponding to a query topic is available.

Index language as checklist.
Indexing = judging relevance against user concepts.
Relevance rather than aboutness

Implementation: Knowledgeable indexers or an expert system using syntactic & semantic analysis & inference.

Fig. 5a. User-centered / request-oriented indexing

Document

The drug was injected into the aorta

User concept: Systemic administration

Document:

The percentage of children of blue-collar workers going to college

User concept: intergenerational social mobility

Document:

CSF studies on alcoholism and related behaviors

User concept: longitudinal study
(Longitudinal not mentioned in the document; determined through careful examination of the methods section.)

Fig. 5b. **Request-oriented indexing. Examples**

This perspective on indexing has implications for cross-language retrieval: The conceptual framework must be communicated in every participating language to allow a meeting of minds to take place, regardless of the languages of the user and the indexer. This is particularly salient in the context of indexing images with descriptors that capture imponderables, such as the mood of an image: One needs to make sure that, as far as possible, the term used by the indexer in one language communicates the same mood as the term given to the user in another language for searching.

3 Thesaurus structure

After a brief review of general principles (3.1), we discuss issues specific to multilingual thesauri (3.2).

3.1 Brief review of thesaurus structure principles

Thesaurus structure consists of the terminological structure that relates terms to concepts (by establishing synonym relationships and disambiguating homonyms) and conceptual structure. We discuss each in turn.

Terminological structure (Figure 6)

Controlling synonyms

Term	*Preferred synonym*
Alcoholism	Alcohol dependence
Inheritance	Heredity
Ultrasonic cardiography	Echocardiography
Black	African American
Afro-American	African American
Pregnant adolescent	Pregnant teen

Disambiguating homonyms

administration 1 (management)

administration 2 (drugs)

Läufer 1 (Sportler)	Engl.: runner (athlete)
Läufer 2 (Teppich)	Engl.: long, narrow rug
Läufer 3 (Schach)	Engl.: bishop (chess)

discharge 1 (From hospital or program)
German: Entlassung

discharge 2 (From organization or employment)
Preferred synonym: Dismissal
German: Entlassung

discharge 3 (Medical symptom)
German: Absonderung, Ausfluss

discharge 4 (into a river) German: Ausfluss

discharge 5 (Electrical)
German: Entladung (which also means unloading)

Fig. 6. Terminological structure examples

The terminological structure is equally important in controlled vocabulary and in free-text searching. In free-text searching, synonym expansion of query terms is important, and homonym indicators can trigger a question to the user on the intended meaning of the query term.

Conceptual structure

A well-developed conceptual structure is a *sine qua non* for user-centered indexing and is very useful for free-text retrieval as well. The two principles of conceptual structure are facet analysis and hierarchy.

Facets. Semantic factoring or feature analysis

Semantic factoring means analyzing a concept into its defining components (elemental concepts or features). This gives rise to a concept frame with facet slots. See Figure 7 for examples.

liver cirrhosis	
Pathologic process:	inflammation
Body system:	liver
Cause:	not specified
Substance/organism:	not specified
alcoholic liver cirrhosis	
Pathologic process:	inflammation
Body system:	liver
Cause:	chemically induced
Substance/organism:	alcohol
hepatitis A	
Pathologic process:	inflammation
Body system:	liver
Cause:	infection
Substance/organism:	hepatitis A virus

Fig. 7. Facet analysis examples

A facet groups concepts that fall under the same aspect or feature in the definition of more complex concepts; it groups all concepts that can be answers to a given question. In frame terminology: The facets listed above are slots in a disease frame; a facet groups all concepts that can serve as fillers in one slot.

Using elemental concepts as building blocks for constructing compound concepts drastically reduces the number of concepts in the thesaurus and thus leads to conceptual economy. It also facilitates the search for general concepts, such as searching for the concept *dependence*, which occurs in the context of medicine, psychology, and social relations.

Facets can be defined at high or low levels in the hierarchy, as illustrated in Figure 8.

Top-level facets

organism
body part
chemical substances by function
chemical substances by structure.

Low-level facets

route of administration
- route of administration by scope of drug action (local/topical or systemic)
- route of administration by body part
- route of administration by method of application (injection, rubbing on, etc.)

liver
- liver tissue (hepatocyte, Kupffer cell, etc.)
- liver part (hepatic lobule, portal lobule, etc.)

Fig. 8. Facets at different hierarchical levels

Hierarchy

A sample hierarchy was presented in Fig. 3a. For another example, consider a search for a broad concept and the more specific concepts that should be included in the query as illustrated in Figure 9.

groups at high risk of drug use
- suicidal or physically or mentally disabled
- persons from unstable or low-cohesion families
- children of alcoholic or other drug-abusing parents
  - SN Adult or still under age
- children of single teenage mothers
- persons subjected to abuse or neglect (now or past)
  - persons subjected to abuse/neglect by parents
    - latchkey children
  - persons subjected to abuse/neglect by spouse
- single teenage mothers
- school dropouts or those at risk of dropping out
- unemployed or in danger of being unemployed
- economically disadvantaged
- homeless
  - runaway youth
- gateway drug users
- persons engaged in violent or delinquent acts

Fig. 9. Hierarchy for searching

Uses of facet analysis and hierarchy

Through facet analysis and hierarchy building, the lexicographer often discovers concepts that are needed in searching or that enhance the logic of the concept hierarchy; he then needs to create terms for these concepts. Examples are *traffic station* as the semantic component common to *train station*, *bus station*, *harbor*, and *airport* or *distinct distilled spirits* (as the semantic component common to *gin*, *whiskey*, *cherry brandy*, *tequila*, etc., the counterpart of the already lexicalized *neutral distilled spirits*), or *analytic psychotherapy* as an umbrella term for a host of methods (such as *insight therapy*, *Gestalt therapy*, and *reality therapy*) that all seek to assist patients in a personality reconstruction through insight into their inner selves.

Fig. 10 lists the most important uses of facet analysis and hierarchy. These uses will be more fully discussed in Section 4.

Help to organize the concept space and establish concept relationships.

Assist the user in analyzing and clarifying a search problem both in terms of the facets involved and the hierarchical structure within each facet.

Facilitate the search for general concepts, such as

inflammation or

dependence (which occurs in the context of medicine, psychology, and social relations)

Hierarchic query term expansion

These functions are useful in both controlled vocabulary and free-text searching.

Fig. 10. Uses of facet analysis and hierarchy

3.2 Special issues in multilingual thesauri

A multilingual thesaurus for indexing and searching with a controlled vocabulary can be seen as a set of monolingual thesauri that all map to a common system of concepts. With a controlled vocabulary, indexing is concept-based; cross-language retrieval is simply a matter of providing designations for these concepts in multiple languages so that queries can be written in multiple languages. However, as the example in Fig. 2 illustrates, conceptual systems represented in the vocabulary of different languages do not completely coincide.

The crux of the matter, then, is which concepts to include. Ideally, the thesaurus should include all concepts needed in searching by any user in any of the source languages. Language differences often also imply cultural and conceptual differences, more so in some fields than in others. We need to create a classification that includes all concepts suggested by any of the languages. At a minimum this includes all relevant concepts lexicalized in at least one of the source languages. Also, different languages often suggest different ways of classifying a domain; the system needs to be hospitable to all of these. The problem that has bedeviled many developers of multilingual thesauri is that a concept lexicalized in one language may not be lexicalized in another and that the terms that do exist often vary slightly in meaning, possibly giving rise to different relationships. Starting from the misguided notion that a thesaurus should include only concepts for which there is a term in the language and that term meanings cannot be adjusted for purposes of the thesaurus, they had difficulty making the system of concepts the same for all languages. But, as we have seen, even in a monolingual thesaurus the lexicographer often discovers concepts needed in searching or to enhance the logic of the concept hierarchy and then needs to create terms for these concepts. In multilingual thesauri this necessity arises more often, particularly when different languages differ in the hierarchical levels at which they lexicalize concepts. The principle proposed here is to establish a common conceptual system, which may require an arduous. and expensive, process of negotiation, and then arrange for the terms in all languages to fit, giving proper definitions, of course.

It is clear that the problems discussed here and illustrated in Fig. 2 and in Section 3.1 have major implications for cross-language free-text searching: Each query term should be mapped from the source language to its multiple equivalents in the target language; each of these equivalents may have other meanings in the target language, presenting potential problems for retrieval. The query term may not have a precise equivalent in the target language; one may need to map to broader or narrower terms, distorting the meaning of the original query.

4 Implementing thesaurus functions in retrieval systems

4.1 Controlled vocabulary

With a controlled vocabulary there is a defined set of concepts used in indexing and searching. Cross-language retrieval simply means that the user should be able to use a term in his own language to find the corresponding concept identifier that is used to retrieve documents (or whatever the retrieval objects are). In the simplest system, this can be achieved through manual look-up in a thesaurus that includes for each concept corresponding terms from several languages and has an index for each language. In more sophisticated systems the mapping from term to descriptor would be done internally. As an example, consider a library catalog using the Library of Congress Subject Headings, for which French and Spanish translations are available. In the VTLS automated library system, each subject heading is identified by a number that is used in the document record. The authority file includes for each subject heading the preferred term and any synonyms; this information can be included in multiple languages. One could use this arrangement for assisting the user in finding subject headings or automatic mapping of user terms to subject headings as follows: Do a free-text search on authority records to find any subject heading for which either the preferred term or any synonym contains the user's query word or phrase in any language. Once appropriate subject headings are found, they can be used to retrieve documents.

Whenever the mapping from user terms to descriptors is done "behind the scenes", transparent to the user, the system should ask the user for clarification whenever the query word or phrase has multiple meanings and cannot be disambiguated automatically. Beyond that, showing the user the descriptor(s) the system came up with in their hierarchical context might improve the accuracy of the query formulation and thus retrieval. The success of this type of interaction depends on the quality of the hierarchy and the interface.

If the user has voice input available, one might even include the spoken form of terms in the thesaurus to enable voice input of query terms which would then be mapped to the appropriate descriptors.

A cross-language retrieval system with controlled vocabulary must also support indexing of documents or other objects, that is the assignment of controlled vocabulary descriptors, in the various languages. For manual indexing, this is accomplished by having thesaurus versions in each of the languages so that each indexer has a version in her own language. But that is not enough. The thesaurus version in each language must make sure that the indexer in that language fully understands the meaning of a descriptor that originated from another language;

otherwise, the indexing of such a descriptor will not be consistent across the database.

Automated indexing with a controlled vocabulary, particularly if it is to take a request-oriented slant, can be accomplished with a knowledge base that (1) allows recognition of important words and phrases (for spoken text this requires the inclusion of spoken forms) and allows for homonym disambiguation and (2) gives mapping rules that lead from the (possibly weighted) set of words and phrases identified for a document to a set of descriptors that should be assigned.

Such mapping rules can take many forms. In their simplest form, they specify a direct mapping from text words or phrases to the appropriate descriptors for each word or phrase (and possibly even word or phrase combinations). To increase accuracy, the mapping can be made dependent on context (Hlava 97). A more complex mapping relies on association strengths between terms (words and phrases) and descriptors. Broadly speaking, the association strength between term T and descriptor D could be seen as the predictive probability that the document containing term T should be indexed with descriptor D. Such association strengths can be computed from a training set of indexed documents. This is the approach often taken in automated text categorization, where often, but not always, the goal is to index each document by only one descriptor (assign it to one of a set of non-overlapping categories). An advanced version of this approach is the use of "topic signatures", profiles consisting of a set of terms with weights; a document is assigned the topic if its terms match the topic signature (Lin 1997). In effect, a topic signature is a query which identifies documents relevant to the topic.

As the foregoing discussion illustrates, the knowledge base needed to support automated indexing is more complex than a thesaurus for manual indexing. It must include more terms and term variants so that the words and phrases important for indexing can be recognized in the text, and it must include information needed for the disambiguation of homonyms (which often requires determining the part of speech of a text word).

For indexing and searching, a controlled-vocabulary cross-language retrieval system can be seen as a set of monolingual systems, each of which maps the terms from its language to a common system of concepts used in indexing and searching. For manual indexing and query formulation, this is accomplished through a multilingual thesaurus, which may in fact consist of multiple monolingual thesauri linked through common descriptor identifiers (such as Dewey Decimal class numbers). Automated indexing in cross-language text retrieval with texts in multiple languages means mapping from each language to the common conceptual structure represented in the controlled vocabulary. The knowledge base component dealing with identification of words and phrases for automated indexing can be developed independently for each language. Mapping rules that are entirely term-based can also be developed independently for each language. However, some mapping rules, for example rules based on context or topic profiles, may include conceptual elements that could be shared across languages.

There are a number of controlled-vocabulary cross-language retrieval systems based on manual indexing in use in bilingual or multilingual areas such as Switzerland, Belgium, Canada, and areas of the US with large Spanish-speaking populations; in international organizations, such as the European Community; and in international collaborative systems, such as AGRIS. These systems are based on the Universal Decimal Classification, which has been translated into many languages (library of the ETH, Zurich); on the Library of Congress Subject Headings, which have been translated into French; on EUROVOC, an EC thesaurus in 9 languages; and AGROVOC, a thesaurus in three languages created by translation from its original English-only version. There are a large number of thesauri that either have been developed as multilingual thesauri or have been translated into several languages.

4.2 Free-text searching

Cross-language free-text searching, finding texts in one language that are relevant for a query formulated in another language without relying on controlled vocabulary indexing, is a more complex proposition. It requires that each term in the query be mapped to a set of search terms in the language of the texts, possibly attaching weights expressing the degree to which occurrence of a search term in a text would contribute to the relevance of the text to the query term. To assist with this task, a thesaurus must include the mapping information. If the thesaurus includes fine-grained definitions that deal with subtle differences of meaning, distance between such definitions can be used to derive term weights.

A major difficulty of this mapping is that a homonym used in the query gives rise to multiple translations, each corresponding to one of its meanings. The target terms may in turn be homonyms in their language and thus retrieve many irrelevant documents unless text terms are disambiguated. (This problem exists in synonym expansion in one language as well but is exacerbated in cross-language text retrieval.) When the mapping goes to a term that has multiple meanings, the specific meaning should be identified, possibly in interaction with the user. For best retrieval results the terms in the texts should also be disambiguated so that only documents that include the term in the right sense score

The issue of homonymy in retrieval is not as straightforward as it may seem at first glance (Sanderson 1994). First of all, quite a bit of disambiguation may occur "naturally", in that a given term may assume only one of its meanings in the specific domain of the collection and there-

fore in the queries. Second, in a multi-component query, a document that includes a homonymous term from the first query component in a meaning other than that intended in the query is unlikely to also include a term from another query component, so excluding irrelevant documents may not require disambiguation in either the query or the texts. On the other hand, with single-concept query to a general collection (such as the World Wide Web), disambiguation can be expected to have a beneficial effect on retrieval performance. Failing that, a system might be able to suggest to the user an additional query component that would separate out the documents that include the query term but in a different meaning. Note that information extraction is much more dependent on homonym disambiguation.

In any event, for best support of free-text retrieval a thesaurus should flag homonyms, give their senses, and include rules for disambiguation.

The greater difficulty of free-text cross-language retrieval stems in no small measure from the fact that one must work with actual usage, while in controlled-vocabulary retrieval one can, to some extent, dictate usage.

4.3 Thesauri for knowledge-based search support

Whether searching is by controlled vocabulary or by free text, it is often helpful to the user to browse a well-structured and well-displayed hierarchy of concepts, preferably with the option of including definitions. A more sophisticated system may guide a user through a facet analysis of her topic. These aids provided by the system enable the user to form a better idea of her need and to locate the most suitable descriptors or free-text search terms. The guidance through facets and their hierarchical display must be available in the language of the user. These suggestions are based on the assumption that browsing a hierarchy is natural to most users and that users will appreciate the structure provided. This assumption rests on the belief that people try to make sense of the world and that guided facet analysis and browsing well-structured hierarchies help them do so. There is anecdotal evidence to support this assumption, but it needs to be investigated by building prototype systems and studying users' success (see, for example, Pollitt 1996).

This is one example of using a thesaurus as a knowledge base to make searching more successful. The assistance provided does not require that the user be an expert in classification and thesauri. This is even more true for "behind-the-scenes" assistance. There is no need to teach users about following a cross-reference from a synonym to a descriptor if the system searches for the descriptor automatically. There is no need to tell the user to look under narrower terms also if the system can do a hierarchically expanded search. There is no need to tell the user about strategies of broadening the search if the system, in response to a user input that not enough was found, can suggest further descriptors to be searched based on cross--references in the thesaurus. Sophisticated retrieval software can make the use of thesauri in retrieval independent of the user's knowledge and thereby can get much more mileage out of the investment in thesauri.

5 Thesaurus construction

Building a thesaurus, especially a multilingual thesaurus, takes a lot of effort. Some term relationships can be derived by statistical analysis of term occurrence in corpora, but this will not result in the kind of well-structured conceptual system described above. Developing such a structure requires intellectual effort.

A common method for thesaurus construction in a single language is to work bottom-up: One collects a list of terms (words and phrases), preferably from search requests, but also from documents, free-term indexing, and other thesauri. These terms are then sorted into increasingly fine-grained groups, until a group contains only synonyms or terms that, for purposes of the thesaurus, can be considered synonyms. In this process at least some homonyms will be detected; they must be disambiguated into several senses, each expressed by its own (possibly newly coined) term having one meaning and being grouped accordingly. A group of synonyms can be considered to represent a concept; usually a preferred term to designate the concept is selected, but some other concept identifier can be used. A first rough hierarchy of concepts emerges from this process.

This is followed by conceptual analysis, especially facet analysis at various levels, resulting in a well-structured faceted hierarchy. Next, one needs to write definitions (scope notes), in the process of which one may rethink the hierarchy, and introduce relationships between concepts that complement the hierarchy.

The development of a multilingual thesaurus is, naturally, an even more complex undertaking; the basic approaches are summarized in Figure 12. The ideal way to develop a multilingual thesaurus is to start from a pool of terms in all covered languages and carry out the process without regard to the language of the terms. This will bring together terms from different languages that have the same meaning into one group. This process gives all languages an equal chance to contribute concepts and concept relationships. It also forces a careful analysis of the meaning of each term in each language to determine the degree of equivalence, making it possible to develop the fine-grained structure of definitions that has the potential of providing powerful support to free-text cross-language retrieval.

Of course, this process would require a lexicographer knowledgeable in the subject matter of the thesaurus and fluent in all covered languages, not a very practical requirement. A more practical variation that still maintains the

spirit of this approach is to start with two languages and develop the conceptual structure — a bi-lingual lexicographer is needed in any event. Definitions should be written in both languages. One would then work on a pool of terms in a third language and fit it into the structure, creating new concepts as necessary. This is not at all the same as translating the thesaurus into the third language. This requires a lexicographer fluent in one of the starting languages and the third language. Following the same principle, one can now add other languages.

The result of such a process is a conceptual system that brings the conceptual structures embedded in the different languages under one roof, so to speak.

The most common approach to the construction of a multilingual thesaurus is to translate an existing monolingual thesaurus into one or more languages. But this approach is problematic: The original language and its vocabulary determine the conceptual structure, and one merely looks for equivalent terms in the second language without covering its terminological richness . In some multilingual thesauri, only one term in the target languages is provided, making the thesaurus unsuitable for query term expansion in free-text searching.

In between is an approach in which one starts with a monolingual thesaurus as the center and fits terms from one or more other languages into the structure of this central thesaurus without changing the concepts or the hierarchy. EuroWordNet (Gillaranz 1997) takes an improved variation of this approach, working with the English WordNet as its central thesaurus. In EuroWordNet, separate and independent word nets are constructed in each language in parallel efforts, each identifying synonym sets in that language (A synset can be considered a concept). Each individual language project then independently maps its sysnsets to the corresponding WordNet synsets; no changes are made to WordNet. In addition to identity, this mapping allows for hyponym and hypernym relationships, thus indicating that the concept identified in the language being worked on is not included in WordNet, but giving at least the hierarchical location. EuroWordNet also uses a very weak variation of approach 5: The participants developed a “top ontology”, which presumably reflects and integrates perspectives from their individual cultures. In addition to being mapped to WordNet, the individual language synsets are also mapped to this top ontology.

Requirements

Must cover all concepts of interest to the users in the various languages, at a minimum all domain concepts lexicalized in any of the participating languages.
Must accommodate hierarchical structures suggested by different languages.

Approaches (by increasing complexity and quality)

(1) Start from monolingual thesaurus and translate. This approach does not capture concepts lexicalized only in another language and is biased to the conceptual structure underlying the starting language. May not produce all synonyms in the second language.

(2) Start from a monolingual thesaurus as the center. Collect terms from a second (third, ...) language and establish correspondences of these terms to the central thesaurus. Suffers from similar bias toward the starting language as (1), but may cover more synonyms in the other languages.

(3) Work with a central thesaurus as in (2), but after collecting terms from a second language first group them into synsets, that is, derive concepts each of which is represented by a set of terms, and then map each concept to the corresponding concept in the central thesaurus or indicate that the concept is new and give the nearest broader or narrower concept in the central thesaurus. Note that the central thesaurus remains unchanged.

(4) As (2), but add concepts not in the starting thesaurus. This mitigates bias, but the central thesaurus now becomes a moving target.

(5) Start from a pool of terms from all participating languages and organize them into a conceptual framework, establishing term correspondence in the process. This approach results in a true "conceptual interlingua" not biased to any one language, but offering a home to multiple conceptual perspectives. This approach requires most effort.

Fig. 12. Building multilingual thesauri

6 Affordable implementation of knowledge-based approaches

The effort needed for constructing and maintaining any knowledge base, especially a well-structured multilingual thesaurus using method 4, is often forbidding, whence the attempts at constructing thesauri by statistical analysis of corpora. Fortunately, there is another way to reduce the effort, often drastically: Capitalize on the intellectual effort already available in a multitude of existing thesauri and dictionaries by automatically merging term relationships from many sources, as is done in UMLS (Unified Medical Language System) or analyzing dictionary definitions to extract term relationships (Ahlswede 1988). Learn from the structure of text by creating hypotheses on the part of speech and semantic features of words during parsing (Sonnenberger 1995), or deriving term relationships from user queries. A further expansion of this approach calls for collaborative development of thesauri and more comprehensive databases of concepts and terms made possible by computer technology as proposed in Soergel 1996. These approaches are summarized in Fig. 13.

Knowledge-based approaches require major investment for constructing the knowledge base.
Solutions
Use what is available (e.g. WordNet).
Reformat and integrate available sources into structured knowledge bases.
Use machine learning techniques based on text or query analysis for building or adding to a knowledge base, perhaps followed by human editing.
Provide integrated access to multiple sources and an environment for distributed collaborative knowledge base development.

Fig. 13. **Implementation of knowledge-based approaches**

Conclusion

It was the intent of this paper to present a high-level review of the contribution knowledge-based systems can make to cross-language retrieval, as exemplified in particular through the structure and function of thesauri and ontologies. Many of the ideas presented have been applied in operational or experimental systems, even though empirical results need to be interpreted with caution (Soergel 1994); others await application and testing

References

Ahlswede, Thomas, Martha Evens (1988). **Generating a relational lexicon from a machine-readable dictionary**. *International Journal of Lexicography* 1, no. 3 (Fall 1988): 214-237.

Ahlswede, Thomas, Martha Evens (1988). **Parsing vs. text processing in the analysis of dictionary definitions**. *Proceedings of the Association for Computational Linguistics (ACL) 26th Annual Meeting* (June 7-10, 1988): 217-224.

Ahlswede, Thomas, Martha Evens, Kay Rossi, Judith Markowitz (1986). **Building a lexical database by parsing Webster's Seventh New Collegiate Dictionary**. In JOHANNESEN, Gayle (ed)., *Advances in Lexicology: Proceedings of the University of Waterloo (UW) Centre for the New Oxford English Dictionary 2nd Annual Conference.* (November 9-11, 1986): 65-78.

Anders, Monika (1986). **Probleme der erarbeitung des mehrsprachigen thesaurus staat und recht des ISGI. Teil 1 [Problems encountered in the building-up of the multilingual thesaurus on state and law of the IISSS (Part 1)]**. *Informatik* 33, no. 2 (1986): 54-7.

Brodie, N.E. (1989) **Canadians use a bilingual union catalog as an online public catalog** . *Library Trends* 37 no. 4 (1989): 414-431.

Buchinski, E.J., W.L. Newman, M.J. Dunn (1976). **The automated subsystem at the National Library of Canada**. *Journal of Library Automation* 9 no. 4 (1976): 279-298.

Canisius, P.W., P.M. Lietz (1979). **Massnahmen zur Überwindung der sprachbarrieren in mehrsprachigen dokumentationssystemen** . *Datenbasen,* Datenbanken, Netzwerk. Hg. R. Kuhlen. München, 1979:127-153.

Commission of the European Communities (1977). *Third European Congress on Information Systems and Networks: Overcoming the Language Barrier, Luxembourg, 3-6 May 1977.* 1. München: Verlag Dokumentation, 1977.

Belal Musstafa, A.A., T.S. Tengku Mohd and M. Yusoff (1995). **SISDOM: A multilingual document retrieval system**. *Asian Libraries* 4, no. 3 (September 1995): 37-46.

Bennett, P.A., R.L. Johnson, John McNaught, Jeanete Pugh, J.C. Sager and Harold Somers (1986). *Multilingual Aspects of Information Technology*. Brookfield, VT: Gower Publishing Company, 1986.

Blake, P. (1992). **The MenUSE System for multilingual assisted access to online databases**. *Online Review* 16, no. 3 (June 1992): 139-46.

Boisse, J. A. (1996). **Serving multicultural and multilingual populations in the libraries of the University of California**. *Resource Sharing and Information Networks* 11, no. 1/2 (1996): 71-9.

Brendler, Gerhard (1970). **Der mehrsprachige thesaurus: Ein instrument zur rationalizierung des informationflusses [The multilingual thesaurus: An instrument for increasing the efficiency of information flow]**. *Informatik* 17, no. 4 (1970): 19-24.

Byrd, Roy J. (1987). **Dictionary systems for office practice**. *Automating the Lexicon: Research and Practice in a Multilingual Environment. Proceedins of the Linguistics Summer Institute Lexicon Workshop* (July 1987): 207-19.

Cousins, S.A. and R.J. Hartley (1994). **Towards multilingual online public access catalogs**. *Libri (International Library Review)* 44, no.1 (March 1994): 47-62.

Crofts, B.A. (1976). *Problems Associations with the Development and Maintenance of a Multilingual Thesaurus*. Report by Verina Horsnell, Chairperson. London, UK: British Library Research and Development, 1976.

Dijk, Marcel van (1966). **Un thesaurus multilingue au service de la corporation internationale [A multilingual thesaurus in the service of international corporation]**. *Bull De L'A.I.D.* 5, no. 4 (October 1966): 85-7.

Djevalikian, Sonia (1980). **Multilingual biblioservice**. *Bulletin-ABQ/QLA* 21, no. 1 (January-April 1980): 18-9.

Evens, Martha (1989). **Computer-readable dictionaries**. *Annual Review of Information Science and Technology (ASIS)* 24 (1989): 86-118.

Food and Agriculture Organization of the United Nations (1995). *Multilingual Soil Database/Food and Agriculture Organization of the United Nations, International Soil Reference and Information Centre, Institute of Natural Resources and Agro-Biology*. Rome: FAO, 1995.

Gilarranz, Julio, Julio Gonzalo, Felisa Verdejo (1997). **An approach to conceptual text retrieval using the EuroWordNet Multilingual Semantic Database**. *Cross-Language Text and Speech Retrieval.* Stanford University, California: American Association for Artificial Intelligence. (1997): 51-57.

Hlava, M.K. (1993) **Machine-aided indexing (MAI) in a multilingual environment**. *Proceedings of the Fourteenth Mational Online Meeting* (May 1993): 197-201.

Hlava, Marjorie, Richard Hainebach, Gerold Belonogov, Boris Kuznetsov (1997). **Cross language retrieval - English/Russian/French**. *Cross-Language Text and Speech Retrieval.* Stanford University, California: American Association for Artificial Intelligence. (1997): 66-72.

Horsnell, Verina (1976). *Workshop on Multilingual Systems: Report No. 5265 HC*. West Yorkshire: British Library Research & Development Reports, 1976.

Humphreys, Betsy L., Catherine R. Selden (1997). **Unified Medical Language System (UMLS): January 1986 thorugh December 1996: 280 citations**. *Current Bibliographies in Medicine*. Washington, D.C.: GPO (1997). [May 17, 1997: World Wide Web (WWW) Uniform Resource Locater (URL) is http://www.nlm.nih.gov/pubs/cbm/umlscbm.html.]

Hutcheson, H.M. (1995) **Preparation of multilingual vocabularies**. *Standardizing and Harmonizing Terminology: Theory and Practice*. Philadelphia, PA: American Society for Testing and Materials. (1995): 102-114.

ISO R1149-1969e. *Layout of Multilingual Classified Vocabularies.*

ISO Standard 5964. 1985. *Documentation Guidelines for the Establishment and Development of Multilingual Thesauri.*

Johannesen, Gayle (ed). *Advances in Lexicology: Proceedings of the University of Waterloo (UW) Centre for the New Oxford English Dictionary 2nd Annual Conference* (1986). Waterloo, Canada: UW Centre for the Oxford English Dictionary, 1986.

Lavieter, L., J.A. Deschamps, und B. Felluga (1991). **A multilingual environmental thesaurus** . *Technology Work in Subject Fields (Proceedings).* (October 12-14, 1991): 27-44.

Layout of Multilingual Classified Vocabularies. 1st ed. Geneva: ISO, 1969.

Lin, Chin-Yew 1997 (cyl@isi.edu, personal communic.)

Loth, K. (1993) *Aufbau eines hierarchisch strukturierten Thesaurus* . Juni 1993.

Loth, K. (1990) . **Aufbau eines eidgenössischen Bibliotheksthesauruus**. *ARBIDO-Revue* 5 no. 4 (1990): 111-113.

Loth, K. (1990) *Gebrauchsanweisung zur ETHICS-Sachkatalogisierung*. Zürich: ETH=Bibliothek. Dezember 1990.

Loth, K. **Computer searching of UDC numbers**. *Encyclopedia of Library and Information Science* 51.

McCallum, Sally and Monica Ertel (eds) (1994). *Automated Systems for Access to Multiliingual and Multiscript Library Materials: Proceedings of the Second IFLA Satellite Meeting, Madrid, August 18-19, 1993*. Germany: International Federation of Library Associations and Institutions, 1994.

Muraszkiewics, M., H. Rybinski and W. Struk (1996). **Software problems of merging multilingual thesauri** . *Compatibility and Integration of Order Systems (TIP/ISKO Meeting, Warsaw, 13-15 September, 1995).* Warsaw: 1996: 58-67.

Paschenki, N.A., S. Ya Kalachkina, N.M. Matsak and V.A. Pigur (1981). **Osnovnye printsipy sozdaniya mnogoyazychnykh informatsionno-poiskovykh [Basic principles in producing multilingual thesauri]**. *Nauchno Tekhnicheskaya Informatsiya* 2, no. 5 (1981): 17-9.

Pollitt, Steven 1997. http://www.hud.ac.uk/schools/cedar/-hibrowse

Ratzinger, M. (1991). **Multilingual product description (MPD), a European project** . *Terminology Work in Subject Fields (Proceedings)* 45. (October 12-14, 1991): 334-345.

Ritzler, C. (1990) **Comparative study of PC-supported thesaurus software**. *International Classification* 17 no. 3/4 (1990): S138-147.

Roelants, J.F.D. (1987). **Le catologage bilingue automatise en Belgique**. *International Cataloging* 16 (1987): 16-18.

Rolland-Thomas, P., G. Mercure (1989). **Subject access in a bilingual online catalogue**. *Cataloging and Classification Quarterly* 10 no 1/3 (1989): 141-163.

Roulin, Corentin (1996). **Bringing multilingual thesauri together: A feasiblity study**. *Compatiblity and Integration of Order Systems*, Warsaw: (1996): 123-35.

Sanderson, Mark (1994). **Word sense disambiguation and information retrieval**. *Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, 3-6 July 1994. Dublin, Ireland*. London: Springer-Verlag (1994): 142-151.

Schubert, K. (1985). **Parameters for the design of an intermediate language for multilingual thesauri**. *Knowledge Organization* 22, no. 3/4 (1995): 136-40.

Schuck, H.J. (1997). **Sprrachliche aspekte bei der Übersetzung von thesauri** . *Die Überwindung der Sprachbarrieren*. Bd.1 Hg. München: Komission der Europaischen Gemeinschaften, (1997): 399-414.

Semturs, F. (1997). **Information retrieval from documents in multilingual textual data banks**. *Third European Congress on Information Systems and Networks: Overcoming the Language Barrier: Luxembourg* I (May 1977): 463-76.

Soergel, Dagobert (1974). *Indexing languages and the thesauri: Construction and maintenance.* Los Angeles, CA: Melville; 1974. 632 p., 72 fig., ca 850 ref. (Wiley Information Science Series)

Soergel, Dagobert (1985). *Organizing Information: Principles of Data Base and Retrieval Systems.* Los Angeles, CA: Academic Press; 1985.

Soergel, Dagobert (1994). **Indexing and retrieval performance: The logical evidence**. *Journal of the American Society for Information Science* 45, no. 8 (1994): 589-599.

Soergel, Dagobert. 1996. **SemWeb: Proposal for an open, miltifunctional multilingual system for integrated access to knowledge base about concepts and terminology**. *Proceedings of the Fourth International ISKO Conference, 15-18 July 1996*, Washington, D.C. Frankfurt/Main: Indeks Verlag: 1996. (Advances in Knowledge Organization, v.5, 165-173).

Sonnenberger, G. (1995). *Der wit-interpreter: Ein textanalysesystem, das sein eigene wissenbasis durch die analyse von texten erweitert (The wit-interpreter. A*

text analysis system that adds to its own knowledge base during text analysis) (1995). Konstanz: Hartung-Gorre Verlag, 1995. (PhD. Thesis).

Stancikova, Pavla. (1996) **International integrated database systems linked to multilingual thesauri covering the field of environment and agriculture**. *Compatiblity and Integration in Order Systems (TIP/ISKO Proceedings)*. Warsaw, (1996): 163-9.

Still, J. M. (1993). **Multilingual and international databases**. *CD-ROM Professional* 6, no. 6 (November 1993): 72-6.

U.S. Department of Health & Human Services, National Institute of Health, National Library of Medicine (NLM). (1993) **Unified Medical Language System (UMLS) Semantic Network** . (May 17, 1997: World Wide Web Uniform Resource Locater is http://www.gsf.de/MEDWIS/pg_term/umls_net.html)

Velho Lopes, Roseane R. (1989). **Automated access to multilingual information: A Brazilian case study (Science and CIENTEC)**. *Information Development* 5 (July 1989).

VTLS, Inc. (1997) *Library Automation and Information Management. [*May 17, 1997: World Wide Web (WWW) Uniform Resource Locatoer (URL)- http://www.vtls.com]

Walker, Donald E., Antonio Zampolli and Nicoletta Calzolari (eds) (1995). *Automating the Lexicon: Research and Practice in a Multilingual Environment* . Oxford, New York: Oxford University Press, 1995.

Warwick, Susan. (1987). **Automated lexical resources in Europe: A survey**. *Automating the Lexicon: Research and Practice in a Multilingual Environment. Proceedings of the Linguistics Summer Institute Lexicon Workshop* (July 1987): 329-63.

White, John (1988). **Determination of lexical-semantic relations for multi-lingual terminology structures**. *Relational Models of the Lexicon,* 326. Cambridge, UK: Cambridge University Press, 1988.

Wolff-Terroine, M. (1975). **Multilingual systems** (1975). *Second European Congress on Information Systems and Networks: Luxembourg* (May 1975): 149-58.

Wolff-Terroine, Madeleine (1983). **SABIR, Un logiciel de gestion de thesaurus multilingues" [SABIR, a program for managing multilingual thesauri]**. *MIDIST Bulletin D'Information* 2 (April 1983): 8.

Mini-Exhibit of Multilingual Thesauri

International Standard 5964: Documentation guidelines for the establishment and development of multilingual thesauri. First edition; 61 p; American National Standards Institute; 1985.

Layout of multilingual classified vocabularies: ISO Recommendation R 1149. First edition; 23 pages; International Organization for Standardization, Switzerland; 1969.

Molho, Emanuel. **The dictionary catalogue**. Second edition., 178 pages; French & European Publications, Inc., New York; 1989. (A bibliography of mono-, bi-, and multilingual dictionaries)

International classification and indexing bibliography; ICIB 1; Classification systems and thesauri, 1950-1982. 143 p; INDEKS Verlag, Frankfurt; 1982.
Includes bibliography of editions in multiple languages of
Universal Decimal Classification (UDC) Library of Congress Classification (LCC)
Dewey Decimal Classification (DDC) Library of Congress Subject Headings (LCSH)

Thesaurus EUROVOC: Official journal of the European communities. Office for Official Publications of the European Communities; 1995.

Viet, J. and Georges van Slype. **EUDISED Multilingual thesaurus for information processing in the field of education**, English version. 307 pages. Mouton Publishers, Berlin,, New York, Amsterdam; 1984.

EUDISED R&D Bulletin, volume 45, ISSN 0378-7192;. 127 pages.; K.G. Saur, Munich; 1993.

Food and Agriculture Organization of the United States. **AGROVOC multilingual agricultural thesaurus**. Second edition, English version; 798 pages; APIMONDIA, Rome; 1992. (Not latest)

International Atomic Energy Agency. **INIS: Thesaurus**. 887 p. and **INIS multilingual dictionary**. 314 p. IAEA, Vienna; 1993, 1983 (not latest editions).

Organization for Economic Cooperation and Development. **Multilingual dictionary of fish and fish products**. Fourth edition., 352 pages; Fishing News Books, Cambridge; 1995. LCC Ql64.7.M84.1995

Centre for Computer-Aided Egyptological Research. **Multilingual Egyptological thesaurus**. http://www.ccer.ggl.ruu.nl/thes/thsaur.html. 1995.

Department of the Army, the Navy, and the Air Force. Medical Service multilingual phrase book. United States Government Printing Office, Washington, D.C.; 1971. Su Doc D 101.22:40-3

*****

The Alcohol and Other Drug Thesaurus: A guide to concepts and terminology in substance abuse and addiction, vol. 1-4. 2.ed.; U. S. Department of Health and Human Services; 1995. approx. 2,500 p. (Monolingual, example for structure) :http://etoh.niaaa.nih.gov/AODVol1/Aodthome.htm