Large multilingual vocabularies. Structure and software requirements

Dagobert Soergel
College of Library and Information Services
University of Maryland

ds52@umail.umd.edu
www.clis.umd.edu/faculty/soergel/

Outline

Purposes

Structure

Construction

Software requirements

Purposes of multilingual thesauri

Controlled vocabulary for a multilingual IR system

Support for cross-language free-text searching

Support for translation

Structure of multilingual thesauri
Key issues

Conceptual systems in different languages differ

What concepts are lexicalized differs from language to language

Translation of an English thesaurus into German does not make a German thesaurus

Multilingual thesaurus problems

English	**German**
simian monkey ape	Affe *niederer Affe* Menschenaffe

bone *bone (vertebrate other than fish)* *fish bone*	*Knochen, Gräten* *Knochen* *Gräten*

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

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

Multilingual thesaurus problems

English	**German**
timepiece	Uhr
clock	*Wanduhr, Standuhr, Turmuhr*
wall clock	Wanduhr
standing clock	Standuhr
tower clock	Turmuhr
watch	*Taschenuhr, Armband-uhr*
pocket watch	Taschenuhr
wrist watch	Armbanduhr
alarm clock	Wecker

Multilingual thesaurus problems

English	**German**
b*lanket, rug, carpet* blanket *rug, carpet* rug (or carpet) *long, narrow rug* wall-to-wall carpet *hanging rug*	Teppich Betteppich Bodenteppich *loser Bodenteppich* Läufer Teppichfußboden Wandteppich

Multilingual thesaurus problems

Other difficulty:
Two terms mean almost the same thing but differ slightly in meaning or connotation:

English: *alcoholism*
French: *alcoholisme*

English: *vegetable* (includes potatoes)
German: *Gemüse* (does not include potatoes)

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.

Multilingual thesaurus problems
Term-concept vs term-term relationships

English term 1		German term 1
English term 2	Concept	German term 2
English term 3		German term 3
English term 1		German term 1
English term 2	Concept	German term 2
English term 3		German term 3

Building a multilingual thesaurus 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.

Principles

Develop common conceptual structure integrating perspectives from multiple languages.

Harmonize concepts where possible, keep concepts where necessary

Invent a term if a concept is not lexicalized in a language

Building a multilingual thesaurus
Approaches

(by increasing complexity and quality)

(1) **Start from monolingual thesaurus and translate**

- ! does not capture concepts lexicalized only in a target language
- ! biased to the conceptual structure underlying the starting language.
- ! Misses synonyms in target language.

(2) **Start from a monolingual thesaurus as the center.**
Collect terms from other languages
Map to concepts in the central thesaurus.
Suffers from similar bias toward the starting language as (1), but may cover more synonyms in the other languages.

Building a multilingual thesaurus
Approaches 2

(3) Start from central thesaurus as in (2)
Collect terms from other languages
Group terms in each language into synonym sets, each corresponding to a concept
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.
The central thesaurus remains unchanged.

(4) As (3), but
add concepts not in the starting thesaurus.
Mitigates bias, but the central thesaurus now becomes a moving target.

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

Software requirements for multilingual thesauri 1 Mundane

Character sets

Term normalization (esp. singular/plural) in multiple languages

Term ID: Language + character string

D: rot (the color red)
E: rot (to rot)

Spell check in multiple languages

Translating definitions and scope notes

Software requirements for multilingual thesauri 2 More difficult

Support for term collection from multiple sources in multiple languages

Support for inferring suggested relationships

Excellent support for developing structured hierarchies

- **Support for alternate hierarchies**
- **Support for maintaining a meaningful sequence**

Concept-term and term-term relationships with indication of semantic closeness