Multilingual Thesaurus Construction: Integrating the Views of Different Cultures in One Gateway to Knowledge and Concepts[1]

Michèle Hudon

Chargé d'enseignement, École de bibliothéconomie et des sciences de l'information, Université de Montréal, Canada

Michèle Hudon consults widely in the area of bilingual thesaurus construction. She was the first editor of the *Canadian Education Thesaurus*, and, more recently, coordinated the development of the bilingual *Canadian Literacy Thesaurus*. She is a Ph.D. candidate at the Faculty of Information Studies, University of Toronto.

Hudon, M. (1997). **Multilingual thesaurus construction: Integrating the views of different cultures in one gateway to knowledge and concepts.** *Knowledge Organization*, 24(2), 84-91. 11 refs.

ABSTRACT: General linguistic and specific semantic problems arising in multilingual thesaurus construction are well defined in the various textbooks and in the guidelines covering this area. Many details are provided on the "conceptual equivalence" issue, and various ways of dealing with conceptual divergence are described. But when discussing semantic solutions, display options, management issues, or use of technology, specialists and guidelines seldom, if ever, go as far as commenting on whether or not a particular option is truly respectful of a language and its speakers. This paper, based on the premise that in a multilingual thesaurus all languages are equal, reviews the options and solutions offered by the guidelines to the developer of specialized thesauri. It also introduces other problems of a sociocultural, and even of a truly political nature, a prominent feature in the daily life of the thesaurus designer with which the theory and the guidelines do not deal very well.

1. Introduction

With the growing number of information databases now available on world wide electronic networks, the "language barrier" has become an even more critical issue than it has ever been before. One solution to increasing communication difficulties is to create semi-artificial controlled-access languages, which, if they are efficient, will allow foreign users to access our data and allow us to access theirs. Multilingual thesauri appear as potentially powerful tools in such a context; they are widely used in information transfer systems maintained by the European Community, and in officially bilingual countries, such as Canada.

Most information users are aware of the very real problems which have traditionally been associated with multilingual thesauri: 1) that of stretching a language to make it fit a foreign conceptual structure to the point where it becomes barely recognizable to its own speakers; 2) that of transferring a whole conceptual structure from one culture to another whether it is appropriate or not; 3) that of translating literally terms from the source language into meaningless expressions in the target language, etc.

New developments in the field of multilingual thesaurus construction make it possible, and it is becoming common practice, to build multilingual thesauri from the ground up, in complete respect of all languages involved, with results that reflect better the various conceptual and terminological structures with which potential end-users (indexers and searchers) are most familiar.

It is useful to remember at all times that there is more to multilingual thesaurus development than finding equivalents for concepts and terms. There is a definite cultural dimension to the process, and in fact it might soon be more appropriate to refer to multicultural thesauri, rather than to multilingual thesauri. There is also a political dimension to multilingual thesaurus construction, especially in dealing with languages which are not, contextually, on the same "standing". Canada, for example, has a good grounding in multilingual thesaurus construction, but it remains a struggle to make sure that French (the minor-

ity language) and English are given equal treatment in the many thesauri designed and used in the country.

Designers of multilingual thesauri face many substantial challenges and obstacles; some are of an administrative nature, some are of a linguistic/semantic nature, some are technology-related. The more specialized a thesaurus, the more specific its descriptors, the more difficult it is going to be to develop and manage in a multilingual environment.

Thesaurus designers are provided with formal guidelines that can help them in their task. In this paper, we will refer to the *Guidelines for the establishment and development of multilingual thesauri* [ISO 5964:1985]. The *Guidelines* define and illustrate problems, and describe a range of optional procedures for dealing with them. Some of these proposed options and solutions are reviewed in the following pages, from a perspective of giving equal treatment to each language represented in the thesaurus.

2. Nature and Functions of the Multilingual Thesaurus

The multilingual thesaurus is more than just the "putting together" of several monolingual thesauri. Each linguistic version of a multilingual thesaurus can be used independently from the others, but is connected with all the others and would not exist without them.

The true multilingual thesaurus offers full conceptual and terminological inventories for each language represented; most importantly, it presents a fully developed thesaurus structure (i.e. all semantic relationships of equivalence, hierarchy, and affinity) in each one of the languages of the thesaurus, so that a user consulting whichever linguistic version is most appropriate for her/him gets an equal amount of valuable semantic information. A thesaurus which adopts a source language, and then provides descriptor equivalents in other languages, but not a full semantic structure, is not, in the perspective of language equality, a true multilingual thesaurus.

Multilingual thesauri serve mainly as indexing and retrieval aids in multilingual information systems. When a multilingual thesaurus is available, documents can be indexed in one or more of several languages (that of the document, of the information centre, etc.) Searches can be conducted in a different language (most often the language of the user). The thesaurus then plays the role of switching language, and facilitates interlinguistic communication.

The multilingual thesaurus is also very useful to the individual who wants to query a database which "understands" only a foreign language. This user will find in the appropriate multilingual thesaurus the controlled terms needed to build a search strategy.

3. Developing a Multilingual Thesaurus: Three Approaches, Two Perspectives

3.1 Approaches:

There are three standard approaches to developing a multilingual thesaurus.

A. Translation in one or more new languages of an existing monolingual thesaurus:

This approach has been very popular in the past, mostly for economic reasons. The approach, obviously, does not allow for equal treatment of all languages involved. The source language naturally becomes the dominant language, and the resulting product cannot reflect adequately the target culture(s). A monolingual thesaurus is always culturally biased, and a straight translation might lead to a form of "cultural imperialism".

B. Merging and/or reconciliation of several existing monolingual thesauri:

Although more acceptable already, this second approach comes with very serious practical problems; it is the most difficult to manage intellectually, as each one of the thesaural structures available might, and probably will, differ considerably as to extent and depth of coverage, degrees of pre-coordination, levels of specificity, etc. And here again, it is likely that the language and structure of the larger or most developed thesaurus will become dominant, and that the structures of the other thesauri will be adjusted to make them "fit" the dominant one.

C. Simultaneous development of distinct linguistic versions:

The third approach offers stronger guarantees for equal treatment of all languages. Each language becomes in turn source language, so that the target language, the one which is often artificialized, is not always the same. Each culture described in thesaurus terms contributes to the structuring of the tool; adjustments and concessions are not always made by the same party. In multilingual terminology work, a similar process is called "harmonization of terminology"[2].

3.2 Perspectives:

A most important decision has to be made early regarding identity and symmetry of semantic structures in the various linguistic versions of the thesaurus.

There are two views on this matter.

A. Identical and symmetrical structures:

The most common view is that all linguistic versions of a multilingual thesaurus must be identical and symmetrical; each descriptor must have one and only one equivalent in a target language (no single-to-multiple equivalence is allowed), and be related to the same terms. Complete structural identity seems neces-

sary in computerized systems (as they stand today) where analysis, indexing, searching, etc. are done without human intervention. Unfortunately, this artificializes all the languages involved, by forcing equivalences where they do not exist (when one source concept/term = no target concept/term), by eliminating true equivalences where they do exist (when one source concept/term = two or more target concepts/terms), by generating semantically incorrect or illogical hierarchies (when a concept/term belongs to a hierarchy in the source language, but to a different hierarchy in the target language), etc.

B. Nonidentical and nonsymmetrical structures:

It seems preferable to accept nonidentical and nonsymmetrical structures in a multilingual thesaurus. The number of descriptors in each linguistic version should be allowed to vary: concepts which exist in a culture are represented in its language, but if those same concepts do not exist in another culture, it is unlikely that equivalent verbal representations will be available. Paradigmatic links, hierarchical relationships for example, are not necessarily recognized as valid in all natural languages. In a multilingual thesaurus, when two top terms or broad terms are inexact or partial linguistic equivalents, they may have a slightly different extension, and consequently different subordinate terms. A multilingual thesaurus in which the structures are allowed to differ is more likely to faithfully reflect the universe of concepts and terms in each one of the cultures and languages represented.

4. Managerial Isues

4.1 The Administrative Structure

The development of a thesaurus, in one or more than one languages, is necessarily a team effort within a more or less centralized and rigid administrative structure. A semi-centralized managerial structure, with decision-making delegated to a small group of designated representatives from all organizations involved, is likely the most appropriate for multilingual thesaurus design. Within such a structure, all parties are given equal responsibility with respect to the end product, but the development work is centralized. Discussions are more productive, and decisions can be made more quickly and more efficiently than within a totally decentralized structure. In a decentralized structure, the development work is done at various sites, decisions may be inconsistent, and consensus is ultimately more difficult to reach.

4.2 The Thesaurus Workers

Each linguistic version of a multilingual thesaurus must be developed by individuals who possess a deep knowledge of the conceptual and terminological makeups of their first language. Knowledge and experience of one or more of the other languages represented in the thesaurus are an asset, especially when it becomes necessary to determine whether or not different linguistic versions of a descriptor are truly equivalent.

4.3 The Process

All parties involved in the development of a multilingual thesaurus must share a common view of the resulting product. All must subscribe to the principle of linguistic equality.

In practice, there are two ways in which the various linguistic versions of a multilingual thesaurus can be developed simultaneously:

A. One person/one team is working on all linguistic versions of the thesaurus;

B. Several persons/several teams are working independently, but according to the same specifications, on distinct linguistic versions of the thesaurus, and merging/reconciling the results of their work at critical points in the process (eg. after initial compilation of candidate descriptors, after preliminary classification of candidate descriptors, after identification of linguistic equivalences, etc.) When these individuals/teams meet, communication difficulties are most likely to arise. Even if all participants in the group know a common language, it will normally be the first language for some, and a second or third language for the others. Ways of ensuring that everybody has equal opportunity to make personal views known and personal decisions understood must be devised.

The second one of the above options is undoubtedly more conducive to the production of a tool in which the principle of language equality has been respected.

4.4. The Sources

In a multilingual thesaurus development context, all candidate descriptors must be extracted from original language sources rather than from translations. Distinct termbanks should be developed independently for each language represented, and reconciled at the end of a set term collection period.

5. Linguistic/Semantic Issues

Natural languages are more than inventories of words: they are a true reflection of conceptual universes which vary from one culture or society to another. Conceptual differences appear more frequently in references to specific entities, processes, and relationships. Life would be easier for thesaurus designers

if all terms used in multilingual networks represented always general and simple ideas. The tendency in modern thesauri, however, is towards more depth of coverage, greater specificity, representation of increasingly complex concepts, and consequently more precoordination.

In the multilingual thesaurus construction process, this translates into well-documented difficulties in determining interlinguistic equivalence, an operation that is at best delicate, and can be at times controversial. The establishment of equivalences is especially difficult if concepts do not have a stable lexical support, as is often the case in the special languages of the social sciences and the humanities. Every natural language carries its own denotative, connotative, evaluative, and emotional implications. In a multilingual thesaurus, equivalent descriptors, ideally, should have equivalent implications.

The *Guidelines* describe five degrees of interlanguage equivalence among concepts and terms: a. the exact equivalence (interlinguistic synonymy); b. the inexact equivalence (interlinguistic quasi-synonymy, with a difference in viewpoint); c. partial equivalence (interlinguistic quasi-synonymy, with a difference in specificity); d. single-to-multiple equivalence (too many terms or not enough terms); e. non-equivalence.

All cases of nonexact equivalence (b., c., d., and e. above) must be looked at with care in a multilingual thesaurus environment. The last two cases (i.e. single-to-multiple equivalence, and non-equivalence) are the most difficult for a thesaurus designer to deal with, especially in multilingual thesauri with identical and symmetrical structure, in which every descriptor must have an equivalent and cannot have more than one equivalent. Various potential solutions to the problems caused by nonexact equivalence are recommended by the *Guidelines*; some of them are more appropriate in a perspective of giving equal status to all languages in the thesaurus.

5.1 Single-to-Multiple Equivalence

There are two distinct cases of single-to-multiple equivalence.

5.1.1 The target language contains more than one equivalent to the source term (too many target terms).

The solutions offered by the *Guidelines* are:
Solution a: the creation of a precombined descriptor in the target language (Figure 1).

Source	Target
Problem FUELS	CARBURANT COMBUSTIBLE
Solution FUELS	CARBURANT + COMBUSTIBLE EP Carburant EP Combustible

Figure 1. Single-to-multiple equivalence - Case 1 - Solution a

Solution b: a modification/specification of the source term, eg. by addition of a qualifier (Figure 2).

Source	Target
Problem FUELS	CARBURANT COMBUSTIBLE
1.1 Solution FUELS (MOTORS) FUELS (HEATING)	CARBURANT COMBUSTIBLE

Figure 2. Single-to-multiple equivalence - Case 1 - Solution b

Solution c: the establishment of one or more non-descriptor(s) in the target language, with a link to the preferred term (Figure 3).

Source	Target
Problem FUELS	CARBURANT COMBUSTIBLE
Solution FUELS	CARBURANT EP Combustible

Figure 3. Single-to-multiple equivalence - Case 1 - Solution c

In the case of single-to-multiple equivalence where the target language offers more than one equivalent to a source term, it appears that solution c. above, since it does not affect the wording or form of any terms, is the most acceptable option in a perspective of language equality. If it happens that one of the target equivalents is identical in wording to the source term, it should preferably be selected as descriptor, even if it is not the most commonly used term in the target language, to avoid the confusion and the processing difficulties that might arise if the same term is a valid term in a language, but a nonpostable term in another.

5.1.2 The target language can only represent the source concept through a combination of terms (not enough target terms).

The solutions offered by the *Guidelines* are:

Solution a: a formal recommendation, in the target language, to use many descriptors, if the system allows it (Figure 4).

Source	Target
Problem	
HEATING	CHAUFFAGE
SOLAR ENERGY	ÉNERGIE SOLAIRE
SOLAR HEATING	?
Solution	
HEATING	CHAUFFAGE
SOLAR ENERGY	ÉNERGIE SOLAIRE
SOLAR HEATING	CHAUFFAGE + ÉNERGIE SOLAIRE

Figure 4. Single-to-multiple equivalence – Case 2 – Solution a

Solution b: the creation of an equivalent (neologism/coined term) (Figure 5).

Source	Target
Problem	
HEATING	CHAUFFAGE
SOLAR ENERGY	ÉNERGIE SOLAIRE
SOLAR HEATING	?
Solution	
HEATING	CHAUFFAGE
SOLAR ENERGY	ÉNERGIE SOLAIRE
SOLAR HEATING	CHAUFFAGE SOLAIRE

Figure 5. Single-to-multiple equivalence – Case 2 – Solution b

Solution c: the establishment of one or more non-descriptor(s) in the source language, with a link to the preferred term(s) (Figure 6).

Source	Target
Problem	
HEATING	CHAUFFAGE
SOLAR ENERGY	ÉNERGIE SOLAIRE
SOLAR HEATING	?
Solution	
HEATING	CHAUFFAGE
SOLAR ENERGY	ÉNERGIE SOLAIRE
SOLAR HEATING	
USE HEATING AND SOLAR ENERGY	

Figure 6. Single-to-multiple equivalence – Case 2 – Solution c

Option c. is, again, the least artificial solution to the problem.

5.2 Non-Equivalence

"Orphans" (i.e. descriptors appearing in one linguistic version of a multilingual thesaurus, but without equivalent in at least one of the other versions) are naturally not tolerated in thesauri where identity of structures is required.

A simple solution to the non-equivalence problem is the removal of the "orphan" from the source language lexicon, if it appears to represent a much more specific concept than the rest of the vocabulary. But cases of non-equivalence can obviously not always be solved so easily.

The solutions offered by the *Guidelines* are:

Solution a: a change of status for the "orphan", which is transformed into a non-descriptor and linked to a descriptor with which it shares many essential characteristics (Figure 7).

Source	Target
Problem	
TEENAGERS	?
Solution	
ADOLESCENTS	ADOLESCENT
UF Teenagers	

Figure 7. Non-equivalence – Solution a

Solution b: the import of the source term into the target language (Figure 8).

Source	Target
(French) RÉGIME PÉDAGOGIQUE	*(English)* RÉGIME PÉDAGOGIQUE
(French) MARKETING	*(English)* MARKETING
	EP Mercatique

Figure 8. Non-equivalence – Solution b

Solution c: the creation of an equivalent (neologism) (Figure 9).

Source	Target
LATCHKEY CHILDREN	ENFANT À CLÉ DF Enfant dont les parents ne sont pas à la maison et qui est muni d'une clé pour rentrer chez lui après l'école
ECOFEMINISM	ÉCOFÉMINISME

Figure 9. Non-equivalence – Solution c

In cases of non-equivalence, solution a., which does not affect the wording or form of any of the terms, might look like the ideal option, providing that the "orphan" actually has a close relative in the thesaurus lexicon. The import of a source term into a target language is more regularly retained as a preferred option however; users from a particular culture might still have to access information on entities, processes, and relationships that exist only in another society.

The creation of neologisms is never the best solution. A thesaurus is not a terminological termbank. The role of a thesaurus is not to bring about changes in a language, it is rather to reflect the specialized use of that language in certain segments of a society.

6. Technology-Related Issues

Multilingual thesauri were for a long time the preserve of major national and international organizations. It is probably safe to say that most of these organizations developed and maintained their thesauri with the help of custom-designed software. Smaller organizations, which joined more recently the ranks of information providers and heavy information users, are now also investing important resources into multilingual thesaurus development. Looking for off-the-shelf software to facilitate the thesaurus building process, they realize quickly that software which is perfectly suitable and efficient for monolingual thesauri is not necessarily appropriate for multilingual ones.

Software which permits the creation of one or more fields for linguistic equivalents, but does not permit the creation of separate linguistic versions for each language represented in a multilingual thesaurus, does not qualify as multilingual thesaurus construction software in the perspective of language equality. Software vendors will tell you that their products do handle other languages (with diacritic even). But do they allow you to rotate source and target languages? To separate records for descriptors in one language from records for descriptors in another? Do they simply perform a translation operation once a descriptor record has been created in the source language, thus commanding automatically identical and symmetrical structures in all target languages?

Two standard file structures would appear to give equal status to all languages in a multilingual thesaurus:

A. In the first type of file, distinct records are established for each linguistic version of a descriptor, with a possibility of sorting on language for report production;

B. In the second type of file (Figure 10), a single record contains all linguistic versions of a descriptor and of its related terms. More complex sorting operations are needed to produce distinct linguistic versions.

T100	Descriptor (source language)
T110	Descriptor (target language 1)
T400	UF (source language)
T410	UF (target language 1)
T500	BT (source language)
T510	BT (target language 1)
T600	NT (source language)
T610	NT (target language 1)
T700	RT (source language)
T710	RT (target language 1)
T800	SN (source language)
T810	SN (target language 1)

Figure 10. Integrated record structure for multilingual thesaurus construction

Although the second type of file structure still allows for variations in the content of each field and authorizes feedback from target language to source language, it should be noted that it does require that a source language be designated.

7. Thesaurus Display

Display of thesaural data, whatever presentation format is chosen, must be complete and equally clear in all linguistic versions of a multilingual thesaurus. Standard codes (BT, NT, etc.), which indicate the nature of terms and of relationships among terms, must be given in the language of the descriptor or as a language independent symbol (>, <, etc.) Introductions and instructions for use, classified displays, keyword indexes, etc. must be available in each language.

In multilingual thesauri available in print form, data have been presented in one of two ways.

A. Parallel presentation of all linguistic versions on one page (Figure 11). In a perspective of language equality, this type of presentation is acceptable only if each one of the languages of the thesaurus appears in turn in the "left column", which commands the filing sequence.

FAMILLE	FAMILIES
EQ Families	EQ Famille
TS Famille à faible revenu	NT Low income families
Famille d'accueil	Foster families
Famille monoparentale	Single parent families
TA Droit de la famille	RT Family law
Finances familiales	Family finances
Violence familiale	Domestic violence
FAMILLE À FAIBLE REVENU	LOW INCOME FAMILIES
EQ Low income families	EQ Famille à faible revenu
FAMILLE d'ACCUEIL	FOSTER FAMILIES
EQ Foster families	EQ Famille d'accueil
TA Adoption	RT Adoption

Figure 11. Display of thesaurus data – parallel presentation

B. Physical separation of all linguistic versions (Figure 12), in distinct volumes or in separate sections in a volume, each with its own introduction, indexes, appendices, etc. This type of presentation is the most respectful of all languages and cultures involved. This has been the preferred option in Canada for quite some time. Electronic versions of multilingual thesauri are now most often displayed according to this basic model.

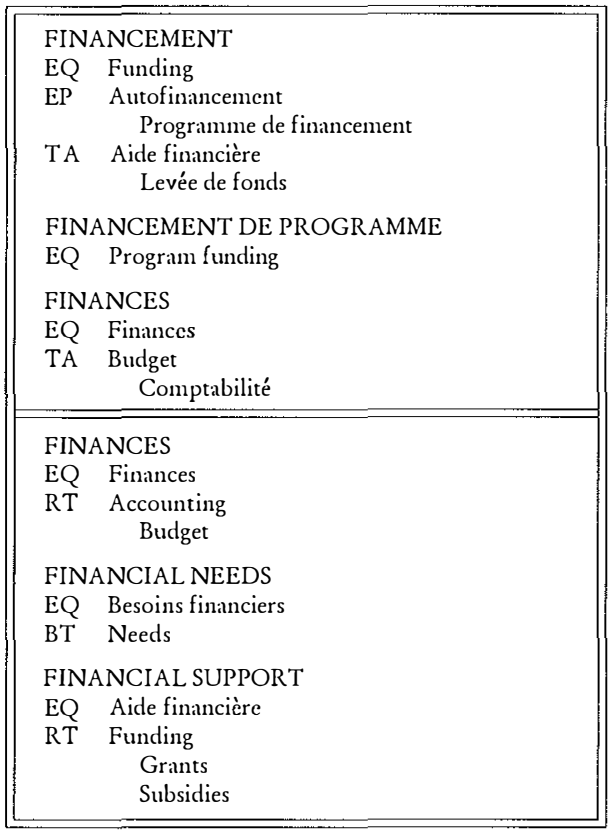

Figure 12. Display of thesaurus data – separation of each version

8. Conclusion

It is very likely that every decision made during the course of developing a thesaurus affects in some way the resulting product. In a multilingual thesaurus development context, it must be remembered that every decision may affect the status of each language represented in the thesaurus.

The difficulties and real problems described in this paper are only a few of the many obstacles which the multilingual thesaurus designer will encounter. The analysis and eventual selection of a solution to a specific problem, whether this solution is a recommendation of the *Guidelines* or not, must always take into account the issue of language equality.

True equality for all languages in a multilingual thesaurus has a better chance of being achieved if the following global requirements are met:

- the thesaurus is built within a semi-centralized administrative structure, with representatives of each language/culture on the decision-making team;
- all linguistic versions of the thesaurus are developed simultaneously from the ground up;
- the thesaurus designers are native speakers of the language in which they work, with a good knowledge of the other languages involved;
- distinct terınbanks are built independently for each language with terms found in source language documents;
- identity and symmetry of structures are not required across the various linguistic versions of the thesaurus, and single-to-multiple equivalence, "orphans", and variations in hierarchies, etc. are allowed;
- the use of neologisms is very restricted if allowed at all;
- thesaurus development software which allows for nonidentity of descriptor records and for rotation of source and target languages is used;
- physically separate displays for each language represented are produced.

If these requirements are satisfied, the resulting product should represent more accurately the various conceptual environments described, and consequently be more readily accepted, and ultimately more useful to all its potential users.

9. Notes

1. This paper was delivered at *Research and Development in Electronic Access to Fiction, Multicultural Knowledge Transfer and Cultural Mediation via Networks*, a research seminar sponsored by the Royal School of Librarianship, Copenhagen, Denmark, November 13, 1996.

2. For a detailed presentation of this process, see Gilreath, C. T. (1992). Harmonization of terminology: An overview of principles. *International Classification*, 19(3), 135-139.

10. Bibliography

Aitchison, J., & Gilchrist, A. (1987). *Thesaurus construction: A practical manual.* (2nd ed.). London: Aslib.

Association française de normalisation. (1990). *Principes directeurs pour l'établissement des thésaurus multilingues, Z47-101.* Paris: AFNOR.

Hudon, M. (1995). *Le thésaurus: conception, élaboration, gestion.* Montréal: Asted.

International Organization for Standardization. (1986). *Documentation – Guidelines for the establishment and development of monolingual thesauri, ISO 2788.* (2nd ed.). Geneva: ISO.

International Organization for Standardization. (1985). *Documentation – Guidelines for the estab-*

lishment and development of multilingual thesauri, ISO 5964. Geneva: ISO.

International scientific symposium on multilingual thesauri, 8-10 October 1973 in Berlin: proceedings. (1974). Berlin: Leitstelle Politische Dokumentation.

Lancaster, F. W. (1985). *Thesaurus construction and use: A condensed course.* Paris: Unesco.

Lancaster, F. W. (1986). *Vocabulary control for information retrieval.* (2nd ed.). Arlington, Va.: Information Resources Press.

Soergel, D. (1974). *Indexing languages and thesauri: Construction and maintenance.* Los Angeles: Melville.

Somers, H. L. (1981). Observations on standards and guidelines concerning thesaurus construction. *International Classification, 8(2),* 69-74.

Van Slype, G. (1987). *Les langages d'indexation: Conception, construction et utilisation dans les systèmes documentaires.* Paris: Ed. d'Organisation.

Michèle Hudon, École de bibliothéconomie et des sciences de l'information, Université de Montréal, C.P. 6128, Succursale Centre-ville, Montréal, Qué. H3C 3J7, Canada.