Accessing documents and information in a world without frontiers

Michèle Hudon

This extended version of a paper presented at the 30th Annual Conference of the American Society of Indexers (Seattle, 13–16 May 1998) presents an overview of common problems affecting the transfer of information across natural languages. A brief description of the most promising solutions for overcoming information retrieval difficulties in worldwide multilingual networks is provided. Selected quotes dating back to the 1970s and early 1980s show that problems and solutions in this area remain pretty much the same today as they were then.

Multilinguality in the new information world

Indexers are now working in a world that is very different from that of their predecessors of no more than a few decades. Rapid transportation and increasingly sophisticated technology have shrunk the size of the globe, enabling instantaneous communication between individuals, and providing seemingly easy access to foreign sources of information. Technologically speaking, it is now possible to make information accessible to almost anybody, almost anywhere, and at almost any time. If information flows freely across most frontiers, we must be reminded, however, that 'much remains to be done before linguistic barriers can be surmounted as effectively as geographic ones' (Oard 1997). It is one thing to know that potentially useful information sources are available, and quite another to find them and to make sense of them! We know how difficult it can be to retrieve pertinent information when resources are in a single natural language; what a challenge it becomes when several natural languages are in simultaneous use, as is the case, for example, on the Internet. Most of the world's population is bilingual or multilingual at various degrees, but most people are only elective bilinguals or multilinguals; and if

> a passive knowledge of a foreign language will most likely be sufficient to a user to peruse an abstract or even a complete document (...), unless one is prepared to go to a great deal of trouble, it will certainly not be sufficient to define the optimum strategy for the formulation of a complex question put to a machine-readable database, and especially if the latter has only free-text capabilities (Iljon 1978, 130).

The language barrier has been presented as a set of difficulties experienced in ascertaining the information wrapped in foreign language for utilization in the solution of local problems (Yasmin 1977). Although proposed 22 years ago, this definition is still valid, as are most comments made at the time on the necessity of developing user-friendly multilingual information systems. It is remarkable that decades later, so many problems identified in the early 1970s by European information specialists remain unsolved. A renewal of interest for these 'old' issues is currently observed, spurred by the expansion of wide-area networks and the increasing participation of former developing countries to scientific research and reporting.

Over time, many natural languages have had their turn at being 'the' language for cultural, scientific, and business purposes. This language was naturally that of the dominant nation or civilization of the time. It was likely in the interest of the reigning authority to make information available in the language of the elite, one that the masses were not likely to understand, and only occasionally was there any attention paid to the needs of those who could not read the dominant language. Wellisch reminds us that 'the earliest examples of indexes in more than one language and in several scripts are those found in herbals of the late 15th and early 16th century' (1978, 81).

Multilingualism in information systems became a major preoccupation of Europe in the 1960s. The global information network was still a dream then, but researchers knew that such a network would necessarily be multilingual. The importance of providing multilingual access became clear as the consequences of not providing such access were revealed. A far-reaching consequence of not having access to scientific reports published in lesser known and read languages could be a considerable delay in scientific progress and a considerable waste of time and funds in re-doing research which had already proven conclusive or inconclusive.

It appeared early on that English would become the *lingua franca* for the communication of scientific and technical information by the end of the 20th century, and to a great extent it has. But this is no longer considered an acceptable means of removing the language barrier, 'for political and nationalistic reasons if for no other' (Lancaster and Smith 1983, 69). It is a fact that if 'from a strictly objective viewpoint, language is a purely utilitarian medium for the transmission of ideas between individuals (...), on the other hand, it is the most conspicuous expression of the cultural life of a nation or group of nations' (Kertesz 1977, 274).

The dominance of English is not permanently established. On the Internet, more and more people publish in their own language. A recent estimate puts as high as 40% the number of Web pages which may now be in languages other than English (Unesco Observatory on the Information Society 1998). By 2005 at the latest, six Internet users out of ten will not be native English speakers (*Computer Economics* 1999). The Internet Society and many researchers are busy trying to figure out ways of reducing the impact of the language barrier on information flow. Oudet believes that 'if the Internet does not allow multilingual conversations (...), mistakes and misunderstandings will become rampant, and many users will be cut out of the tremendous opportunities that international communication has to offer' (1997). This sounds like a modern echo of Wellisch, who was already writing, 26 years ago:

> the users in developing countries might even come to feel that they are deliberately excluded from the storehouses of the world's knowledge by being presented with information in a language which is seemingly simple and universal yet does not allow for full exploitation of the sources because its hidden difficulties and shifting vocabulary constitute serious obstacles to the formulation of queries and to information seeking (1973, 159–60).

On the World Wide Web itself, the presence of sites with such evocative names as *Babel* (http://www.isoc.org:8080/index.html), *Multilingual demo — Any language on any computer anywhere in the world!!!* (http://www.v-serve.com/mlit97/showcase/htmls/links.html), and *Multilingual Information Society* (http://salt.essex.ac.uk/salt/general/europe/docs/mlis-0.html) reflects the amount of current interest for the issue. New terms have been coined: cross-language information retrieval (CLIR) and trans-lingual information access are now key expressions in the specialized discourse of information organization and retrieval.

The main objective pursued by proponents of multilingual information systems has always been to ensure that all who needed to inform themselves were able to do so, whatever their skills at foreign languages. To attain this objective, they have designed, and at least partially implemented over time, three types of multilingual systems, all of which involve of necessity more or less complex linguistic transformations:

1. Systems that speak a single language but appear to understand many. Such systems provide multilingual access to what is essentially a monolingual resource base. In the world of print indexes, the equivalent would be the provision of cross-references in a second (third, etc.) language to access an index created in one language only. This model, unconcerned with issues of language equality, remains the most inexpensive of all.

2. Systems that truly speak and understand a few preferential languages. Such systems provide multilingual access to a multilingual resource base, one in which documents have been indexed, manually or automatically, in all the languages of the system. Many European print indexes and online databases, for example, are accessible in most or all of the official languages of the European Community.

3. Systems that truly speak and understand a large number of natural languages. Conceived as refined extensions of the previous types, such systems rely greatly on automatic processing and on external language tools to provide access in any language to source documents in the same or in any other natural languages. These systems have yet to be fully implemented successfully.

Language-related problems

Not surprisingly, the standard protocols and software on which the Internet is based are beginning to show their limitations in terms of the number of user languages they can efficiently process. Here, as in more traditional information transfer environments, four categories of language-related problems affect both processing and outcomes. They are: encoding problems, morphological problems, lexical/terminological problems, and conceptual problems.

The first category of problems has much to do with script and with character recognition. Much of the software commonly used today is still not handling diacritics very well, and cannot even begin to process non-Latin alphabets. A potential solution to the encoding/recognition problem resides in the wide implementation of Unicode, a coded character set that currently assigns unique numbers to 39,000 characters used in various natural languages (for more information on Unicode, see http://www.unicode.org/).

The second category of problems, the morphological problems, stems from the fact that most words and terms may appear in more than one form in any flexional language. Some of the variations are simply due to authorized multiple spellings for the same verbal representation (e.g. Behaviour/Behavior, Clé/Clef). In the Western world, multiple spellings are frequent occurrences in information bases which have to be transliterated; many romanization schemes exist, and no two authors seem to be using them consistently. Other morphological variations are grammatical in nature, the various forms of a word indicating its number, gender, and in some languages, its function in the discourse (e.g. Student/Students, Étudiant/Étudiante/Étudiants/Étudiantes, Knabe (nom.)/Knaben (acc.)). Morphological diversity makes it difficult to achieve high recall in retrieval.

Lexical/terminological problems relate to the vocabulary itself that is used to produce and to index documents, and to search for information. A word is the verbal representation of a concept (an object existing either in the physical or in the abstract world). The same concept, however, is not always represented by the same word. Within the same natural language, different verbal representations will clearly mark national, regional, or local use (e.g. in general language, Lift/Elevator; in the language of a specialty, Cancer/Carcinoma). Across languages, the variations are as obvious as they are inevitable (e.g. Grandmother/Grand-mère/Abuela). Lexical variations will also affect recall in retrieval.

There is a progression in the impact of the first three categories of problems on access to information. The technical encoding problems can and will likely be solved in the near future, thus providing software with the means to identify any language, and making possible the implementation of multilingual information systems able to process a large number of natural languages. The reluctance to use Unicode will gradually fall away as 'the advantages of global language interoperability are found to far outweigh the trade-off in heavier storage requirements and the potential effect on response times' (Peters and Pichi 1997). The morphological problems will be solved with much research and standardizing work, as well as with consistent and complete cross-referencing systems. The impact of lexical/terminological problems will be reduced with the help of controlled switching languages. But neutralizing the effects of the fourth category of problems, the conceptual problems, may prove more of a challenge.

Languages are much more than lists of words and sets of rules to combine them. Languages are above all organized conceptual and lexical structures which reflect the way their speakers see and interact with the realities of the world. It is simplistic to believe that everything in the world can be organized in categories or classes distinct from one another, recognized in every culture, and adopted as a basis for each language. In every language, a word or term covers a certain area of the conceptual world, and this area can and will vary slightly from one language to another. When individuals describe the world in different languages, they are not describing exactly the same world and, in the best of cases, there still does not exist an exact coincidence between two languages in their ways of defining concepts and terms, of categorizing and relating concepts, of characterizing abstract entities, etc. (Maniez, 1997). This obviously makes it difficult to translate one natural language into another, a situation with which translators, terminologists, indexers, and multilingual thesaurus designers are all too familiar.

The passage from one language to another may lead to much ambiguity when a term available in a source language has more than one cross-lingual equivalent, representing different concepts, in the target language (e.g. the French *beau-père* is equivalent to both ‘stepfather’ and ‘father-in-law’ in English). Conversely, one term may exist in one language for which no equivalent is found in another, presumably because the concept represented in the source language does not exist in the target culture. Finding appropriate equivalents may be especially difficult if concepts do not have a stable lexical support; this is often the case in the special languages of the social sciences.

Conceptual problems will evidently hinder retrieval if it proves difficult or even impossible to identify accurate cross-language equivalents, thus preventing the extension of a search into more than one database or into databases including documents in several languages.

Potential solutions

Throughout the centuries, many different solutions have been used to overcome the language barrier, with varying degrees of success. The solutions have reflected the diverse requirements of important fields of human endeavour (commerce, literature, science, etc.), but the foremost solution has always been translation in oral and written form. This critical need to translate has been beneficial to the study of natural languages and of the linguistic phenomenon more generally.

With the recognition of the potential negative impact of the language barrier in expanding networks, these well-proven solutions are now being applied to facilitate the transfer of information. In a slightly modified form, often a more automated one, the proposed solutions are pretty much the same today as they were years and years ago:

1. Increasing the individual linguistic competence of information searchers and of information intermediaries. This humanistic solution has severe limitations: not everybody has the time and the inclination to learn foreign languages; besides, how many languages can one master in a lifetime? And the intermediaries are not as numerous today as they used to be!
2. Improving the quality and the amount of meaningful content in secondary information — information about information. In scientific journals, it is now common to add informative abstracts in several languages. In the world of the Internet, however, there is often no secondary information available, aside from the few index terms or category labels which have been used to retrieve the primary information in the first place.
3. Improving the quality and increasing the availability of language tools such as terminological banks, multilingual glossaries and thesauri.
4. Improving the quality and increasing the availability of machine-assisted translation services, the only viable solution for the Internet, according to Oudet (1997). A refinement of machine translation procedures is still needed, of course, since automatic translation currently operates at about 75%–85% intelligibility (Kay 1996). At this time, machine translation might be useful for translating access points (titles, subject descriptors, etc.), but it is not yet truly efficient for translating contents.

All of the above solutions are useful in providing access to documents in a multilingual environment, but only solutions 1 and 4 have the potential to be fully efficient in providing access to contents.

All four solutions have been adopted by the European Community in an ambitious programme aimed at facilitating information exchange among its members. Within a well-defined framework and with the support of evolving policies, Europe is proposing to: 1) better coordinate the creation, validation, and distribution of language resources and tools; 2) strengthen the language engineering industry and support the development of marketable products for the computer-based handling of languages; 3) develop an advanced translation industry; and 4) create an educated and advanced user community (European Commission 1995).

The vocabulary solution: a future for controlled indexing and searching vocabularies?

Much faith is being put in a combination of controlled vocabularies and machine-assisted translation to facilitate access to information in multilingual contexts. Among the controlled language tools being tested with some measure of success are lexical banks (multilingual dictionaries of the general language), termbanks (multilingual inventories of languages of specialties, with definitions and equivalents within and across languages), and multilingual thesauri (structured sets of terms used to index and retrieve documents in several languages in a specific domain of knowledge). The use of switching languages, which allow for the translation of an index term from one scheme into another via a neutral code system in which concepts are represented, is also explored; library classification schemes, such as the Dewey Decimal Classification, might be appropriate in this role (Mitchell 1997).

Truly multilingual tools are respectful of the essential equality of all natural languages when it comes to representing concepts, independent of the extent of their use in the world. Most information users are aware of the very real problems traditionally associated with the use of multilingual controlled vocabularies for information transfer: 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. To avoid such problems, it is recommended that multilingual tools be built from the ground up, starting with distinct banks of terms (one for each language represented) and developing distinct structures through semantic relations; these structures must be faithful reflections of the way speakers of a language see and represent their own world.

Indexing in a multilingual environment

Indexers have a major role to play in information transfer, not only by the work they do, but also by their beliefs about the importance of information, of quality access to information, and of truly equal access to information. Along with other information specialists, indexers should be acutely sensitive to language issues.

In the world of print, too many publishers still have to be convinced of the necessity of dual indexing for bilingual documents, for example, which would force them to recognize that not only costs, but also quality of access have to be taken into account in an information dissemination project such as publishing. Indexers must themselves recognize that texts in different languages, even if they are translations of one another, deserve full analysis and processing so that the integrity of all contents can be respected. Indexers must believe in and promote the bilingual/multilingual index as one which gives equality of treatment and recognition to all languages represented.

In back-of-the-book indexing, the issues related to multilingualism might not appear as critical. Indexers, however, should remain aware that people who will consult their index might not all have the same linguistic background, and that even within the same language, barriers do exist. Completeness of representation (i.e. making sure that the many different verbal representations of a concept do appear in the index), and consistency of expression (i.e. making sure that terms used to represent concepts are always the same) will be beneficial to all information seekers.

Indexers will be helped in their task by a growing number of language tools that will supplement what terminology is found in the document itself. A large number of these tools are already available on the Web, where can be found, next to traditional dictionaries and thesauri, *Just cows — The word cow translated into many languages* (http://www.arrakis.es/~eledu/justcows.htm), and *Silent night, holy night — Information about the Christmas carol, including the song translated into several languages* (http://silentnight.web.za/).

As information workers, let's make it a regular practice to look at multilingual tools when searching for the correct word, for synonyms, for related terms. It can only help to see how others organize the world and name its parts. We can then make sure that they do find what they are looking for when they search our information systems in a way that seems to them logical and natural. And when we search for information that they have produced and organized, we might stand a better chance of finding interesting and useful sources!

References

Computer Economics (1999). English will dominate Web for only three more years. *Computer Economics*, 9 June. Available: http://www.computereconomics/com/new4/pr/pr990610.html

European Commission (1995). A multilingual information society for Europe: the need for complementary action. In *The multilingual information society: a communication from the Commission*. Available: http://salt.essex.ac.uk/salt/general/europe/docs/mlis-5.html

Iljon, Ariane (1978). Compiling multilingual thesauri for sectors connected with agriculture: a European Community experience. In *Congrès régional européen des bibliothécaires et documentalistes agricoles*, 129–41. München: Saur.

Kay, Martin (1996). Machine translation: the disappointing past and present. In *Survey of the state of the art in human language technology*. Available: http://www.cse.ogi.edu/CSLU/HLTsurvey/HLTsurvey.html

Kertesz, F. (1977). Differences between the European and the American approach to overcoming the language barrier in science. In *Third European congress on information systems and networks: overcoming the language barrier...,* 271–91. München: Dokumentation.

Lancaster, Frederick.W. and Smith, Linda C. (1983). *Compatibility issues affecting information systems and services*. Paris: UNESCO.

Maniez, Jacques (1997). Fusion de banques de données documentaires et compatibilité des langages d'indexation. *Documentaliste — Sciences de l'information*, **34**(4–5): 212–22.

Mitchell, Joan (1997). Classification as a multilingual access tool. Paper presented at *Information and restructuring for democracy*, Warsaw, Poland, 5–7 November 1997. [unpublished]

Oard, Douglas W. (1997). Serving users in many languages: cross language information retrieval for digital libraries. *D-Lib Magazine: the magazine of digital library research*, December. Available: http://www.dlib.org/dlib/december97/oard/12oard.html

Oudet, Bruno (1997). Multilingualism on the Internet. *Scientific American*, March 1997. Available: http://www.sciam.com/0397issue/0397oudet.html

Peters, Carol, and Picchi, Eugenio (1997). Across languages, across cultures: issues in multilinguality and digital libraries. *D-Lib Magazine: the magazine of digital library research*, May. Available: http://www.dlib.org/dlib/may97/peters/05peters.html

Unesco Observatory on the Information Society (1998). *Multilingualism.* Available: http://www.unesco.org/webworld/observatory/themes/multilingual/home.html

Wellisch, Hans (1973). Linguistic and semantic problems in the use of English language information services in non-English speaking countries, or How to install an elevator in the tower of Babel. *International Library Review* **2**: 147–62.

Wellisch, Hans (1978). Early multilingual and multiscript indexes in herbals. *The Indexer* **11**(2): 81–102.

Yasmin, Nuzhat (1977). Language barrier in dissemination of scientific information. *Pakistan Library Bulletin* **8**(1): 37–46.

Michèle Hudon is Assistant professor at the École de bibliothéconomie et des sciences de l'information, Université de Montréal. Her main area of interest, research and expertise is multilingual and multicultural thesaurus design and use. She is a former president of the Indexing and Abstracting Society of Canada. Email: michele.hudon@umontreal.ca.