Abstract: The existence of huge amounts of information available in information systems and networks worldwide imposes the creation of adequate tools able to efficiently organize it and allow its retrieval across geographical, linguistic and cultural boundaries. An indexing language covering all areas of knowledge and converting the language-independent structure of a classification system like the Universal Decimal Classification into a thesaurus structure in more than one language seems to be a solution. Among the key attributes of the indexing language thus obtained we can mention: consistency in indexing, control on terms, user-friendliness. The paper presents the great potential in information retrieval of the combined retrieval method by means of a case study.

1. Introduction

Among the consequences of the rapid development of the global information society a major one is the existence of huge amounts of information stored in information systems and networks across geographical, linguistic and cultural boundaries. The need was imposed to create tools and technologies able to efficiently organize and allow retrieval of information in this universal context. Information professionals had to cope not only with the multitude of knowledge organisation and representation systems but also with the multitude of languages the available information is stored in order to provide the users with effective information retrieval tools.

For this purpose a real language industry has been developed, theoreticians and researchers making considerable efforts to find feasible solutions to problems of multilingual access by way of natural language processing and machine translation methodologies. Such corporate efforts belong to the CoBRA+ working group for multilingual access to subjects (MACS) or to the cross-language information retrieval (CLIR) tracks of the Text Retrieval Conferences that annually report the progress made in multilingual information access and retrieval. The encouraging results they have obtained so far are still confined to discipline/domain restrictions and most of their achievements are based on language pairs rather than multiple language combinations.

The multilingual access facilities offered by combining two different types of information languages in order to improve information retrieval make the substance of this report. The approach going to be described has two component parts:

- mapping a traditional classification system onto an interdisciplinary multilingual thesaurus
- making information retrieval possible by means of either UDC codes or UDC-based descriptors in any of the languages included in the project.
2. The Universal Decimal Classification and its potential to switch between languages

Along with the remarkable efforts made in the 1970's to prove the benefits of using the UDC as an intermediate or switching language, some other efforts were concentrated in the direction of building alphabetical indexes to the UDC tables in order to facilitate browsing and searching. An example of such an index is one Lorphevre (1973) made to the French International Medium Edition of the UDC. According to what the author argues in the introduction, this index includes all the expressions in the tables taken in alphabetical order. Yet, he goes on, the absence of some words do not imply that there is no way to classify such subjects but they can be expressed by combinations of numbers according to the UDC grammar rules.

Example:

<table>
<thead>
<tr>
<th>Développement des bibliothèques</th>
<th>021</th>
</tr>
</thead>
<tbody>
<tr>
<td>des plantes</td>
<td>581.14</td>
</tr>
<tr>
<td>équipement photographique</td>
<td>771.4</td>
</tr>
<tr>
<td>ontogénétique</td>
<td>577.95</td>
</tr>
<tr>
<td>organique, physiologie</td>
<td>612.64</td>
</tr>
<tr>
<td>organisation générale</td>
<td>65.016</td>
</tr>
<tr>
<td>photographie</td>
<td>77.023</td>
</tr>
<tr>
<td>urbanisme</td>
<td>711.12</td>
</tr>
<tr>
<td>zoophysiologie</td>
<td>591.16</td>
</tr>
</tbody>
</table>

The comparison Lorphevre makes in the introduction between his index and a thesaurus (or "trésor", as he calls it) is what we are concerned with. The first difference is a rather formal one, of succession of the two sections of the indexing tools, i.e. the classifications start with the systematic tables going from general to particular and they are completed by an alphabetical table, while the thesauri go the other way around: the alphabetical part comes first and it is followed by the systematic part. Another difference is relating to the way the coordination of terms is made: in the case of the UDC precoordination is the intrinsic feature of its own structure and additionally there are also "see" and "see also" references provided for this purpose, whereas in the thesaurus the coordination is indicated by the codes used in the alphabetical part. In the end the author announces his intention to transform the alphabetical indexes of the UDC into "trésors", structuring them differently in the first place, and then coordinating the terms. This way there will be no difference, he says, between the two methods, the last word belonging to the UDC "qui se joue des langues qui traduisent ses notations chiffrees".

Once again here, the advantage of numeric systems over alphabetical systems is underlined. Without saying it straightforwardly, the main advantage lies in the lack of ambiguity of the numeric notations as is the case with the problematic use of the natural language words that translate them.

3. Description of the case study

The great potential in information retrieval of the combined retrieval method will be demonstrated by a case study. An overview of the main issues covered by our project is following:
3.1. Structure of the database

A database has been built out of bibliographic records taken from a real-life online catalogue. The multilingual document collection thus constituted has subjects represented initially by UDC notations in all the bibliographic records and manually assigned descriptors in Romanian in some of them. To these bibliographic records the UDC Master Reference File (MRF) and a shortened version of this containing only the upper subdivisions of the UDC classes were added. The latter section of the database is meant to provide context to the captions. All in all, the database has 159553 records. A number of 97922 are bibliographic records with UDC notations, of which 40627 have manually assigned descriptors and 84110 have UDC-derived ones (see Table 1).

<table>
<thead>
<tr>
<th>Number of records in the bibliographic database</th>
<th>Number</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>97922</td>
<td>61.37%</td>
</tr>
<tr>
<td>Number of records in the Master Reference File (MRF)</td>
<td>61457</td>
<td>38.52%</td>
</tr>
<tr>
<td>Number of records in the short MRF</td>
<td>174</td>
<td>0.11%</td>
</tr>
<tr>
<td>Total number of records</td>
<td>159553</td>
<td>100%</td>
</tr>
<tr>
<td>Number of bibliographic records indexed with manually assigned descriptors</td>
<td>40627</td>
<td>25.46%</td>
</tr>
<tr>
<td>Number of bibliographic records indexed with automatically assigned multilingual descriptors</td>
<td>84110</td>
<td>52.72%</td>
</tr>
</tbody>
</table>

Table 1. Configuration of the experimental database

To serve our purposes of integrating a multilingual thesaurus into a bibliographic database in order to facilitate multilingual access, an indexing tool was created from a pre-established list of descriptors based on an abridged version of the UDC. The list of 1254 Romanian descriptors was developed into a trilingual thesaurus in Romanian, English and French that eventually had well over 3,000 terms (descriptors and non-descriptors) in each contributing language and a total of 4827 terms.

3.2. Automatic assignment of multilingual descriptors

Several programs have been created to perform each of the following activities in the experimental database such as:

- integrate the multilingual thesaurus terms in the MRF by matching the UDC-derived descriptors with the UDC numbers in the file;
- merge the resulting product with the bibliographic database to which successive other programs were previously applied in order to decompose the UDC notations given in the classified catalogue into their separate parts and add the corresponding text to them.

The bibliographic record in Figure 1 illustrates the additional access points to the subject:
MFN: 089574
Descriptors: Bibliologie.
    Istoria cartii.
    Activitate editoriala.
UDC-descriptors: 703: Great Britain – Grande Bretagne – Marea Britanie
    709: Graphic industries – Industries graphiques – Industrii poligrafice
    709: Documentation – Documentation – Documentare
UDC subject notation: 002
655(410)
Aux. for place: ^a(410)^eGreat Britain. United Kingdom of Great Britain and Northern Ireland
   Authorship
   ^xProlegomena. Fundamentals of knowledge and...

Figure 1. Bibliographic record with additional access points to the subject of the document

The traditional search method of using the precoordinated UDC numbers as queries is enriched by additional search possibilities represented by two types of descriptors in our experimental database namely: monolingual Romanian descriptors assigned by the indexers at the moment of subject indexing and multilingual descriptors in English, French and Romanian assigned automatically. Likewise, the UDC numbers in the classified catalogue decomposed into their component parts and particularly their textual meaning can be successfully used in searching.

3.3 Consequences of the automatic indexing on information retrieval

The search we present as an example used queries in Romanian and in French and it was first conducted by using automatically assigned multilingual descriptors in a) and b) and then by manually assigned Romanian ones existing in the bibliographic database in c). A comparison between these search results will show the higher recall rate of automatically assigned multilingual descriptors used as queries as against the manually assigned ones [see below the responses to the search statement 'Cardiovascular diseases': 101 hits for 1 query statement vs. 80 hits for 6 query statements].

The results are quite promising in as much as information retrieval is concerned if we compare the search results for a given topic using each of the two search possibilities in turns, i.e. the manually assigned monolingual descriptors and the automatically assigned multilingual ones.
Consider another example. It is often difficult to pinpoint exactly the right term for a particular concept. Let us consider the following UDC numbers in the MRF and see what the consequences are when they are used in indexing and mapped (manually) onto descriptors.

<table>
<thead>
<tr>
<th>UDC</th>
<th>Descriptors</th>
</tr>
</thead>
<tbody>
<tr>
<td>159.923.3</td>
<td>Composition of the personality. Character traits. Psychogram</td>
</tr>
<tr>
<td>159.925</td>
<td>Study of expression. Physical expression of character</td>
</tr>
</tbody>
</table>

For the first UDC number there are 3 different descriptors used in the bibliographic records in the database: Tipologie (Psihologie), Psihologie individuale and Caracter (Psihologie).

The second UDC number is represented in the bibliographic records by descriptors such as: Personalitate (Psihologie), Caracter (Psihologie), Characterologie, Tipologie (Psihologie).

The third UDC number has as textual correspondents the following terms: Psihologie individuala, Fizionomie, Caracter (Psihologie), Morphopsihologie - here it was the title of the document that inspired the use of this descriptor: ABC de la morphopsychologie : [connaître sa personnalité par les traits du visage] / Carleen Binet.

Needless to say that the diversity of manually assigned descriptors can only produce confusion because of total inconsistency in indexing but the final result is loss of information having as source information scattering throughout the catalogue. The necessity of control on terms is imperious here and one way to
provide it is to make cross references between descriptors based on the principle of likeness. A much 'safer' way to ensure consistency in indexing is to map unique descriptors onto UDC notations. This way the control on terms is (automatically) imposed and this is just what our approach is about.

One of the requirements of the above-described approach is that there has to be a close relationship between the specificity of the classified catalogue of the given database and the specificity of the UDC-based thesaurus for better evaluating the system. If this request is ignored then we have a problem of compatibility between the classified catalogue and the selection of UDC numbers the thesaurus is based on. The direct consequence of this situation is information loss as in the following example. The main topic represented in the existing UDC notation 616.281-089.843 is too detailed compared with the general level of specificity of the UDC-based thesaurus. However, in a system like this there are ways to avoid such shortcomings given the alternative IR methods which can be used. One of those is the textual counterpart of the UDC notation itself (fields 702-709 in Figure 2).

Looking back at the table presenting the structure of the experimental database we can easily notice the difference between the number of records in the bibliographic database (97922) and the number of bibliographic records indexed with automatically assigned multilingual descriptors (84110). Presumably all bibliographic records are indexed with UDC numbers. The automatic indexing resulted in a lower number of records indexed with multilingual descriptors. The reason has to do with the difference in specificity between the two indexing languages i.e. the classification notations found in the database and the selection of classification numbers used as a base for the multilingual thesaurus.

4. Advantages of a multilingual UDC-based thesaurus

These examples are meant to argue the following:

- Being given a bibliographic database with subjects indexed by UDC numbers and descriptors assigned manually according to the meanings of those UDC numbers, the automatically assigned descriptors derived from a UDC-based thesaurus turn out to give better results in information retrieval. Therefore the improved retrievability is the main advantage of our approach.

- The multilingual character of the UDC-based thesaurus implies automatically that it addresses a broader range of users according to the number of languages involved.

- The manually assigned descriptors are not so reliable as the automatically assigned ones given the inconsistencies and lack of control likely to occur when more than one descriptor is linked with the same UDC number; the UDC-based descriptors will always be the same for a particular UDC number. It is the number of lead-in terms in each of the contributing languages that provides lexical richness to the indexing language anticipating the users' formulations therefore enhancing predictability.
The user-friendliness of this indexing tool is beyond any doubt. The accurate use of classification notations combined with the necessary respect of the UDC grammar rules in the classifying process are the basic requirements for an optimal result in the automatic conversion of the UDC notations and the assignment of the descriptors from the multilingual UDC-based thesaurus. Such a way the highly preferable natural language words replace the inconvenient numeric codes in information retrieval.

The automatic assignment of UDC-based descriptors in the bibliographic records has an extra advantage derived from the user-friendliness of this indexing method. If numbers do not say much when it comes to appreciating the correctness of a given UDC notation, a descriptor that is automatically converting that notation will immediately struck the eye if not correctly assigned. An apparently minor typing mistake can tremendously distort the subject of a document affecting thus the recall rate. Example: A switch
between two digits in the UDC notation: 239.18 Dogmatic theology instead of 329.18 Fascist attitude for a document having the title “The last days of Hitler” may at least generate confusion in the searcher’s mind. Such an error is much easier to be identified in natural language words than in classification codes. Under such circumstances the keeping up of the database will demand less efforts particularly if global change procedures are available for that purpose.

Our approach illustrates how traditional resources can be integrated into newly-designed models of knowledge organizers by mapping terms in several languages onto the same concept expressed in a language-independent structure. The UDC in our case plays a creative role in information retrieval participating in new strategies for improved access to information in multilingual environment.

Reference