The impact of specificity on the retrieval power of a UDC-based multilingual thesaurus

The Impact of Specificity on the Retrieval Power of a UDC-Based Multilingual Thesaurus

Victoria Frâncu

SUMMARY. The article describes the research done over a bibliographic database in order to show the impact the specificity of the knowledge organising tools may have on information retrieval. For this purpose two multilingual UDC-based thesauri having different degrees of specificity are considered. Issues of harmonising a classificatory structure with a thesaurus structure are introduced and significant aspects of information retrieval in a multilingual environment are argued in an extensive manner. Aspects of complementarity are discussed with particular emphasis on the real impact produced by alternative search facilities on IR. Finally a number of conclusions are formulated as they arise from the study.

KEYWORDS. Information languages; intermediate languages; UDC; compatibility of information languages; multilingual thesauri; information retrieval

Victoria Frâncu is a librarian at the Central University Library of Bucharest (Romania) completing her PhD study in library and information science at the University of Antwerpen (Belgium). (E-mail: francu@bcub.ro).

The author wishes to acknowledge Dr. Gerhard J. A. Riesthuis from the University of Amsterdam for his assistance in preparing the database to make automated indexing possible.

INTRODUCTION

The challenge of easier access to information contained in bibliographic databases in which the subjects of documents are represented by unfamiliar (to average users) classification codes – in our case the UDC notations – makes the key issue of our study. Our purpose is to demonstrate that multilingual access to information contained in bibliographic databases is possible via multilingual descriptors mapped onto the UDC numbers. This can be automatically done while subjects of the documents are indexed with UDC notations alone.

More specifically, this study is intended to discuss problems of compatibility and integration of information languages and to emphasise major aspects of multilingual access to information through an intermediate language. In our case this position is undertaken by the Universal Decimal Classification. Compatibility of information languages being important to our study, particular importance will be given to aspects of full and partial compatibility along with a few problems regarding the complementarity of the indexing languages involved.

A multilingual thesaurus based on the UDC Pocket Edition (1999), built according to the existing international standards – ISO 2788 (1986) and ISO 5964 (1985), will be compared with a smaller one in size, characterised by some particular features generated by its wide coverage on the one hand and its lack of specificity on the other. Going from a lower level of specificity to a higher one affects the building process of the indexing language. At the same time, the use of this thesaurus, wider in coverage than the previous one, has some impact on information retrieval.

CONFIGURATION OF THE EXPERIMENTAL DATABASE AND OBJECTIVES OF THE RESEARCH

In order to accomplish our objectives an experimental database was built from a number of bibliographic records in a real-life classified catalogue. To this bibliographic database four other databases were added so that in the end the experimental database as such has five sections according to our requirements: 1) Bibliographic records – 97918; 2) Records in the Master Reference File (MRF) – 61457; 3) Records in the short MRF – 174; 4) Records in the abridged (little) UDC-bases thesaurus (LTHES) – 1254; 5) Records in the Pocket Edition UDC-based thesaurus (PTHES) – 2033.

Our objectives are the following:

1. The logical structure of the UDC can be successfully converted into a thesaurus structure;
2. There are difficulties in building the UDC-based multilingual thesauri connected with the relational structure given the classification system is discipline-oriented and in many parts enumerative; likewise, hierarchies are not always present (for example Class 2 Religion, before revision);
3. Once the thesaurus is implemented in the bibliographic database the use of UDC-based thesaurus terms give better search results than the manually assigned descriptors;
4. There are commonalties and particularities in using the two different thesauri in searching according to their specificity level;
5. What the UDC can and the thesauri cannot, or the importance of keeping separate fields for unique entities;
6. The alternative search methods are complementary to each other;
7. Recall and precision rates and the way they are granted in keeping with the specificity of the search method used in IR.

METHODOLOGY

The experimental database described in the foregoing has in its bibliographic section documents indexed with UDC class numbers in all records, part of them having also manually assigned Romanian descriptors given by the indexers at the time of indexing (Frâncu, 2002). Given the application of several successive programs to the UDC notations in order to decompose them into their separate parts and add the corresponding captions to them, the text of the UDC MRF for each UDC notation is also found in the bibliographic database (Riesthuis, 1999). With these pre-existing elements we started our approach according to the following methodology:

- A UDC-based multilingual thesaurus based on an abridged version of the UDC was built according to the existing international standards: ISO 2788 (1986) and ISO 5964 (1985). The selection of UDC numbers became a list of 1254 Romanian descriptors that was further developed into a trilingual thesaurus in Romanian, English and French. The resulting thesaurus eventually had about 3,000 terms (descriptors and non-descriptors) in each contributing language and a total of 4827 terms (Frâncu, 2002: 406). This thesaurus has some characteristics given its wide coverage on the one hand and its lack of specificity on the other.
- Apart from this, a multilingual thesaurus in the same three languages was built, this time based on the UDC Pocket Edition. Not all the classes of the UDC are covered by the newly built thesaurus. This thesaurus has a total number of 7554 terms of which 2033 are descriptors, the ratio between preferred and non-preferred terms amounting at 3.7. The number of terms in each of the contributing languages is 4033 for English, 3735 for French and 3849 for Romanian. The thesaurus covers all the auxiliary tables, Classes 0, 1, 2, 3, 61 and 8 of the UDC Pocket Edition. Therefore a more specific information retrieval tool was created to be used in parallel with the initial one and, by comparison, enable a better perception on the difference specificity makes in information retrieval. Aspects of information retrieval are going to be examined comparatively for documents in the database

for which multilingual descriptors in both thesauri have been automatically assigned by programs specially created.

- Since the bibliographic records in the database had the subjects of documents initially represented by UDC notations and by subsequent processing also by the textual counterpart of these, the ultimate goal of our research is to emphasise the difference in search results generated by the difference in specificity of the two multilingual UDC-based thesauri and additionally compare the three ways of accessing the information: UDC notations and their textual counterpart, the automatically assigned multilingual descriptors, and the manually assigned descriptors given in by the indexers.

The first two in our list of objectives having been extensively dealt with in previous articles (Frâncu, 1996, 1999, 2000), in the ongoing we try to explore our goals beginning with the 3rd: the use of UDC-based thesaurus terms in information retrieval and the search results of those compared with the manually assigned descriptors.

AUTOMATICALLY VS. MANUALLY ASSIGNED DESCRIPTORS

In order to test the retrieval power of the multilingual descriptors in the experimental database we made several searches and compared the results. We give here a few examples of these searches.

Search query No. 1: 3 different UDC numbers of which the second is a subdivision of the first: 159.923, 159.923.3, 159.925.

Along with the UDC numbers, as previously said, the subject matter of documents in the database is also represented by Romanian subject headings assigned to bibliographic records at the moment of indexing. Since our intention is to illustrate the overlaps and inconsistencies likely to occur in case of manual indexing, we give below the correspondence between the captions of the selected UDC numbers and the Romanian descriptors as far as they were used to represent subjects relating to their meanings (Table 1).

UDC numbers	UDC Captions	Manually assigned descriptors
159.923	Type psychology. Individual psychology. Psychology of individualities. Individuality. Personality. Character psychology. Characterology. Idiosyncrasies. Personal equation. Personality types	*Tipologie (Psihologie)* *Psihologie individuala* *Caracter (Psihologie)*
159.923.3	Composition of the personality. Character traits. Psychogram	*Tipologie (Psihologie)* *Caracter (Psihologie)* Caracterologie Personalitate (Psihologie) *Psihologie individuala*
159.925	Study of expression. Physical manifestation of mentality. Bodily expression of character	*Caracter (Psihologie)* *Morfopsihologie* *Fizionomie*

Table 1. Sample of UDC numbers and captions from the Master Reference File and corresponding descriptors

Discussion:

At first glance the descriptors have very much in common. If only we look at ‘Caracter (Psihologie)’ we see it mentioned in all three sets of Romanian descriptors. A closer look will reveal that also the texts of all three UDC numbers include the word ‘Character’ and furthermore, the second UDC number is a subdivision of the first. Therefore, the need was felt for the presence of other two descriptors as counterparts of the first two UDC numbers: ‘Tipologie (Psihologie)’ and ‘Psihologie individuala’. Additionally a differentiating element was introduced accounting for the higher level of specificity in the second case i.e. the term ‘Personalitate (Psihologie)’ that is the exact meaning of the 2nd UDC number. In such cases the consequences for information retrieval will be an increased recall rate but a lower precision rate. Many of the documents got in response to such queries will have other collateral subjects apart from the intended search formulation. As far as we see it the solution here is to have *one-to-one correspondence between the UDC numbers and the descriptors used in searching*. In other words, a UDC-based thesaurus will alleviate this kind of shortcomings. If such a thesaurus is imbedded in the bibliographic database as in our experiment, the descriptors will automatically be assigned to the bibliographic records in keeping with the UDC numbers formerly assigned. Each time a UDC number is ascribed to a bibliographic record, its corresponding descriptor and only that will be attached to it. Such a way consistency but also control on the indexing terms is ensured.

Search query No. 2: acoustics. We give below the results of different searches conducted in the database to illustrate the advantage of automatically assigned descriptors over the manual ones, starting from a descriptor in our multilingual thesaurus this time:

```
Set Hits    Query element              Current Data Base name = VFDAT
--  ------  -------------
 1   131    "MDE=ACOUSTICS"

 2   159    ?v709^e : 'acoustics'
 3   148    ?v709^e : 'acoustics' and v709^x : 'physics'
 4    10    ?v709^e : 'acoustics' and v709^x : 'music'
 5     2    ?v709^e : 'acoustics' and v709^x : 'linguistics'

 6    26    "BDE=ACUSTICA" + "BDE=ACUSTICA APLICATA" + "BDE=ACUSTICA
             ARHITECTURALA" + "BDE=ACUSTICA FIZIOLOGICA" + "BDE=ACUSTICA
             MEDICALA" + "BDE=ACUSTICA MUZICALA"
```

Discussion:

As previously argued, this example demonstrates the improved result in terms of precision and lack of ambiguities when automatically assigned descriptors are used as search elements. The number of retrieved records stand proof for our statements. And indeed, if we compare the number of hits resulted after the first search, using as query element the term ‘Acoustics’, an English descriptor from LTHES, the number of retrieved records is very close to those resulting from a key term from the UDC caption ‘Acoustics’. The next three are combined searches in which one of the elements is ‘Acoustics’ and the other represents the discipline acting as context for the initial term such as: ‘Physics’, ‘Music’ and ‘Linguistics’.

The relatively small difference between the UDC caption search mode (2nd query element) and UDC-based thesaurus term mode (1st query element) on one hand compared with the big difference between the former and the manual descriptor mode (6th query element) pleads in favour of our approach. In addition to that, the 3rd, 4th and 5th queries that differentiate instances in the use of our query statement compared with the use of various combinations in manual descriptors (the 6th query) point out in an even clearer manner that the recall rate is in the last case quite low. It is true though, that one of the objective reasons for this very low number of retrieved records is that the manually assigned descriptors have only been used in indexing from a certain moment on in the history of the catalogue.

COMMONALTIES AND PARTICULARITIES IN USING THE TWO THESAURI

One of the main distinctions between LTHES and PTHES is that the first was built from a pre-established list of selected UDC numbers whereas the second was based on the UDC Pocket Edition. For all that, many of the rules applied in building these thesauri are shared by both.

In mapping the thesaurus terms onto the classification structural framework the limitations were mostly imposed by the differences in specificity. A solution suggested in an earlier study to overcome this deficiency (Frâncu, 2000) is what was called '*upward posting*' (Aitchison & Gilchrist, 1987: 12) or '*generic posting*' (NISO, 1994) for an expanded category of terms as compared with what this method was initially established for. According to this device we treated narrower terms as sibling terms providing 'see' references to the preferred term. This does not prevent such terms from being access terms however, but they lose their quality as indexing terms. The control on the thesaurus terms is yet maintained and the recall rate is not affected by any means. Precision has a lower rate in such circumstance but, as we shall see further, this inconvenient can be overcome by the use of alternative search facilities that coexist in our experiment. Consider this example that illustrates the way we approached the upward posting (Frâncu, 2000):

ENGLISH
Field crops
UDC 633
UF Aromatic plants
UF Beverage plants
UF Cereals
UF Condiment plants
UF Edible roots and tubers
UF Forage grasses
UF Industrial plants
UF Leguminosae
UF Medicinal plants
UF Oleaginous plants
UF Plants yielding stimulants
UF Sugar plants
UF Tanning plants
UF Textile plants
BT Agriculture

FRENCH
Plantes de culture
CDU 633
EP Céréales
EP Leguminosae
EP Plantes à boissons
EP Plantes à tanin
EP Plantes aromatiques
EP Plantes fourragères
EP Plantes industrielles
EP Plantes médicinaux
EP Plantes oléagineuses
EP Plantes stimulantes
EP Plantes sucrières
EP Plantes textiles
EP Racines comestibles et tubercules
TG Agriculture

ROMANIAN
Plante de câmp
CZU 633
UF Cereale
UF Leguminosae
UF Plante aromatice
UF Plante de cultură
UF Plante de zahăr
UF Plante furajere
UF Plante industriale
UF Plante medicinale
UF Plante oleaginoase
UF Plante textile
UF Rădăcini comestibile şi tuberculi
TG Agricultură

Another of the differences between the two thesauri in discussion is that the first was created for practical purposes, having in mind the usefulness of the concepts, in other words their adequacy in indexing and information retrieval. This premise was not true for the second thesaurus that followed strictly the UDC structure it was derived from. While the first has a quite enumerative configuration, the other is both enumerative and faceted. This will have an influence on information retrieval that we shall examine later.

As previously said, all the auxiliary tables, as far as they exist in the UDC Pocket Edition, have their matching terms in PTHES. This was beneficial for Class 8 – Linguistics and Literature whose main class numbers are built by means of parallel subdivision using the Auxiliary table for languages (Table 1c). An instruction given in the database is meant to bring together the term ‘Literature’ (821 for individual literatures in UDC) and terms that correspond to the common auxiliaries of language (having this form: =…) to denote the literature of a particular language.

This does not apply in LTHES, where the list of most used literatures was already established at the moment the list of Romanian descriptors was made. Which brings about the inconvenient of having only a given number of literatures available for whom the automatic indexing is possible, the rest of them not being indexed at all. The alternative would be an excessively long list of languages and literatures, enumerating all those that are likely to occur in a library collection, but who can grant that?

WHAT THE UDC CAN AND THE UDC-BASED THESAURI CANNOT

The flexibility of the UDC permits the user to tailor or accommodate any given concept even though there is no class number provided for that. This can be done through *alphabetical additions* that, in theory, can accompany any class number. In practice, though, there are few numbers that have this indication in the UDC tables. For these numbers followed by always different alphabetical additions there is no possible way to issue thesaurus terms since they cannot be predicted. Furthermore, they cannot be controlled.

Another way the UDC denotes flexibility is that it allows for expressing the category of time by means of the *common auxiliaries of time*. A prescribed list of such auxiliaries of time is given in Table 1g but other mentions of this category (such as particular dates and ranges of time) are possible according to the UDC grammar. These auxiliaries of time, just like the alphabetical additions, cannot be predicted neither controlled. In terms of specificity this is very much increased by devices such as these. Both types of concepts discussed above are considered as so-called ‘unique entities’ (NISO, 1994) and a way to handle this category of information so that it can be retrievable is to keep it in separate fields. Field 706 in the example below is such an example.

COMPLEMENTARITY OF THE ALTERNATIVE SEARCH METHODS

Let us now explore more closely the search methods we only introduced in a few words so far. Consider this example in order to have a clear image of how a bibliographic record in the experimental database looks like:

```
+ 4 / 8   -------------------------------------------------- Format: VFDAT  -+
¦TIT:Grammatica albanese con le poesie rare di Variboba / Vincenzo Librandi. ¦
¦    - 2a ed. -  Milano, Ulrico Hoepli, 1928. -  XV,381p., 16cm. -           ¦
¦P-DES: ^1709!^eAlbanian language^fLangue albanaisee^rLimba albaneză         ¦
¦P-DES: ^1709!^eAlbanian literature^fLittérature albanaisee^rLiteratura      ¦
¦       albaneză                                                             ¦
¦P-DES: ^1709!^eGrammar^fGrammaire^rGramatică                                ¦
¦P-DES: ^1709!^eLiterary studies^fEtudes littéraires^rStudii literare        ¦
¦P-DES: ^1709!^ePoetry^fPoésie^rPoezie                                       ¦
¦UDC:  811.18`36                                                             ¦
¦UDC:  821.18.09-1                                                           ¦
¦UDC:  821.18.09Variboba, G.                                                 ¦
¦676:  811.18`36                                                             ¦
¦676:  821.18.09-1                                                           ¦
¦676:  821.18.09Variboba, G.                                                 ¦
¦706:  ^eVariboba, G.                                                        ¦
¦709:  ^a811.18`36^eAlbanian - Grammar^xLinguistics. Languages               ¦
¦709:  ^a821.18.09^eAlbanian - Literary criticism. Literary                  ¦
¦       studies^xLiterature                                                  ¦
¦709:  ^a821.18-1^eAlbanian - Poetry. Poems. Verse^xLiterature               ¦
¦MFN: 38009                                                                  ¦
+----------------------------------------------------------------------------+
```

The bibliographic record shows what we have discussed earlier: different kinds of access represented by *PTHES descriptors* (prefixed by P-DES), *UDC notations* (prefixed by UDC) and *700 type fields* (706 and 709) that illustrate the conversion of UDC notations and, if complex ones, of their component parts, into natural language words. There is one more thing to be said here, namely that the *alphabetical addition* (Variboba, G.), difficult to manage in a thesaurus as unique entity is placed in a separate field and thus made available to the user. We deal here with an example of complementary information adding value to the search methods used.

The missing piece is the representation of subject by way of LTHES. The absence of this search mode is caused by the low level of specificity that information retrieval tool has which did not allow for such relatively seldom met literature to be included among the descriptors existing in the list. This is one of the shortcomings of insufficient tuning between the specificity of the two information languages basically used: UDC notations and the UDC-based thesaurus terms. Manual descriptors are not found here either because the record is in the first half of the database.

The bibliographic record earlier shown belongs to the second set of retrieved records in which we used field 675 for the UDC notation as query element (see the display of the query formulations underneath). An identical search result was obtained as response to the third query that was initiated by using field 709 for the text of the UDC notation as query element. Such searches using fields that were designed for UDC notations are hardly supposed to be used by the non-experienced user of the database. The first search, though, uses one of the PTHES descriptors in English and the search result is rather close to the other two. ‘Albanian language’ that goes together with ‘Albanian literature’ as query element makes the difference in this case.

```
Set      Hits   Query element                Current Data Base name = VFDAT
---     ------  -------------
1         9   "PDE=ALBANIAN LANGUAGE" + "PDE=ALBANIAN LITERATURE"
2         8   "675=821.18(05)" + "675=821.18-1=111(082)" +
              "675=821.18-31=133.1" + "675=821.18-31=135.1" +
              "675=821.18-822=133.1" + "675=821.18.09" +
              "675=821.18.09-1" + "675=821.18.09<14/19>" +
              "675=821.18.09VARIBOBA, G."
3         8   "709=821.18" + "709=821.18-1" + "709=821.18-31" +
              "709=821.18-822" + "709=821.18.09"
```

Consider this example in which the subject is represented by terms in LTHES and PTHES:

```
+   3 / 37 ------------------------------------------------- Format: VFDAT --+
¦TIT:  Japanese financial markets, deficits, dilemmas, and deregulation /    ¦
¦      Robert Alan Feldman. -  Cambridge. Mass, The MIT Press, 1986. -       ¦
¦L-DES: ^1709:^eFinance^fFinances^rFinanțe                                   ¦
¦LNDES: ^1709;^eFinancial policy^fPolitique fiscale^rPolitică fiscală        ¦
¦P-DES: ^1703-^eJapan^fJapon^rJaponia                                        ¦
¦P-DES: ^1709-^eFinance^fFinances^rFinanțe                                   ¦
¦PNDES: ^1703>^eNippon^rNippon                                               ¦
¦UDC:  336(520)                                                              ¦
¦676:  336(520)                                                              ¦
¦703:  ^a(520)^eJapan. Nippon (Nihon Koku)                                   ¦
¦709:  ^a336^eFinance. Public finance. Banking. Money^xEconomics. Economic   ¦
¦       sciences                                                             ¦
¦MFN: 14403                                                                  ¦
+----------------------------------------------------------------------------+
```

The subject is represented by descriptors in each of the multilingual thesauri along with the UDC notation and its textual meaning. The difference between the terms of the two thesauri is that LTHES does not provide a descriptor for the country code, whereas PTHES does. Furthermore, we can notice here the presence of non-descriptors in both cases (prefixed by LNDES and PNDES).

Another round of searches are meant to compare the manually assigned descriptors with the automatically assigned ones in both variants:

```
Set       Hits    Query element            Current Data Base name = VFDAT
---      ------   -------------
1          36      "BDE=CARDIOLOGIE"
2         101      "MDE=CARDIOVASCULAR COMPLAINTS"
3         101      ? v709 : 'cardiovascular complaints'
4          60      "PDE=CARDIOVASCULAR DISEASES"
```

The most comprehensive results are given by the LTHES searches and by the 709 field for textual meaning of the UDC notation (101 hits). The highest recall rate does not give any guarantee about the precision of the retrieved records. An analysis of this set of search results will reveal that we have here bibliographic records classed not only with 616.1 for ‘Cardiovascular complaints’ but also with subdivisions of this: ranges with “/” in which this number is included, colon relations in which this number can be the first or the second member, plus multiple common auxiliaries and special auxiliaries accompanying the basic UDC number.

A lower result is given by the PTHES search (60 hits). Although restricted numerically, this search has the advantage of the highest precision rate.

The Romanian manual descriptor corresponding to our search topic was found in only 36 records but this number will not say much since Romanian descriptors were assigned to bibliographic records from a certain moment on and not from the beginning of the database (the first record that has manual descriptors – other than proper nouns like country names and personal names – is record no. 45695 out of a total of 97918). In addition to that, it is not certain that this descriptor was consistently assigned to all documents dealing with “cardiology” or “cardiovascular complaints” as counterpart of 616.1 as UDC number.

Summing up, the bibliographic records in the experimental database have subjects represented by different indexing tools with different coverage degrees, therefore different degrees of specificity. In the forthcoming part of the study we shall point out the impact the various information languages and particularly the terms from the two thesauri have on information retrieval in terms of recall and precision.

RECALL AND PRECISION RATES AND SPECIFICITY

Most definitions of thesauri underline one commonly agreed characteristic i.e. they are indexing languages that contain controlled vocabularies used in post-coordinated search of information. An equally important characteristic of thesauri is that terms within their structure are related to each other both hierarchically and associatively in a way that may be helpful for their users.

One of the most accurate definitions of a thesaurus is given by the MDA Archaeological Objects Thesaurus (1997): “A thesaurus is a tool which helps indexers and searchers to choose words consistently to describe things or concepts. The thesaurus is structured in such a way that related words are grouped together and cross-referenced to other groups of words which may be relevant to the subject. Where there is a choice of words with the same or similar meanings, the thesaurus provides a single preferred word and, by arranging terms in a hierarchy, allows the selection of more general or specific words. The purpose of the thesaurus is to standardise the use of terminology, which not only helps in indexing information but also in retrieval. Furthermore, it is a dynamic tool, one which can be developed through the addition or amendment of hierarchies, terms and relationships according to need.”

To what extent this complex definition applies to our study?

The experimental database has subjects of the documents in the bibliographic records represented by:

- one or more UDC notations in every one of them – field 675;
- text corresponding to the UDC notations or their broken up parts, if those are correctly assigned from the UDC grammar point of view and they are found in the UDC MRF included in the database – fields 701-709;
- one or more Romanian descriptors manually assigned, beginning with record no. 45695 out of a total of 97918 bibliographic records – field 610, access prefix BDE=;
- zero or more UDC-based multilingual descriptors automatically assigned from a thesaurus based on a selection of 1254 UDC numbers from the whole UDC schedule (LTHES) – field 620, access prefix MDE= for English, MDR= for French and MDR= for Romanian; additionally there are also non-descriptors accompanying them in field 625, prefixed by MNE=, MNF= and MNR= ;
- zero or more UDC-based multilingual descriptors automatically assigned from a thesaurus based on a selection of classes of the UDC having 2033 descriptors (PTHES) – field 630 access prefix PDE= for English, PDR= for French and PDR= for Romanian; these too have non-descriptors in field 635 prefixed by PNE=, PNF= and PNR=.

The two different thesauri imbedded in the database do not keep all the relations among their terms but only the non-descriptors which can be used as access points to the information in the bibliographic records. Therefore one of the two major characteristics of the thesauri mentioned above does not function here unless the printed versions of LTHES and PTHES go together with the database thus constituted and are used as search guiding tools by the users of the database.

Our findings on the experiments and tests that have been conducted over the database – of which some are shown above – are the following:

- the highest recall rate in all the searches conducted belongs to searches in the 700 fields for broken up UDC notations and their textual meaning;
- the highest precision rate is given by searches with PTHES provided that searches are conducted for classes that exist in PTHES; precision is estimated here by comparing the number of hits as result of searches made with manually assigned descriptors and automatically assigned descriptors in PTHES; compare these search results:

```
Set      Hits     Query element            Current Data Base name = VFDAT
---      ------   -------------
 1       8211    "PDE=ROMANIAN LITERATURE"
 2       2955    "BDE=LITERATURA ROMANA"
 3       8731    ? v709^e : 'Romanian' and v709^x : 'literature'
```

- given the broad coverage but lack of specificity, the use of LTHES in information retrieval can give "no hits" as search result if the UDC numbers in the bibliographic records are too detailed, but gather consistently large number of hits if descriptors in LTHES find the numbers they are based on in the bibliographic records; for example, a search request for 'psychiatry' will give no hits if the query is MDE='psychiatry', but also these results for query elements such as:

```
Set      Hits     Query element            Current Data Base name = VFDAT
---      ------   -------------
 1       102      "675=616.89"
 2       198      "PDE=PSYCHIATRY"
 3         0      "MDE=PSYCHIATRY"
 3       216      "MDE=NEUROLOGY"
 4        95      "BDE=psihiatrie"
```

But what is specificity, after all? What we mean by specificity in this study is the possibility to index the information under a specific subject heading not under the class that particular subject heading belongs to (see the 'neurology' and 'psychiatry' examples above). In his Rule 106, C. A. Cutter (1904) was arguing in favor of an entry for the subject heading and not for the class which included a subject. We too, plead for the possibility to be as precise in indexing as possible and as consistent as possible avoiding ambiguities and overlapping for an enhanced information retrieval.

CONCLUSIONS

This study shows that the information in a classified library catalogue containing multilingual bibliographic data can be accessed by natural language words existing in multilingual UDC-based thesauri. The recall and precision rates for such search tools varies according to the specificity level of the thesaurus terms used in IR. Ideally, the coverage and specificity of the IR tool should be as close as possible to the level of specificity of the classification notations assigned to the bibliographic records. For a database like the one used in this study the alternative search methods are meant to give improved search results providing in many cases ways out if 'no hit' responses are given to particular queries. Furthermore, unique entities like alphabetical additions and dates or ranges of time, permitted by the UDC grammar, can be evidenced in separate fields and consequently accessed if needed, in spite of the thesauri being incapable to handle or control them.

Eventually, the UDC notations in a classified catalogue can be and is the backbone for a variety of search methods based on its structure. Its language independence allow for switching between languages that translate the UDC codes into thesaurus terms in as many languages as needed or merely into words that represent their own meaning.

AUTHOR NOTE

The database used in this study contains bibliographic records from the catalogue of the Central University Library of Bucharest (Romania).

REFERENCES

1. Aitchison, J. & Gilchrist, A. (1987). Thesaurus Construction: a Practical Manual. 2nd ed. London: Aslib
2. Cutter, Charles A. (1904). *Rules for a Dictionary Catalogue*. 4th ed. Washington
3. Frâncu, Victoria (1996). *Building a Multilingual Thesaurus based on UDC*. In: Knowledge Organization and Change: proceedings of the 4th International ISKO Conference, 15-18 July, Washington, DC. Ed. by Rebecca Green. Frankfurt/Main: Indeks Verlag, pp.144-154
4. Frâncu, Victoria (1999). *A Universal Classification System going through Changes*. In: Albrechtsen, H. and Mai, J.-E. (Eds.) Advances in Classification Research : Proceedings of the 10th ASIS SIG/CR Classification Research Workshop, October 31, 1999, held at the 62nd ASIS Annual Meeting, Washington, D.C., pp. 55-71
5. Frâncu, Victoria (2000). *Harmonizing a Universal Classification System with an Interdisciplinary Multilingual Thesaurus: Advantages and Limitations*. In: Dynamism and Stability in Knowledge Organization: Proceedings of the Sixth International ISKO Conference, 10-13 July 2000, Toronto, Canada. Ed. by Clare Beghtol, Lynne C. Howarth, Nancy Williamson. Würtzburg: Ergon Verlag, 409 p.
6. Frâncu, Victoria (2002). *Language-Independent Structures and Multilingual Information Access*. In: Challenges in Knowledge Representation and Organization for the 21st Century. Integration of Knowledge across Boundaries: Proceedings of the Seventh International ISKO Conference, 10-13 July 2002, Granada, Spain. Ed. By Maria J. Lopez-Huertas with the assistance of Francisco J. Munoz-Fernandez. Würtzburg: Ergon Verlag, 607 p.
7. International Organisation for Standardisation (1986). ISO 2788: *Guidelines for the establishment and development of monolingual thesauri*. 2nd ed. Geneva: ISO

8. International Organisation for Standardisation (1985). ISO 5964: *Guidelines for the establishment and development of multilingual thesauri*. Geneva: ISO

9. MDA Archaeological Objects Thesaurus (1997). Available at: http://www.mda.org.uk/archobj/archcon.htm

10. National Information Standards Organization (1994). Guidelines for the Construction, Format, and Management of Monolingual Thesauri. ANSI/NISO Z39.19-1994. New York: NISO

11. Riesthuis, G.J.A. (1999). *Searching with Words: Re-use of Subject Indexing*. In: Extensions and Corrections to the UDC, Vol. 21, pp. 24-32

12. UDC Pocket Edition (1999). *Universal Decimal Classification Pocket Edition*, PD 1000. London, British Standard Institution, 288 p.