Multilingual access to information using an intermediate language: Proefschrift voorgelegd tot het behalen van de graad van doctor in de Taal- en Letterkunde aan de Universiteit Antwerpen

UNIVERSITEIT ANTWERPEN

Faculteit Taal- en Letterkunde
Germaanse Taal- en Letterkunde

Multilingual access to information using an intermediate language

Proefschrift voorgelegd tot het behalen van de graad van doctor in de Taal- en Letterkunde aan de Universiteit Antwerpen

te verdedigen door
Victoria FRÂNCU

Promotor : Willy Vanderpijpen

Antwerpen, 2003

Dedicated to the memory of my father

Acknowledgements

The experiment described in this thesis could not have been possible without the support and assistance and sympathy of many people of which I shall mention some in the following lines.

First of all, I am extremely grateful to my supervisor, Prof. Dr. Willy Vanderpijpen, whose assistance and valuable comments guided me throughout the elaboration of the thesis. Despite his lack of time and the multitude of his responsibilities, he would find a break between two of his tasks to answer my questions when necessary. Our brainstorming working sessions always ended up in productive assignments that meant steady progress for my work whenever we met.

I am also thankful to Dr. Gerhard Riesthuis who configured and implemented (twice) the experimental database I worked with. Had it not been for his programming skills, patience and inspiring challenging questions, the many different ways I needed the experimental database at different stages in the development of my study would have hardly been accomplished.

I owe much of the result of my efforts, materialized in this work, to the support both of them provided by participating in our meetings and critically reading the earlier versions of my thesis. There's nothing more helpful for such an undertaking than having a critical opinion on one's work.

At certain moments in time I was lucky to have met the right persons who, without knowing, had a meaningful influence over me and my work: Hanne Albrechtsen, Clare Beghtol, Michelle Hudon, Jens-Eric Mai, Robert Fugman, Gerhard Knorz, Jennifer Rowley, Nancy Williamson, Johann van der Auwera, mostly belonging to the ISKO (International Society for Knowledge Organization) community, but not necessarily so. Special mention is deserved by Ia McIlwaine who has been my intellectual model for many years.

I am much obliged to the UDC Consortium for giving me the permission to use the UDC Master Reference File for study purposes.

It is time to say a big “thank you” to my principals and colleagues from the Central University Library of Bucharest (BCUB) who accepted me to be away from my day-to-day work each time I had to travel to Belgium for my study. In the first place I thank Ion Stoica for allowing me to use the bibliographic database of the BCUB. Furthermore, I am grateful to each colleague in the Cataloguing and Indexing Department who gave the right answers to my questions whenever I needed them from remote places.

I would certainly not forget to thank my family who helped me by accepting to take over my responsibilities each time I was away from home. I am deeply thankful to my mother, my sons and my sister who took care of everything so that I could concentrate only on my study while abroad.

Thanks to all those who are not mentioned here and know that I entrusted during my study.

CONTENTS

INTRODUCTION

The multitude of information storage and retrieval systems nowadays make the information professionals more and more aware of the necessity to find solutions capable to break the barriers of all kinds standing against the access to information. In a world in which databases and all other types of information providers are reachable (even though situated at great distances from each other) within minutes from the moment the potential user sits in front of a computer screen, the only thing needed is a reliable information language.

While being theoretically so widely available, information can be restricted from a more general use by linguistic barriers. The linguistic aspects of the information languages and particularly the chances of an enhanced access to information by means of multilingual access facilities will make the substance of this thesis.

The main problem of this research is thus to demonstrate that information retrieval can be improved by using multilingual thesaurus terms based on an intermediate or switching language to search with. Universal classification systems in general can play the role of switching languages for reasons dealt with in the forthcoming pages. The Universal Decimal Classification (UDC) in particular is the classification system used as example of a switching language for our objectives.

The question may arise: why a universal classification system and not another thesaurus? Because the UDC like most of the classification systems uses symbols therefore, it is language independent and the problems of compatibility between such a thesaurus and different other thesauri in different languages are avoided. Another question may still arise? Why not then, assign running numbers to the descriptors in a thesaurus and make a switching language out of the resulting enumerative system? Because of some other characteristics of the UDC: hierarchical structure and terminological richness, consistency and control.

The problem will be approached by its two aspects: translatability between the natural languages used in building the thesaurus and compatibility in so far as the two types of information languages are concerned.

Translatability problems will be studied and discussed upon by comparing the three languages involved, (English, French and Romanian) in terms of both interlingual and intralingual aspects of synonymy, homonymy and polysemy as much as other lexical and semantic aspects.

One big problem to find an answer to is: can a thesaurus be made having as a basis a classification system in any and all its parts? To what extent this question can be given an affirmative answer? This depends much on the attributes of the universal classification system which can be favourably used to this purpose. Examples of different situations will be given and discussed upon beginning with those classes of UDC which are best fitted for building a thesaurus structure out of them (classes which are both hierarchical and faceted). The opposite situations, of classes which are not hierarchical and not faceted, will also be considered together with their possible solutions.

Compatibility issues will be discussed in as much as they occur between a classification system and a subject heading system. To be more specific, a comparative study will be made between classification notations and subject headings starting from the classification numbers as they are found in the Master Reference File and the assigned descriptors with examples taken from the online catalogue of the Central University Library of Bucharest (BCUB).

Aspects of compatibility and ways of harmonising classification notations and equivalent subject headings will be discussed considering the paradigmatic structure of UDC and that of a thesaurus based on it.

Before anything else we consider that some theoretical linguistic approaches to the information languages are necessary to clear the way to the purposed goal. This will be followed by a brief presentation of the three languages involved in our research – English, French and Romanian. The presentation of each of the three languages will point out both lexical and semantic aspects as far as they belong to different linguistic families: English being West Germanic and French and Romanian, Romance languages.

The second chapter will put forward the multilingual aspects of information storage and retrieval from the point of view of the search methods used. The discussion will mostly focus on formal cataloguing aspects.

Compatibility of information languages, a major topic and very much discussed in the literature of the field will take us deeper in the study of multilingualism. The possible solution of reconciliation or harmonisation of the many information languages, though there are authors who do not agree on that, could be the creation of an intermediate language for information exchange or a switching language. The revived interest for the subject that made the concern of the information scientists mainly in the 70's will be once again proved here, in the third chapter of the thesis.

The perspective of using the UDC as an intermediate language will be argued in the fourth chapter. This leads us to the key point of this research, namely the possibility of enhancing the search results (precision and recall ratios) by means of combined methods i.e. classification notations + subject headings/descriptors. Here too, aspects of multilingualism will be presented as far as they emerge from building a thesaurus based on UDC and translating it in the aforementioned languages. The resulting multilingual thesauri will be presented as the end products of this research in the format provided by the MTM-3 Macrothesaurus program. Samples of each thesaurus structural configuration are shown in appendices at the end of the dissertation.

In order to fulfil our goal of demonstrating the strong qualities of a UDC-based multilingual thesaurus both as an indexing tool (allowing automatic indexing) and as an information retrieval tool (used postcoordinately in searching) two different thesauri with different levels of specificity have been developed.

Going further we shall introduce some of the new trends in multilingual access. A few major projects of European as much as international interest will make the substance of the fifth chapter. We shall discuss about MACS (Multilingual ACcess to Subjects), a distributed project initiated by the CoBRA+ working group (Computerised Bibliographic Record Actions), about Expo 2000, another project that unfortunately could not be put into practice yet some of its ideas have been developed into an operational system, about ETHICS (the ETH Library Information Control System) created and used at the Eidgenössischen Technischen Hochschule (ETH), Zurich and about multilingual and multicharacter library systems and the way they work in countries like Switzerland, Finland, and Israel. In the end of this chapter we shall consider the recent research results in the field of cross language information retrieval (CLIR) according to the latest Text REtrieval Conferences (TREC).

The online applications of the UDC as of utmost importance to the pursuit of our goals will be discussed in the sixth chapter in the form of a case study made on an experimental database. Our investigation is largely based on previous research done at the University of Amsterdam. Most of the study will be directed towards revitalising the users' interest in the qualities of the UDC, that, despite its respectable age, permits online developments given these qualities, if adequately explored.

An additional research follows with a view to explore the impact of specificity on the retrieval power of a UDC-based multilingual thesaurus. Issues of crucial importance for the performance of an information system are discussed particular attention being given to the main performance indicators like: recall, precision and relevance. Chapter seven demonstrates by multiple examples of searches conducted in the experimental database the way the search results are influenced by different degrees of specificity of the information languages. The more specific multilingual thesaurus developed for this purpose (based on the Pocket Edition of the UDC) proves to meet the requirements of a highly efficient information retrieval tool accomplishing the user's needs in terms of friendliness, high performance in searching and the expectations of the indexer who only has to respect the rules of classification.

The eighth and last chapter will come up with our final remarks after looking back at the methodology used in this dissertation. The general conclusions will bring evidence of the feasibility and effectiveness of our approach advocating its strong points yet not neglecting its weak points.

Attached to the text of the dissertation several appendices are included. The first is intended to give instructions to the user about the way the search codes should be used and briefly shows the field structure of the experimental database used in this research. The second and the third provide samples of the two multilingual thesauri in both alphabetical and systematic arrangement.

CHAPTER 1
INFORMATION LANGUAGES: A LINGUISTIC APPROACH

The languages used in information processing (information storage + information retrieval) have been labelled so differently by different authors that there is a need for making distinction among them.

*Indexing language (*or *index language)* seems to be the most agreed upon, therefore the most used term to denote the language used for the representation of the subject matter of a document by indexers and searchers (Foskett, 1971, Maniez, 1997, Fugmann, 1997). But another term, *documentary language*, is used just as well with the same meaning (Hutchins, 1975). Now, within these terms there can be made a subdivision in order to specify what they comprise. Hutchins, for example, includes in the documentary languages (DLs) the *indexing languages* (ILs) and the *classification languages* (CLs):

DOCUMENTARY LANGUAGES

INDEXING LANGUGAGES

CLASSIFICATION LANGUAGES

Maniez (1997) extends the contents of the indexing languages to information languages and natural languages according to the following two-level scheme:

INDEXING LANGUGAGES

INFORMATION LANGUAGES

NATURAL LANGUAGES

and he does so in order to suggest a different opposition from the conventional one between INDEXING LANGUAGES and NATURAL LANGUAGES.

Speaking about indexing and the languages used for that there is yet another opposition still to be made, namely that between *controlled indexing* and *uncontrolled indexing* or *free indexing* performed with *controlled vocabularies* and *free-term vocabularies*, respectively. This distinction is nothing but another expression of the opposition suggested by Maniez. Controlled indexing operates with controlled vocabularies as intrinsic parts or tools of *information languages* (Ils) while free indexing takes terms as they appear in documents, from *natural languages* (NLs):

INDEXING LANGUAGES	vs.	NATURAL LANGUAGES
CONTROLLED INDEXING		FREE INDEXING
CONTROLLED VOCABULARIES		FREE-TERM VOCABULARIES

Henceforward we shall call *information languages* any and all the languages used in information systems and we shall distinguish within this framework controlled languages or *documentary languages* and uncontrolled languages or *natural languages*. Further, we shall divide the documentary languages in *indexing languages* (based on thesauri and subject heading lists) and *classification languages* (based on different systems of classification) as *Figure 1* shows.

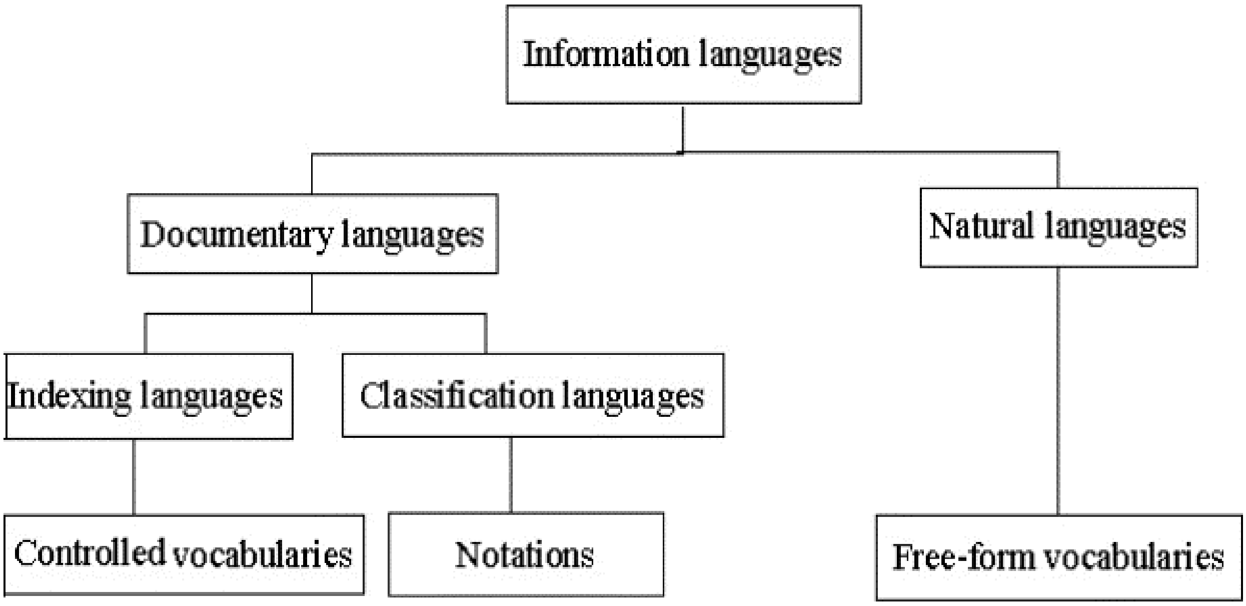


Figure 1. Diagram showing the classification of information languages

If not all authors agree upon the terms used to denote the information languages, they all agree on them being languages given their main characteristics: they have a vocabulary and syntax and they are systems of signs and communication (Lyons, 1970, 10-14, Hutchins, 1975, 3). Lyons gives as examples of languages the sign language, the language of mathematics, the language of the bees and the language of flowers. In the ongoing lines we shall have a closer look at the characteristic features of languages in general and draw a parallel between documentary languages and natural languages.

1.1 Documentary languages vs. natural languages

According to Ferdinand de Saussure (1964), semiotics or semiology has as central goal the theory of signs in all their forms and manifestations. The *signe* (French for sign) has two component parts: *signifiant* and *signifié*, that is to say it contains a *support* and a *concept* (which is random in the case of verbal signs). The example of the game of chess he gives, in which all the figures can be represented by any other object having the same value, will be referred to later on when aspects of semantic equivalence will be discussed. In this case, what counts are the features of each of the figures and the relations between them, therefore the coherence of the system.

Morris (1971) describes the five-place relation within the process of semiosis as a 'mediated taking-account-of' and illustrates it with the language of the bees. The participants in this relation are: the *sign vehicles* (**s**) as mediators of the process, the *interpreters* (**i**), as senders and receivers of the process, the *designata* (**d**), that is what is taken account of, the *effects of the process* (**e**) as the takings-account-of and the *context* (**c**), which are the external factors influencing the process.

In the process of human communication the sign (**s**) is the sequence of physical sounds or written marks (the lexeme or more commonly called word); the interpreters (**i**) are the speakers (writers) or the hearers (readers); the designata (**d**) are the relationships between the physical form of signs and the objects they refer to; the effects (**e**) of signs are the changes evoked by designata in the disposition of interpreters and the contexts (**c**) are the textual and situational environments in which communication takes place.

The semiosis process in the case of an information language can be described similarly. The signs (**s**) are the sequence of physical forms the information language is using, the interpreters (**i**) are the indexers and the users, the designata (**d**) are the relationships between the physical form of the documents and their subject content, the effects (**e**) are the reactions of the users about the relevance of the information language insofar as they can interpret it and the context (**c**) is the physical arrangement of the index and the information system as a whole.

Jakobson formulates another theory of verbal communication in which 6 factors are involved (Jakobson, 1963):

Context
Sender……………..Message……………………Receiver
Contact
Code

Between the sender and the receiver there is a message being sent. The verbal communication takes place inside one language, the sine-qua-non condition in this process. The code in this case is the natural language within which the two participants in the communication process understand each other. Apart from the code i.e. the language they both use permitting thus the contact between them there is one more element of crucial importance and that is the context, the situation or the instance in which the communication process takes place. It is only the context that gives meaning to the words. Taken out of the context the words alone might generate ambiguity thus hampering communication. Compare these two examples (Webster's, 1989):

"… most young Israelis… are tough, confident, hail-and-hearty." (The Economist, 26 July 1985)

"They hail from any number of Western states." (Garry Schmitz, Denver Post, 31 Aug. 1984).

The spelling 'hail' is correctly used for the noun and verb relating to icy lumps of precipitation and for the verb meaning "greet" or "acclaim". Words like 'hail', which have more than one meaning, can be misinterpreted if not placed in a context (see **§1.1.2** for details on homonymy).

Some words of current usage like 'cats' and 'dogs' completely loose their common meaning when placed in an idiomatic phrase (e.g. "it's raining cats and dogs" meaning it rains very hard). And such examples may go on. We deal here with what Wittgenstein (1958) called 'the language game' (see also **§3.4**). The philosopher argues in his theory that language is not strictly held together by logical structure, but consists of simpler sub-structures or language games. He goes on explaining that words do not denote sharply circumscribed concepts but are meant to mark family resemblance between objects identified by the concept. Words in natural languages only have meaning insofar as public criteria for their application exist. Therefore meanings are developed only in the use of words.

The indexing languages like the natural languages play the same role in information transfer as the latter do in verbal communication. They too work as a code in which the message is expressed in order to reach from the sender (the indexer) to the receiver (the end user) once the contact between the two has been established. The meaning of a descriptor or an indexing language element, roughly speaking, is in many if not all occurrences dictated by the context. We shall see further the overwhelming role the context has in the semantic disambiguation of the indexing language terms. In addition to that the semantic relations either hierarchical or associative have themselves much to say about the meaning of an indexing language term too.

Going back to Saussure's theory of language we shall remind here of the dichotomy he makes between *language* (langue) as a system of verbal signs and *speech* (parole) meaning the utterances produced by means of a language. This is very much to mark the distinction between *competence* and *performance* made by Chomsky (1965, 4) and refers in principle to the possibility offered by any language to express a multitude of utterances by means of the lexical units (words) existing in that language.

Many linguists identified *productivity* – or, as Chomsky (1968) calls it, *creativity* – as one of the universal characteristics of human language. This brings us to the basic principle of Chomsky's generative grammar namely that by virtue of this property each of the languages is characterised by the ability of its speakers to construct and understand an indefinitely large number of sentences in a natural way, without conscious application of grammar rules. The creative aspect of language was reduced to the explanation of the way in which 'names' are attached to 'things' or, more generally, the way in which meaning(s) of particular words and utterances is (are) associated with them (Lyons, 1970, 13). So too, in the Information Science we can distinguish information languages (systems of signs designed for describing the subject content of documents) and utterances produced through information languages, indexing formulas (Maniez, 1997). This can be a starting point in drawing the parallel between documentary languages and natural languages.

The first important level at which the comparison between the two types of languages can be made is that of the primary units they are based on, i.e. the descriptors and classification notations in the case of documentary languages (depending on whether they are thesauri and subject heading lists or classification systems) and the lexemes in the natural languages. Either of them are characterised by form and meaning.

Hutchins (1975, 12) uses the word 'descriptor' to denote the vocabulary element of any and all documentary, therefore controlled languages used in information systems. He argues that these 'descriptors' consist formally of combinations of graphic symbols (either numbers or letters plus punctuation marks) in order to compare them with the vocabulary elements of natural languages. In the latter case we deal formally with lexemes as combinations of phonemes. According to Hutchins a descriptor is ä string of one or more graphic symbols having signification within the language system". He continues arguing that "subject description may consist of just a single descriptor phrase and a descriptor phrase may consist of just a single descriptor and a descriptor may consist of just a single symbol". (p. 12). 'Descriptors' then, are, according to Hutchins, either single or compound terms having their own meaning in either an indexing language or a classification system *(Figure 2).*

DESCRIPTORS		
Indexing languages		**Classification systems**
Thesaurus terms		UDC notations
Single terms:	Birds (E) Oiseaux (F) Păsări (R)	598.8
Compound terms:	Songbirds (E) Oiseaux chanteurs (F) Păsări cântătoare (R)	598.81/.84
	Birds of prey (E) Oiseaux de proie (F) Păsări de pradă (R)	598.9

Figure 2. Examples of 'descriptors'

Hutchins therefore considers as ‘descriptor’ any ‘term of an indexing language’. In order to avoid confusion, we shall not call the elementary units of all documentary languages ‘descriptors’ since our purpose is not a comparison between documentary languages and natural languages, however useful it might prove at this point in our study. This is the more so as our intention is to make a comparison between descriptors in the commonly used meaning i.e. as vocabulary elements of a thesaurus on the one hand and classification notations on the other hand. Their compatibility and convertibility, along with the multilingual aspects of information access through them is to a great extent our purpose.

When comparing the vocabulary elements of documentary languages and natural languages we obviously note that in the documentary languages we have basically written forms whereas in the natural languages we have both vocal forms and written forms. This leads us to the distinctions necessary to be made between homonyms, as homographs and homophones and furthermore, to the next level of the vocabulary of these languages, the sememic level.

The common trait of both kinds of languages is that their vocabulary elements have meaning. An analysis at the sememic level of the two will point out the main difference between the polysemic nature generating redundancy and ambiguities in the natural languages compared with the attempted albeit not always achieved alleviation (if not elimination) of these shortcomings by disambiguation in the documentary languages. The existence of synonyms and homonyms shows there is no one-to-one correspondence between lexemes and sememes in natural languages (Hutchins, 1975), which means they have a plurivocal character. This is especially true for the English language, which is highly polysemic *(Figure 3)*:

Course	**Ground**	**Table**
Lessons	Surface of earth	Piece of furniture
Lectures	Floor of a room	Arrangement of columns and rows
Route (of a ship or aircraft)	Area of land (for sports)	List of multiplication of numbers
Series (of events, or treatments)	Reason	
Part of a meal	Basis	
Area in sports	Land	
Flow of a river		

Figure 3. Examples of polysemous words in English

The ideal information language should attempt to unify the lexemic level with the sememic level and thus have a bi-univocal character. That would mean: ”One subject for an utterance, one utterance for a subject” (Maniez, 1997).

The attempt of documentary languages to normalise the semantics of natural languages give the former the attributes of artificial languages: they use symbols as form of expression and they are designed for specific purposes or range of functions. *We can hence conclude that the documentary languages use special notations to express objects or concepts (as the classification systems UDC, DDC, LCC do) and they are standardised or normalised versions of natural languages (as the indexing languages are).* Their main functions are to reduce the redundancy and ambiguity of natural languages and to provide for consistency in indexing. They are also considered as channels of communication between documents and potential users.

In order that the information contained in a document reaches the potential user a translation process is needed and this process takes place at different levels. There has to be an essential kind of compatibility – a conceptual compatibility as Maniez (1997) calls it –

between the searcher's and the indexer's discourses i.e. on the meanings of the words they both use.

On one hand the information need has to be translated from the query (formulated in natural language terms) into index terms as they exist in the information system. Here the translation process is influenced by the context, that is to say by linguistic factors as much as by psychological factors. The user is trying to convert his own information need into index terms according to his own interpretation.

On the other hand the user finds in the information system the representation or the result of a double translation process. Firstly, the content of the document is reduced to its essential after the concept analysis was made. This is a *conceptual translation*. Secondly, the syntagms representing the essentialisation or summarisation of the document contents are formulated into index terms as neutrally as possible so that most of the major aspects of the contents are represented. This is a *linguistic translation.* The translation processes taking place during the information transfer can be furthermore influenced by the user's beed-back on condition that the information system has a provision for that (*Figure 4*).

The more the index terms and formulas used by the indexer to represent the subject matter of a document correspond to the search terms and formulas used by the user to represent his information need the higher the reliability of the documentary language. Fugmann (1999) strongly argues in favour of predictability as against consistency in subject representation. As long as the index terms and formulas are predictable, the recall rate will increase (see **§5.5** and **5.6**).

As mentioned above the documentary languages are channels of communication between documents and their users. Access tools created for this purpose mediate the information transfer. These tools provide the users (either searchers or indexers) with translation facilities between natural languages and documentary languages in the form of 'bilingual dictionaries' giving equivalents of natural language expressions in the form of classification notations or 'thesauri', showing when a natural language form has been adopted as an indexing language descriptor or how it is expressed in the indexing language (Hutchins, 1975, 9-10).

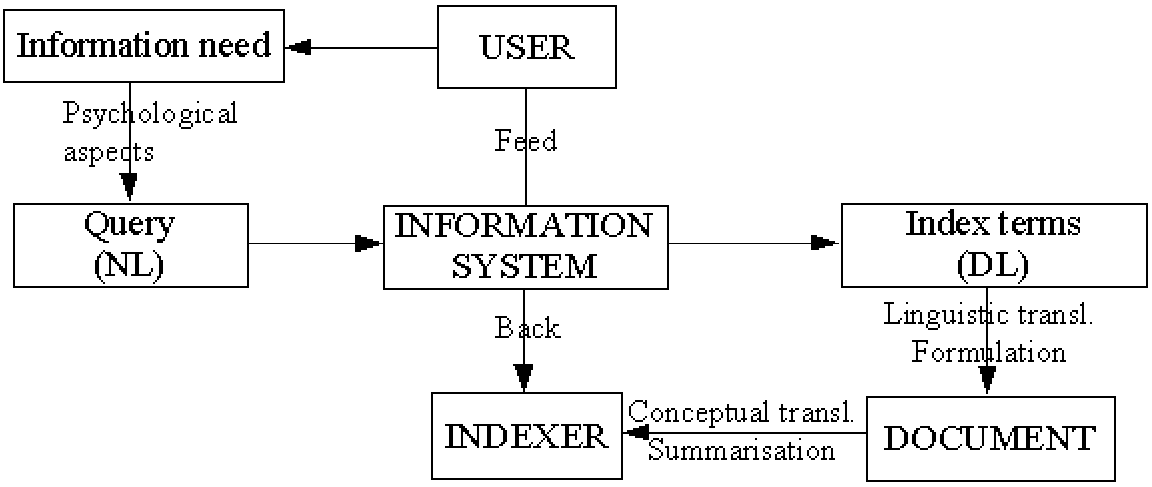


Figure 4. Diagram showing the translation processes involved in information transfer

As a rule, it should be possible to express in a documentary language any subject of a document and any subject of a search request addressed to the information system. Search requests fall into two major types: *specific reference* and *generic survey* (Foskett, 1971).

For 'specific reference' most of the documentary languages must have the capacity to form descriptor phrases for the expression of a complex subject. The searchers want only to retrieve those documents strictly corresponding to their query. This is performed by two methods: (1) the indexing terms are entered separately in an index file and that is done in post-coordinate systems and (2) the descriptor phrases (indexing formulas) are entered as a whole in the index file (pre-coordinate systems).

The syntagmatic organisation of a documentary language can either be a facet of equally its indexing function and searching function and that is the case of pre-coordinate systems or it can function just as a search language in 'pure' post-coordinate systems (Vickery, 1971). In the first case the information retrieval is performed directly, using the indexing formulas to search with whereas in the second there is need for a *search strategy* to mediate the information retrieval procedures.

With 'generic survey' things work differently as enquirers are able to consult documents on related subjects to those represented by the search request either broader or narrower or belonging to related fields. For that purpose the terms in the index files need to be related to each other generically, specifically or associatively in explicitly formulated paradigmatic structures (hierarchies in classification systems and thesaurus structures). Related terms can suggest strings of conceptual associations opening new perspectives for the search. Poly-hierarchies too, can generate such associations (Chmielewska-Gorczyca, 1997). It depends on the paradigmatic structure of each of the thesauri whether they permit or not such poly-hierarchies, i.e. the existence of more than one broader term. However, there are authors who do not agree on that. According to Soergel (1985) for instance, a descriptor must not have more than one broader term.

Some of the factors the designers of information languages should take account of are the type of users (interpreters) and their interests and the environment of the information system (context). In addition to that they should try to be as neutral as possible and not to impose their point of view on the users (critical indexing).

But the more organised the paradigmatic structure of a documentary language the less the possibility, quite convenient sometimes, of finding apparently not related subjects which often prove to fit most to the searcher's topic, the so-called serendipity. We shall see further the effect of very closed paradigmatic structures of indexing languages on their compatibility.

1.1.1 Paradigmatic structures of indexing languages

The paradigmatic structures of indexing languages can be accessed by means of such tools as authority lists, subject headings lists and thesauri (Hutchins, 1975, 89). Each of these tools can be considered as an indexing language lexicon or vocabulary (V), formally made of entry terms (T), descriptors (D) and a set of relations (R):

$$V=\{T, D, R\}$$

Hutchins identifies four types of relations within the vocabulary elements of an indexing language: *identity, substitution, inclusion* and *associative.*

1. The *identity relation* is established between entry terms and descriptors when they are the same. Both being in NL form, once the entry terms are assigned to documents for their subject description they can also provide access to it, acting like descriptors.

2. When the entry term is not a descriptor and the indexing language directs the user or searcher to the descriptor or set of descriptors we have a *substitution relation*. In this case the entry term is a non-descriptor and synonymous with another NL form which is preferred as a descriptor. They are both vocabulary terms but treated differently

e.g. *Familiar language*
see COLLOQUIAL LANGUAGE

3. The *inclusion relation* occurs when the sense of an entry term is included in the sense of another NL word or syntagm which has been preferred as a descriptor, the 'reversed synecdoche':

e.g. *Chronicles*
use HISTORICAL WRITINGS

The inclusion realtion defined by Hutchins as 'reverse synecdoche' is what Aitchison and Gilchrist (1987, 38) call 'upward posting', a technique which treats narrower terms as if they were equivalent to, rather than species of, broader terms. The effect is a decrease in the size of the vocabulary but with the advantage that access is retained via the specific terms to the broader terms used to represent them.

e.g. *SOCIAL CLASS*
UF Elite
Middle class
Upper class
Working class

Elite
use SOCIAL CLASS

Middle class
use SOCIAL CLASS

The same technique is called 'generic posting' in NISO's Guidelines for the Construction, Format and Management of Monolingual Thesauri (1994). This technique was extensively used as building method in one of our thesauri described in the fifth chapter (see **§5.3**).

4. The *associative relation* is established between the vocabulary terms when the sense of the entry term is expressed by a combination of descriptors or a descriptor formula. Lancaster (1972, 115) coins these entry terms as 'specifiers'

e.g. *English grammar*
see ENGLISH LANGUAGE and GRAMMAR

In a simpler and somewhat more practical manner, most of the thesaurus construction manuals and guidelines indicate two major types of relations within the thesaurus terms: hierarchical and associative (Aitchison & Gilchrist, 1987; ISO 2788, 1986; NISO, 1994). The four types of relations discussed by Hutchins are somehow retreieved, in a different formulation in these two. The issue of structural relationships will be resumed and discussed in more details in the following (see **§4.2** and **5.3**).

To say it in different words, the standard structure of an indexing language, either a thesaurus or a subject heading list includes: *terms, relations* and *instructions*. As earlier said, *terms* always refer to concepts and can they be descriptors and non-descriptors; *relations* link concepts to concepts and they are hierarchical – represented by broader terms (BT) or narrower terms (NT) – and associative – represented by related terms (RT). *Instructions* are given 1) in the form of scope notes (SN), intended both to indexers and searchers alike with a double role of giving a definition and indicating the use of the terms, and 2) as 'use' references, for the indexers or 'see' references, for the searchers respectively (*Figure 5):*

TERMS		RELATIONS		INSTRUCTIONS	
Concepts		Concept – Concept			
Descriptors	Non-descriptors	Hierarchical (BT, NT)	Associative (RT)	Scope notes (SN)	‘see’ / ‘use’ references

Figure 5.Thesaurus paradigmatic structure

Summing up the thesauri or subject heading lists have two purposes as they derive from their paradigmatic structure:

1. they indicate which of a number of synonymous terms is preferred in order to be used as a subject heading;
2. they show the logical and semantic relations between terms making references from the chosen term to other terms in the subject heading list or thesaurus structure.

The MDA Archaeological Objects Thesaurus (1997) gives a very comprehensive and accurate definition of a thesaurus underlying the two purposes stated above:

> “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 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 a thesaurus is to standardise the use of terminology, which not only helps in indexing information but also in retreival. Furthermore, it is a dinamic tool, one which can be developed through the addition or amendment of hierarchies, terms and relationships according to the need.”

1.1.2 Homonymy and its effects in natural languages and indexing languages

There are authors who define *homonymy* as an accident. Such an author is Buyssens (1943, 60) who argues:

> “L’homonimie est un defaut de perspective qui ne se produit que lorsqu’on isole artificiellement le signe du discours”.

Other authors, like Weinreich (1963, 178-179), define a homonym as a word-form, whether a phonological word (a homophone) or an orthographic word (a homograph), that has two or more distinctive sememes, i.e. two or more sets of semantic components having no members in common

e.g. *bank* of a river and *bank* as an institution to keep your money in

Polysemes can be defined as word-forms having sememes with a number of common semantic components and only a few distinctive ones; the sememes in effect are quasi-synonyms

e.g. *root* in botany, *root* in linguistics and *root* in mathematics

Distinctions between homonymy and polysemy are rather difficult to be made, yet, there are authors who outlined them (e.g. Cruse, 1986, Panman, 1982 etc.).

> "Homonymy is the phenomenon that two or more words have the same form and polysemy is the phenomenon that a word may have more than one meaning". (Panman, 1982, 107),

e.g. a *fine* woman and a *fine* irony

Lyons (1995, 60) considers that the problem of the distinction between homonymy and polysemy is in principle, insoluble. No matter how 'accidental' may linguists find homonymy and polysemy, semantic ambiguities can create special stylistic effects as in this line from Shakespeare:

> "Ask for me tomorrow and you shall find me a *grave* man" (Romeo and Juliet, Act III, Scene I).

Partial homonyms, i.e. homophone-heterographs, are two or more words which are identical in the phonic medium and different in written medium and meaning (Hanga-Calciu, 1997, 227). This can also result in pleasant stylistic effects for the natural language reader:

> "Mine is a long and sad *tale*!" said the Mouse, turning to Alice in sighing.
> "It is a long *tail*, certainly", said Alice, looking down with wonder at the Mouse's tail, "but why do you call it sad?" (Alice's Adventures in Wonderland).

We can formally distinguish homonymy from polysemy by the fact that the former is concerned with more than one word whereas the latter is considering different meanings within one word only. Hanga-Calciu (1997) argues that it is hard to imagine a natural language without this attribute (i.e. polysemic), as if language were the work of a mathematician or a mere photography of the world and not the result of a complex evolving process.

Since in natural languages homonymy has such a broad coverage we shall have a glance at its basic terms: homonyms, homographs and homophones.

A *homonym* is, according to Webster's Third New International Dictionary, one of two or more words spelled and pronounced alike but different in meaning (*pool* of water and *pool* as the game are homonyms). The same dictionary defines a *homograph* as one of two or more words spelled alike but different in origin or meaning or pronunciation (*fair* = market and *fair* = beautiful). A *homophone* is one of two or more words (like *to, too* and *two*) pronounced alike but different in meaning or derivation or spelling.

A special category of homonyms for which R. Quirk (1985) uses the term *homomorph*, share the same morphological form but belong to different word classes (e.g. *painting* as a noun, referring to the product and *painting* as a verb, referring to the process).

As aforesaid, we can have in English homophone-heterographs (*tale-tail*, *sad-said* or *to-too-two*), homomorphs (*painting*, noun and *painting*, verb). In French, homonyms that are at the same time homophones and homographs are rare (e.g. flic, grog, radar). Identity of pronunciation and spelling is very rare and this is because in the classical period great attention was paid to differentiate homophones by means of spelling (e.g. *dessein* and *dessin*, *compte* and *conte*). In Romanian, a phonetic language belonging to the same language family as French, identical pronunciation and identical spelling go together (e.g. the homonyms and homomorphs *vie (alive)*, adjective and *vie (vineyard)*, noun or the polysemes *rădăcină (root)* in botany, in linguistics and in mathematics). Hanga-Calciu (1997) formulates the conclusion that being highly language specific, homonymy acts differently in different language backgrounds and that also the relations among these terms (homonyms, homophones, homographs) differ from one language to another.

In documentary languages (only concerned with written forms) there is no such problem like homophony but homographs and polysemic words make disambiguation necessary. Being an extremely important topic especially in universal thesauri (Chmielska-Gorczyca, 1997), poly-hierarchy has been largely and vividly discussed on in the literature of the field, being coined in turn, ‘perspective hierarchies’ (Svenonius, 1997), ‘multiple location’ (Classification Research Group, 1957) or implied in the ideas of ‘heteroglossia’ (Jacob and Albrechtsen, 1997) and of ‘spatial analysis’ (Olson, 1997). Polyhierarchy is a recognition that concept terms may be ambiguous and therefore can have different meanings in different contexts (Vickery, 1997). The existence of several broader terms can disambiguate these meanings opening new search perspectives.

As a rule, each sense of a polysemous word or homograph has to be represented by a separate descriptor. The alphabetic index to a classification system may however contain the same term belonging to several classes (e.g. ‘rabbit’ may appear under Zoology, Domestic animals and under Hunting as well). Disambiguation is performed automatically here, each meaning being specified by a different class mark (e.g. in the UDC Master Reference File we find the term ‘rabbit’ in 9 occurrences):

56	**Palaeontology**
569	**Mammalia. Mammals**
569.32	Rodentia and Lagomorpha. Including: Extinct rodents. Extinct relatives of rats, mice, beaver. Extinct *rabbits*, hares
59	**Zoology**
599	**Mammalia. Mammals**
599.325	Lagomorpha. Including: Hares. *Rabbits*. Pikas
599.62	Hyracoidea. Hyrax (rock-*rabbit*). Daman (dassy, rock-badger)
631	**Agriculture in general**
631.225	Houses and structures for *rabbits* and rodents
636	**Animal husbandry and breeding in general. Livestock rearing. Breeding of domestic animals**
636.92	Domestic *rabbits*
637	**Produce of domestic (farmyard) animals and game**
637.55’712	Meat of small furred animals. *Rabbit* and hare meat
639	**Hunting. Fishing. Fish breeding**
639.112	Small furred animals. Small game generally. Including: *Rabbit*. Hare. Beaver
675	**Leather industry (including fur and imitation leather)**
675.31.8	Skinns of small domestic animals (and allied wild species). Including: Hare skins. *Rabbit* skins. Skins of dog, dingo, etc.
677	**Textile industry**
677.354	Hare fur. *Rabbit* fur

Going back to the paradigmatic structures of the indexing languages we find it necessary to say more about the difference, if any, between the sense of a lexical unit in a natural language and that of a corresponding descriptor in an indexing language. Consider the word ‘sonnet’as an example. The lexeme ‘sonnet’ may have two sememes: one denoting the product of the literary genre, a poem, and the other, a fix form of versification of 14 lines having a formal rhyming scheme. Each of the meanings places the term under a different hierarchy: the former, under literary genres, actually under poetry, the latter under prosody, more specifically under Italian verse forms. Disambiguation being necessary for both the indexer and the searcher, each of the two meanings have to be pointed out. This can be done by providing a scope note for each and additionally by making singular/plural distinction.

Therefore, in the alphabetical display of a multilingual thesaurus (English-French-Romanian) we can have the following thesaurus structure:

SONNET
F: Sonnet
R: Sonet
UDC: 801.675.2
SN : Used to denote a fix form of versification of 14 lines having a formal rhyming scheme
BT : Italian verse forms
RT : Sonnets

SONNETS
F: Sonnets
R: Sonete
UDC: 82-193.3
SN : Used to denote the literary genre
BT : Short poetic forms
RT : Sonnet

Many, if not all, of the senses of the descriptors in an indexing language are determined by their position in the paradigmatic structure (Hutchins, 1975, 118). In addition to that, scope notes are intended to direct either the indexer or the searcher, or both of them, to the specific meaning of the descriptor which can be somewhat different from that of the natural language form.

This brings us to the most important requirement of a documentary language, that of achieving the essential kind of information compatibility, the conceptual compatibility.

> “Compatibility of terms largely depend on the decision and delineation of the domains covered by the different term systems to be considered”. (Schmitz-Esser, 1996).

Schmitz-Esser gives clarifying examples of the way the meaning of a term is influenced by the domain it belongs to (see also p. 45): INITIATION is beginning of a process (in physics) or the introduction into manhood (in sociology) and POSITIVE is something good (in ethics), something bad (in medicine) or a piece of film (in photography).

It is more the context than the isolated term or syntagm that gives the meaning of a documentary language vocabulary element. Thereof Hutchins (1975, 118) concludes that to a large extent syntagmatic ambiguity can be eliminated by contextual evidence. We shall see further on how tremendously important context is in indexing languages based on natural language elements.

In pre-coordinated documentary languages the context is pre-established and the structure of the indexing formulas is identical with that of the search formulas (e.g. the classification notations in UDC and DDC and the strings of subject headings in LCSH or MeSH). Never-the-less, pre-coordination is not an absolute feature of these documentary languages since different classification notations can still be combined at the moment of search by Boolean operators. Likewise, if more than one subject heading string are assigned to documents it is possible to combine them and have various search results.

By contrast, in post-coordinated documentary languages, and the typical example here is offered by thesauri, the descriptors assigned by indexers, either single terms or compound terms, can always be combined at the moment of search in a practically unlimited number of variants. If we take, for example, a descriptor like ‘morphology’ and make a search in a bibliographic database[1], we shall trace it associated with many others from different domains as illustrated by the bibliographic records below:

[1] The titles used as examples belong to the online catalogue of the Central University Library of Bucharest (BCUB) and the descriptors have been translated for this purpose

Title 1:

Neflexibile indo-europene / Ioana Costa . – Bucuresti : Universitatea din Bucuresti,1995

UDC notations:	*Descriptors:*
811.1/.2’367.63(043) 811.1/.2-115(043)	Indo-European languages, Comparative linguistics, **Morphology**, Synsemantic words, Doctoral dissertation

Title 2:

Meaning and the English verb / Geoffrey N. Leech. – 2nd ed . - London : Longman,1987

UDC notations:	*Descriptors:*
811.111’367.6 81’22	Linguistics, Meaning, English language, **Morphology**, Verb

Title 3:

La semantique des adjectifs en langues romanes / Sorin Stati. - Saint-Suplice de Favieres : Editions Jean-Favard, 1979

UDC notations:	*Descriptors:*
811.13’367.6 811.13’37	Romance languages, Semantics **Morphology**, Adjective

Title 4:

Studies in Pre- and Proto-morphology / ed. by Wolfgang U. Dressler . – Wien : Verlag der Osterreichischer Akademie der Wissenschaften, 1997

UDC notations:	*Descriptors:*
81’366’276.3-053.2 372.46-53.2 159.922.7:372.46	Sociolinguistics, **Morphology**, Rudiments of speech, Native language, Usage of language, Child psychology

Title 5:

Structuri morfo-sintactice de baza in limbile romanice : pentru uzul studentilor / Sanda Reinheimer Rapeanu . – Bucuresti : Universitatea din Bucuresti, 1993

UDC notations:	Descriptors:
811.13’366/’367(075.8)	Romance languages, **Morphology**, Syntax, Handbook

Title 6:

Morphology of plants and fungi / Harold C. Bold et al. – New York : Harper & Row, 1980

UDC notations:	*Descriptors:*
581.4:582.28 581.16:582.28 582.28:581.16	Botany, **Morphology**, Plant anatomy, Mycology, Phycomycetes, Reproduction

Title 7:

The evolution of man : a brief introduction to physical anthropology / Gabriel Ward Lasker. - New York : Holt, Rinehart and Winston, 1961

UDC notations:	*Descriptors:*
572.5/.7	Anthropology, Somatology, **Morphology**

1.1.3 Synonymy in natural and documentary languages

According to Hornby (1989) a *synonym* is a word or phrase with the same meaning as another in the same language, though perhaps with a different style, grammar or technical use. The example given is ‘slay’ and ‘kill’. For the adjective *synonymous*, the author of the dictionary appreciates: having the same meaning, ‘slay’ is synonymous with ‘kill’ (though it is more forceful and rather dated).

The aforementioned definition admits that similarity is not perfect between the meanings of two words considered as synonyms. Near synonyms are quite frequently met in any of the natural languages. Consider the word ‘morals’. How far goes the degree of similarity between ‘morals’ and ‘ethics’? Lyons (1968, 447) states that in order to be real synonyms, lexemes should be interchangeable in all contexts and have identical cognitive and emotive import (see also **§1.3.2.1**).

Hutchins (1975, 37) considers that a stricter definition of synonymy would be that lexemes have the same sense, meaning that only the cognitive sense is taken into account and not the emotive sense. He concludes that two lexemes can be called synonyms if they have one sense in common. It is therefore only the semantic component, in other words the denotation and not the connotation(s) of the lexeme, which really matter.

Lyons (1977, 198-200) makes a clear distinction between sense and reference. He argues that “expressions may differ in sense but have the same reference” while citing Husserl’s example ‘the victor at Jena’ and ‘the loser at Waterloo’ both of which expressions refer to Napoleon (cf. Coseriu & Geckler, 1974, 147). Lyons states further that expressions with the same reference should not always be intersubstitutable in all contexts. The example he gives is Frege’s (1892) classic example of ‘the Morning Star’ and ‘the Evening Star’ which refer each to planet Venus. They have the same reference (Bedeutung) but they cannot be said to have the same sense (Sinn). Likewise, the author argues, what may be taken pre-theoretically to be non-synonymous expressions (like “my father” and “the man over there”) can be used to refer to the same person and, on the other hand, the same expressions can be employed to refer to distinct persons. Unlike the natural language, the information languages have a set of rules meant to simplify these complicated questions of meaning in the use of languages.

A simpler and, in a way, clearer dictionary definition of a synonym is given by Collins Cobuild (1992): a synonym is a word or expression which means the same as another word or expression and the example given is: “They loved the word ‘storm’ as a synonym for energy.” Such an explanation must have made Aitchison and Gilchrist (1987, 35) argue that in general linguistics synonyms are not common but they occur more frequently in scientific terminology.

In documentary languages the preferred term out of two synonyms is the most neutral one, lacking in connotations and emotional inflections and often, the choice tends to be in favour of the scientific term. In such a case the popular term becomes non-preferred and is mentioned just as an entry term

e.g.

Study of insects
use ENTOMOLOGY

Origins of language
use ETYMOLOGY

Terrestrial magnetism
use GEOMAGNETISM

Practical use of language
use PRAGMATICS

Scientific terms in botany, for instance, are highly recommendable and necessary, for reasons like compatibility of terms belonging to different indexing languages and for precision. (For the same issue see §**1.1.1** about the co-existence of Latin terms and their correspondents with which the former make semantic pairs in order to increase the number of access terms in an indexing language).

In botany and zoology the name variants are so diverse in both plants and animals, local name variants being likely to be met within relatively restricted geographical areas that Latin acts like a 'lingua franca', unifying terminology. Therefore, the preferred term will be in Latin and the non-preferred one, its popular correspondent, will be referred to as a non-descriptor but an access term as well

e.g.

LEGUMINOSA
UF Vegetables

ALLIUM CEPA
UF Onion

ALLIUM URSINUM
UF Garlic

ARACHNIDA
UF Spiders

Acronyms and abbreviations and their expanded forms are also considered as synonyms in information languages and therefore they are treated the same way i.e. cross referenced from each other,

e.g.

ISKO
use: INTERNATIONAL SOCIETY FOR
KNOWLEDGE ORGANIZATION

INTERNATIONAL SOCIETY FOR
KNOWLEDGE ORGANIZATION
UF: ISKO

and also:

AAT
use: ART AND ARCHITECTURE THESAURUS

ART AND ARCHITECTURE THESAURUS
UF: AAT

Another type of synonyms have to do with what is considered at a certain moment to be politically correct:

e.g.

DISABLED / IMPAIRED / HANDICAPPED
AGED / ELDERLY

In most cases the choice of descriptors from among synonymous terms should take account of the needs of the category of users the indexing language is intended to. In order to enhance the recall ratio of the controlled vocabulary, as many equivalents as possible should be included as entry terms.

Quasi-synonyms or *near-synonyms* are terms whose meanings overlap with each other to some extent but they are treated in controlled vocabularies as synonyms (Aitchison and Gilchrist, 1987, 37).

e.g.

URBAN AREAS / CITIES
CAR PARKS / PARKING SPACES
GIFTED PEOPLE / GENIUSES

Antonyms are also considered as a special category of quasi-synonyms (Aitchison and Gilchrist, 1987, 38). Mostly documents which discuss, say, problems on war, have a critical point of view and also discuss problems about peace. Antagonistic concepts co-exist in so many of the documents that this type of reference is almost mandatory to be made. However, if a clear distinction exists between the two opposite terms, they should both be used as indexing terms and references should be made from each other or else precision may be lost.

e.g.

WAR
see also PEACE

LITERACY
see also ILLITERACY

Upward posting treats broader and narrower terms as equivalents (see also **§1.1.1**). This device is more frequently met in thesauri having a rather low level of specificity as we shall see in an ongoing chapter (see **§4.3**).

1.2 Languages and language universals

The unity and diversity of language made the subject of piles of books.

It is generally agreed that the first of the fundamental properties of language is that it is uniquely human (Russell, 1948). Another of its basic attributes is that there are core properties that languages have in common and this is one of the crucial concerns of modern linguistics. ‘Language universals’, as they are referred to, allow us to say that all languages, are, in some sense, the same (Whaley, 1997, 4). This is claimed with the perfect awareness of the existence of roughly 4,000 to 6,000 languages currently in use. The unity of language is due to human biology, to the human inborn capacity for language (Chomsky, 1991). In his opinion, humans are genetically endowed with a ‘language faculty’ that permits the rapid acquisition of a complex and mature grammatical system (a universal grammar). This is the fundamental idea on which the structural grammar is based.

Whaley (1997, 6) mentions the purposes of language usage which are also universals: asking questions, scolding bad behaviour, amusing friends, making comparisons, uttering facts and falsehoods. In order to carry out these functional purposes the speakers need grammars to point out language similarities. The author goes on illustrating his point of view by underlying that it is the common experiences shared by humans which can account for language universals. For this purpose he cites Lee (1988, 211-212):

> “Despite the fact that I come into contact with quite a different set of objects than a Kalahari bushman, the possible divergence between our experiences in the world is circumscribed by a number of factors independent of us both, and even of our speech communities as a whole. For example, we can both feel the effects of gravity and enjoy the benefits of stereoscopic vision. These shared experiences exert a force on the languages of all cultures, giving rise to linguistic universals.”

The unity of language derives from a number of interactive factors, be they innate, functional, cognitive, experiential, social or historical. It is the domain of linguistic typology to define the factors which can account for certain common features of languages and to classify languages accordingly. Whaley defines typology as the classification of languages or components of languages based on shared formal characteristics. As it involves cross-linguistic comparison, one of the goals of typology is to identify cross-linguistic patterns and correlations between these patterns.

As aforesaid, languages have a number of common properties. The typological classification of languages into categories is based on such shared properties. The formal features of languages place them in classes based on (1) genetic relationships, (2) geographic location and (3) demographic features. In the first category we have languages that have a common origin, or belong to the same language family (e.g. Indo-European, Afro-Asiatic etc.). Considering the geographic location we can speak about Australian languages or Baltic languages. In terms of demographic characteristics, we can classify languages according to the number of speakers (e.g. languages spoken by more than 100,000 people).

In the line of the Saussurian theory, the meaning of any 'langue' is given by a combination of 'sèmes' (Saussure, 1964). The lexical meaning expressed by 'parole' is explained by the combination of 'sèmes' but much richer than that. The semantic components (Hutchins, 1975), or 'sèmes' are the minimal semantic elements having differentiated characteristics (Stati, 1979, 11). The minimal difference between two sememes is that of a 'sème' (Pottier, 1963), in which case we speak about a minimal pair: e.g. '*chaud / tiède'* which differ exclusively by the semantic component expressing gradualism (Stati, 1979).

Stati argues that the same semantic components exist in all Romance languages (1979, 14-15). Many other linguists suggest as universal semantic components*: 'animate', 'inanimate', 'human', 'agent', 'place', 'beginning'*. But answers to questions like: "Are there real language universals?" are still reserved and cautious (Chomsky, 1965a).

Further on, Stati (1979, 34) remarks that the same two semantic classes 'human' and 'inanimate' which characterise two senses of certain adjectives can characterise only one sense of some others (i.e. *un homme, film intéressant*, or rather *un enfant, climat insuportable*).

It can just as well happen that the same context formally represents two senses of a word. This has as effect what Quantz (1995) defines as *ambiguity*. He gives the following very simple examples:

Visiting relatives can be boring.

or:

They are *eating apples*.

According to Quantz, in the above given examples the English V-ing N construction produces, in the first case, semantic ambiguity, and in the second case, syntactic ambiguity. Ambiguities arise whenever a representation on a particular level (syntactic, semantic, pragmatic) can be mapped into more than one representation belonging to the subsequent linguistic level.

However, Stati (1979) argues that it is always possible to disseminate the different meanings of words used in different contexts and have for each an appropriate definition. He gives the examples of two adjectives in French, 'obligatoire' and 'difficile' and their contextual semantic variations in the syntagms:

une taxe obligatoire = une taxe qu'on doit payer
une disposition obligatoire = une disposition qu'on doit exécuter
un arrêt obligatoire = un arrêt auquel un moyen de transport doit s'arrêter

and:

une langue difficile (= à apprendre)
un texte difficile (= à comprendre)
une personne difficile (= à supporter ou à satisfaire)
un enfant difficile (= à élever)
une vie difficile (= à vivre)

The relevance and at the same time the importance of context for the information languages will be largely discussed about in the following chapters (see **§3.4**, **§4.3**, **§4.5** and others).

1.3 A brief presentation of three languages: Romanian, English and French

The three languages that we discuss about here have as common property their affiliation with the big Indo-European language family. Moreover, Romanian and French belong to the same sub-family of Romance languages, as we shall see further. Although this is an undisputable fact, there are authors who, by some reason or another, do not include Romanian among the Romance languages. Whaley (1997) argues the strong association between typological and genetic classification of languages by the unsurprising fact that “Spanish (Italic: Spain and Latin America) and French (Italic: France) both have articles that reveal gender or they both have subject agreement marked on verbs because we know that both languages have inherited these traits from Latin (Italic)” (p. 12). He concludes that “the typological similarity of the two languages is a function of their genetic association”. That is perfectly true.

Consider the following examples illustrating the masculine and feminine indefinite articles and the subject agreement marked on the verb in the two Romance languages Whaley mentioned (*Figure 6*). In Romanian the indefinite article differentiation by gender and the subject agreement marked on verbs functions identically as in the other two languages.

Masculine/Feminine	*Singular/Plural*
uno hijo - una hija (Spanish)	Esta una flor muy ermosa. (Nos) vamos a nadar.
un fils - une fille (French)	C’est une fleur très belle. Nous allons nager.
un fiu - o fiicã (Romanian)	Este o floare foarte frumoasã. (Noi) mergem la înot.

Figure 6. Gender opposition and subject and verb agreement in Romance languages

The typological similarity of the three Italic languages – as Whaley calls them – is obvious and it is indeed a function of their genetic affiliation with Latin.

1.3.1 Romanian - a mysterious Romance language

According to Pei (1976) the group of Latin-Romance languages encompasses Portuguese, Spanish, Catalan, French, Provençal, Sardinian, Italian, Rheto-Romansh, Dalmatian and Romanian. A whole chapter in one of his books entitled “The Story of Latin and the Romance Languages” is dedicated to what he calls ‘the mystery of Romanian’. We give below some citations from his book:

> “One more conquest of note was that of Dacia, the modern Rumania [sic!] which occurred under Trajan, around AD 100. This had far-reaching consequences that will appear later.” (p. 11).
>
> “Smallest of the Romance areas, and geographically detached from the others, is Rumania, which occupies approximately the same area as Trajan’s province of Dacia. But the movements of the Romanised Dacians and the Roman settlers are shrouded in mystery by reason of the third century invasion of the Goths. After more than a thousand years Rumania emerged. But the historical development of the country in its formative period is so strangely intertwined with the linguistic development of the Rumanian language and presents such puzzling features, that it is best left for later discussion in a separate chapter.” (p. 32).

Speaking about the difference between typology and areal classification and about the extent to which the structure of one language can be affected by the languages around it, Whaley (1997:13) reaches to the completely astonishing idea that Romanian is a Balto-Slavic language. He admits though that, genetically speaking, it is differently affiliated but he says that without specifying the subfamily of Indo-European languages Romanian belongs to.

Considering the lexical productivity, 78.8% of the Romanian vocabulary contain Latin basic units able to form derivatives (Dinu, 1996). Polysemy, another characterising criterion for a language, can be higher the more a word is built into phrases. Consequently, meanings of words can be phrase-conditioned. According to the number of meanings given in the dictionary, DLRM[2] contains words which can be grouped in 31 classes. The supremacy of the Latin basic fund in the higher ranked classes is even stronger than that of the lexical productivity. By way of statistical methods and mathematical linguistics, the author proves that among the 82 most productive Romanian words, 57 are of Latin origin (i.e. 69,5%) whereas from the 82 richest in meanings Romanian words, the inherited Latin lexical material is represented by 76 units (i.e. 92,7%). It is worth specifying that half of the 6 terms left originate from scholarly Latin: Rom. *linie* < Lat. *linea*, Rom. *punct* < Lat. *punctum* and Rom. *spirit* < Lat. *spiritus*.

In a research made by Constant Maneca (1966) over a corpus of 50,000 words, out of the 6,475 excerpted words, 1,007 lexical units were found to be most frequently used. Of these, 349 were of Latin origin and 54 of Slavic origin. Another research conducted previously by V. Suteu (1959) over another corpus of 50,000 words, the 522 most frequently used words of the total 4,547 excerpted, included 345 of Latin origin and 69 of Slavic origin. The high amount of Latin words found in each of these corpus-based researches make a clear evidence of the Latin origin of the Romanian language. The figures speak for themselves so we shall make no comment on them at this point.

1.3.2 English - "the sea which receives tributaries from every region under heaven"[3]

English, a West Germanic language, was briefly characterised by McCrum (1986, 51) as follows:

> "In the simplest terms, the language was brought to Britain by Germanic tribes, the Angles, Saxons and Jutes, influenced by Latin and Greek when St. Augustine and his followers converted England to Christianity, subtly enriched by the Danes, and finally transformed by the French-speaking Normans."

Contacts between the Germanic tribes of Angles and the people living in Friesland (the marshy islands of coastal Holland) account for the existence in English of words like: cow, lamb, goose, boat, dung and rain corresponding to the Frisian *ko, lam, goes, boat, dong* and *rein* (McCrum, 1986, 58). It was the Norman Conquest of 1066 which greatly produced the separation of English from Dutch and Danish (the language spoken in the land the German tribes originated from). Old English or Anglo-Saxon is still alive in modern English (more than 400 words). Computer-based analysis has proved that 100 most common words in English are of Anglo-Saxon origin (among them some basic lexical units like: *the, is, you, mann*).

After the Norman victory in the Battle of Hastings, Latin became the language of the church and Norman-French the language of the court and government circles. Yet, English survived as the Old English vernacular, both written and spoken, was too well established at

[2] Dictionarul limbii române moderne. Bucuresti, Editura Academiei RPR, 1958
[3] Ralph Waldo Emerson

that time and it was spoken by most of the common people who could not and would not accept the language of the foreign conquerors (McCrum, 1986, 75).

The mixture of all these foreign elements in the language of the inhabitants of the British Isles and provinces can explain the lack of unity of English.

After such a troublesome early history, after the refinement it knew in the Middle Ages, gaining its most brilliant expression through Shakespeare's works, the English language evolved surprisingly. It has become a global *lingua franca* by war, empire, broadcasting and more recently, by Internet and everything that comes with it. English words which are hard to be translated in other languages have penetrated various lexical systems *tale quale* or else, if they were opposed too strong resistance, they became hardly recognizable (e.g. '*logiciel*' is the French for 'software', '*pret-à-manger'* is the equivalent of 'fast food'). Another aspect of the English 'linguistic colonialism' is the existence in languages other than English, of derivatives specific to those languages but having English words as stems (linguistic calques): in Italian they have '*bufferizare'* (to buffer), '*debuggare*' (to bebug) and '*randomizzazione*' (random access); in Romanian, it is rather common nowadays to say '*a se loga'* (to log on), '*a scana'* (to scan), '*softist'* (software specialist), '*hardist*' (hardware specialist), '*a forvarda'* (to forward). This would not be so illegitimate if the respective English words had no correspondent in one or another of the borrowing language. But is it always so?

1.3.2.1 British English and American English

Speaking about the complexity and richness of languages, Alan Gilchrist (1972, 387-388), starts from the number of signs in the alphabet, then makes an inventory of the words in the *Oxford English Dictionary (OED)*, estimating it to contain about half a million of them. Further he estimates the number of words used by average individuals as active vocabulary in normal conversation and in writing, he compares it with the inactive vocabulary and adds the possibility of combining words into phrases. Finally, he states that though the *Thesaurus of Engineering and Scientific Terms (TEST)* contains 23,364 terms, these are generated from only 13,012 unique words.

Since *OED* is English and *TEST* is American, in case they are considered to be based on the same language, he proves it is not quite so by citing a part of a letter published in *The Guardian:* "When I am in Britain, I have a *car*. It has a *bonnet,* a *boot,* a *windscreen, wings* and a *silencer*. I run it on *petrol* and I drive it on the *road*. When necessary, I *mend* a *puncture*. When, as sometimes happens, I am in North America, I have an *automobile*. It has a *hood*, a *trunk*, a *windshield* and a *muffler*. I run it on *gasoline* and I drive it on the *pavement*. When necessary, I *fix* a *flat*".

Consider these pairs of underlined words in the citation above:

car *vs*. automobile
bonnet *vs*. hood
boot *vs*. trunk
windscreen *vs*. windshield
silencer *vs*. muffler
petrol *vs*. gasoline
road *vs*. pavement
mend *vs*. fix
puncture *vs*. flat

Are all these words perfect synonyms? Are they replaceable in any context? How about the last two? The New Merriam-Webster Dictionary gives as the 6th meaning of the word 'flat' used as a noun, 'a deflated tyre'. For the same word, the Collins Cobuild English Language Dictionary (1992) gives the 10th definition as 'a flat is also a tyre that has not enough air in it'. In British English the meaning of the sentence 'I mend a puncture' is somewhat different from the meaning of the same sentence in American English. In the former we have, semantically, the cause and in the latter, the effect. Synonymy is found here, unlike in the previous sentences, at sentence level alone, not at both lexemic and sentence level.

1.3.2.2 Aspects of contrastivity between English and Romanian

Contrastivity is based on the polysemic identity or non-identity relation between words belonging to different languages. Comparison was made between 2,700 most frequently used English words and their Romanian correspondents (Iarovici et al., 1979). The research resulted in a remarkably high number of 'cognates': 510 English words having total semantic identity and very close formal resemblance with their Romanian equivalents (e.g. Eng. *actor* = Rom. *actor*, Eng. *client* = Rom. *client*, Eng. *explorer* = Rom. *explorator*). If we add to this the number of partial cognates, i.e. 418 words, we can reach the conclusion that over one third of the most frequently used English words are semantically and formally identical with their Romanian correspondents. This is the more so if we take into account that some 'partial cognates' do not significantly differ in meaning from one language to another (e.g. Eng. *confuse*$_1$ = Rom. *a confunda* and Eng. *confuse*$_2$ = Rom. *a încurca*, Eng. *button*$_1$ = Rom. *nasture* and Eng. *button*$_2$ = Rom. *buton*).

The problem arises when we look at the list of 50 'deceptive cognates' and 137 'partly deceptive cognates' which are words with similar or identical form but different or partly different meaning across the two languages.

e.g. Eng. *actual* = Rom. *real, efectiv* (compared with Rom. *actual* = Eng. *present*)
Eng. *advertisment* = Rom. *reclam*ã (compared with Rom. *avertisment* = Eng. *warning*)
Eng. *library* = Rom. *bibliotec*ã (compared with Rom. *librarie* = Eng. *book shop*)

These words are also called 'false friends' and can be misleading and give difficulties when it comes to translatability issues. The problem of 'false friends' was discussed on a previous occasion in a comparative study on three language versions of the Universal Decimal Classification (Frâncu, 1997) when presenting the equivalents in French and Romanian for the English word 'rudiments' in the description of a UDC notation *(Figure 7).* The word 'rudiment' has as first meaning in Romanian, 'an organ which can hardly be seen, is growing or under-developed; beginning'; the second meaning is figurative, and usually in the plural, it is 'first elements of a theory, of an art etc.' (Marcu & Maneca, 1978).

UDC notation	English description	French description	Romanian description
372.46	Rudiments of language and speech	Apprentissage du langage et de l'élocution	Notiuni elementare de limbã si vorbire

Figure 7. Aspects of equivalence between different language descriptions of a UDC notation

1.3.3 French - the language of 'calembours'

French is based like all other Romance languages on vulgar Latin. Instead of introducing in a few lines historical facts or statistics about the language as a whole, we give below the nine semantic characteristics of French as they are presented by S. Ullmann (1952, 316-317) as a result of his researches:

- The French word is essentially arbitrary: generally the French words are not semantically motivated;

- The French word is essentially abstract: the language prefers the flexible terms, with general value, which can be interpreted according to the structure of the whole;
- The affective values are accomplished by delicate mechanisms, particularly by intonation and word order;
- The synonymous distinctions are clear and subtle. The French synonymy is a play over two pianos: one of the native's and the other of the scholar's;
- The French word is essentially polysemic (see also **§1.2** about different meanings of the French words *difficile* and *obligatoire*): the multiple meanings of the words, specified by the context, make a discrete device which compensate the lack of explicit motivation; its syntactic transposition and metonymic richness are classical forms of polysemy in French;
- French is a language of homonyms: the phonetic erosion multiplied the number of monosyllabic words and the prevalence of rhythmic groups as phonetic unit increased the danger of *calembours*; in speech, homonym words and word groups are semantically specified by context. At the end of the Middle Ages, the great "rhétoriqueurs" enjoyed the pleasure of inventing "equivocal" rimes (*louange - loup ange*) which are very difficult to be found an equivalent of in a foreign language. One of the masters of the French 'calembour' was Clément Marot (1496-1544) on whose grave stone a true friend had engraved: *C'est Marot, des François le Virgile et l'Homère* (Hofstadter, 1997, 3).
- The frequency of polysemy and homonymy increased the risk of unpleasant associations: the polysemic and homonymic impacts cause obsolescence in French;
- The semantic autonomy of French words is relatively low: many of the words need a context for being understood;
- There is no phonetic unity and no syntactic unity either in the French vocabulary: the function of word in the sentence is not specified in most of the cases other than by its determinants and its position. As it is not semantically motivated neither is it grammatically motivated and tends to become more and more so.

1.4 Conclusions

We start the current research on multilingual access to information stored in bibliographic databases with a comparison between documentary languages (DL's) and natural languages (NL's). Several semantic theories are mentioned among which those of outstanding authors like Ferdinand de Saussure, Jakobson, Chomsky, Morris, Lyons and Hutchins. By doing this we intend to clarify the meanings of the concepts used and to roughly explain the processes that take place during the information transfer with a view to identify which are the characteristic traits the two types of languages share in common and what the differences are between the two.

One of the main common traits found in both types of languages is that they operate with vocabulary elements that have form and meaning. In the first case the vocabulary elements are used in verbal communication among humans, whereas in the second they are used to represent the subjects of documents, in other words they give a secondary image of human knowledge stored in documents.

Whilst in a natural language the multiple meaning of a vocabulary element (i.e. word) is a token of the richness of that language and highly appreciated in an utterance, in a documentary language more than one meaning of a vocabulary element (i.e. term or descriptor) is inevitably problematic and has to be normalised. According to Maniez (1997) the ideal documentary language should attempt at providing "one subject for an utterance and one utterance for a subject". We concluded that the documentary languages use special notations to express objects or concepts (as the classification systems UDC, DDC, LCC do) and they are standardised or normalised versions of natural languages (as the indexing languages are).

The *information transfer* that takes place in any and all interactions between an information searcher and an information retrieval system is regarded as a communication process. While in the verbal communication process we have to do with a sender, a receiver and with a message being sent between them – to put it in a very simple way – in the information transfer the information need formulated by the user goes through a *multiple translation process* before it gets an answer.

First, the *information need* is formulated in natural language words by the *searcher*. Depending on the predictability of the information language these words will match up with the *indexing terms* used in the *information retrieval system* to a higher or lower degree. The higher this degree, the greater the effectiveness of the information language as such. The translation process dealt with here has linguistic in as much as psychological implications.

At the other end of the information transfer the *indexer* translates the *subject matter* of the document by reducing it to its essential. By means of a *conceptual translation*, the contents of the document are converted into index terms. It is beyond any doubt that linguistic aspects are involved here also since the concepts representing the subject of the document have to be expressed into index terms so that the major aspects of the subject are accurately represented (*Figure 4*).

A brief account on the *paradigmatic structure of the indexing languages* points out the four types of relations within their terms: *identity, substitution, inclusion* and *associative*. Some controversial issues like *homonymy, synonymy and language universals* are looked upon comparatively in natural and documentary languages. The way these particular linguistic categories are treated in documentary languages is argued with special concern.

In the end the three contributing languages used in this research – *Romanian, English and French* – are concisely presented with a few hints as to their history and characteristic features. Special reference is made to the strong *Latin character of the Romanian language*, the relatively recent penetration of English in quite a lot of lexical systems, some *contrastivity aspects of Romanian and English* and the *major semantic features of French*.

CHAPTER 2
MULTILINGUAL ASPECTS IN INFORMATION STORAGE AND RETRIEVAL

In a broader sense, according to the definition in the Collins Cobuild English Language Dictionary (1992), '*multilingual*' is 'something written or said in several different languages' or 'someone who is able to speak more than two languages very well'. If this definition is applied to the subject of this thesis then the multilingual aspects of the language of the document, that of the catalogue, the language of the OPAC as much as that of the user have to be considered.

Whenever we talk about the language of the catalogue we have in mind the help messages, the dialogue language used in the online public access catalogue, the added information introduced by the cataloguer in the bibliographic description and the bibliographic annotations. Depending on the performances of each library program, the help messages can be in more than one language. This feature can be used to the advantage of both the indexers and the users of the library system in that it permits clearer guidance in the particular functionality of that system. Such a system is the VUBIS integrated library system used in a large number of both public and academic libraries in Europe. Having as working languages English, French and Romanian (in the case of the product adapted to the Romanian market), the dialogue languages used in the OPAC can be selected according to the user's native or preferred language. The added information and the bibliographic annotations can provide for enhanced access to the bibliographic information in a situation like the following:

If the database contains, for instance, a book in Chinese with bibliographic annotations in the language of the catalogue and subject headings in 3 languages (say English, French and Romanian) the user has no access to the contents of the document because of its unknown language. Therefore that document will not be used. However, the user knows from the annotations that there is an extended abstract of the contents of the book included there. Hence a solution should be a translated title of the document in the same annotation field which prompts to the user that there is a translation available. It has always proved useful that the catalogue gives the translated title in as many languages as possible for any user who would thus have access to either of them.

As far as the information retrieval is concerned, two situations will be studied: one, in which the user is searching for a known item and another one in which the user is interested in a particular subject without any knowledge of authors or titles. In the first case there are three search methods available in almost every library systems that can be used:

a - title/uniform title and title/uniform title words
b - personal author
c - corporate author

In the second case, the information contained in the documents can be accessed by means of three other search methods:

a - words from the title
b - subject headings
c - classification notations

Except for the last method, language problems have to be dealt with. The second case, specifically, the multilingual aspects of subject access, make the substance of most of the research in this thesis.

2.1 Title/uniform title and title/uniform title words used as search methods

There are instances when the title of a book, even being known word for word, gives frustrating search results. The title of a book can be different in its second edition compared with the first. For instance, Gerald F. Corey's book entitled *'Issues and ethics in the helping professions'* was published as the 2nd edition of *'Professional and ethical issues in counselling and psychotherapy'* and Arthur C. Guyton's *'Basic human physiology'* had as its 3rd edition a completely different tile, i.e. *'Human physiology and mechanisms of disease.* A simple mention in the annotation field will be not enough for the user to find them both in one search. There are library systems which have a provision for 'related editions' as VUBIS has, but that field only works for different editions published under the same title. For different titles the solution is to make a uniform title as shown below.

With *title* and *title words* as search methods (which can be expanded to abstracts and full text words) the search result can be quite satisfactory provided that:

a - the searcher is using a very specific word for the query, that is to say the word used as search key should be as meaningful as possible for the subject of the document;

Discussion:

For this purpose, a word like 'technology' or 'engineering' will be of no relevance for the query since they retrieve too many titles[1]. It would make sense here to perform a second search using the Boolean operator AND in order to restrict the search result.

b - the input is correctly spelled (differences may occur between British-English spelling and American-English spelling (e.g. *catalogue* vs. *catalog*, *colour* vs. *color*, *organisation* vs. *organization*; compound words can be spelled with blanks and hyphens placed differently, e.g. *'pre-coordination'* but also *'preco-ordination'* and *'precoordination'*);

Discussion:

Spelling and typing errors, occurring both on the cataloguer's side and on the searcher's side, can generate false answers to the queries. If, for instance, the title of a book about a person has the name of that person mentioned in the title but the name is wrongly typed in, (e.g. 'C.S.Louis') the book will not be retrieved by the last name 'Louis' because the blank space is missing in front of it.

Stop words too, can be a source of search failure if not explicitly displayed or made known to the users of the catalogue. An entire search can sometimes consist of stop words. Such an example is the journal entitled "And" (Yee, 1998, 83).

Other characters like the Greek letters "φ" or "π", or particular symbols like "C++", when used as search key for known items will never give a satisfactory search result. An example of such a title is that of a Romanian novel by Dumitru Radu Popescu, which consists of only one character, "F". Another example is "Istoria numărului π" (The Story of Number π) by Florica Câmpan. For all these situations a solution should be found so that the expected information is retrieved (in the latter case, a search using the author's name will retrieve also this title).

[1] 372 hits and 401 hits respectively, as a result of such a search in the BCUB catalogue on February 28, 1999

c - the truncation sign is placed after the host word and before the grammatical clitics so that all possible forms are brought together

Discussion:

A search based on words from title using the truncated form ‘mycolog$’ will retrieve the following:

MYCOLOGIA, MYCOLOGICAL, MYCOLOGIE, MYCOLOGIST, MYCOLOGY

As for the search result, it may be different, depending on the words selected: a distributed result of 9 hits will include titles in English and French. The result will have no Romanian in the list because In Romanian, the word is spelled with an “i” and not with a “y” after the initial “m” (Rom. Micologie). Some more titles will be displayed when a further inquiry will use a subject heading as search key. The term ‘mycology’, used as a descriptor for a query, will retrieve all documents on mycology, regardless of their language, provided they were indexed with it.

The number of languages in which the user is formulating the query affects the search method and the result of its use. In theory, the user will only search for documents whose language is known to him. In practice this can be misleading for various reasons. Here is an example where language plays quite an important role: Douglas A. Hofstadter’s book ‘Le ton beau de Marot’. The title of the book is in French and the contents in English in the first place. Then, the subject has little to do with the French poet Marot, but much more with the theory of translation applied to translating poetry. The first two meaningful words can be easily misinterpreted if not seen the way they are spelled. In other words, ‘le ton beau’ meaning ‘the sweet tone’ can be taken for ‘le tombeau’, meaning ‘the tomb’. What we have here is contextual homophony and the example proves how a title, or words from that title, can be a source of errors when relating its meaning(s) to the subject of that document.

An even clearer example is following. The title of the book is “SPICE”[2] and it is the Romanian translation of “The SPICE book”[3]. The Romanian reader will first think of the meaning of the word ‘spice’, i.e. ears of cereal plants and take it for a book on agriculture or related topics. So will do the English reader, considering it a book about spices and the like. But surprisingly, the subject of both the Romanian and the English book is thoroughly different from what first comes to one’s mind: it deals with integrated circuits, particularly with what the author calls **S**imulation **P**rogram with **I**ntegrated **C**ircuit **E**mphasis. Therefore the title is a mere abbreviation of the name of this program. To make it even more misleading, but so much more interesting, the cover of the Romanian version has a nice picture of ears of barley in a field on it (see page 50).

Uniform titles and *uniform title words* can make a separate issue here though most of the problems are the same as in the case of title and title words used as search method. One of the differences is that there can be more than one variant of the same title put together under one authorised form. The 2nd revised edition of the Anglo-American Cataloguing Rules (AACR2R, 1998) points out the purposes of the uniform titles:

- To bring together all catalogue entries for a work when various manifestations (e.g., editions, translations) of it have appeared under various titles;

[2] Vladimirescu, Andrei. SPICE. Bucuresti: Editura Tehnica, 1999

[3] Vladimirescu, Andrei. The SPICE book. John Wiley & Sons, cop. 1994

- To identify a work when the title by which it is known differs from the title proper ;
- To differentiate between two or more works published under identical titles proper;
- To organise the file.

Since title information is considered a very strong and much used retrieval tool, the more title information is made available, including parallel titles in other languages and original title information, the greater the chance that the information needs of the library user are fulfilled (Goossens, 1993).

According to Yee and Layne (1998, 110) a uniform title brings together all the editions of a work, both by language and by chronological order. If the uniform title is not displayed during the search session, the user may be confused as to the reason why a record has come up at a particular point. Here is the example Yee and Layne give as single record displays including the uniform title:

The user browses through the works of Oscar Wilde and he decides to look at the editions of "The Selfish Giant":

```
Wilde, Oscar, 1854-1900
    1.          The selfish giant, 1911
    2.          The selfish giant, 1932
    3.          The selfish giant, 1945
    4.          The selfish giant, c1954
    5.          The selfish giant. Portuguese. O gigante egoista, c1982
```

If he chooses line 5 without any mention of the uniform title, then the display will be:

```
Wilde, Oscar, 1854-1900
    O gigante egoista / Oscar Wilde ; illustrado por Joana Isles . – Lisboa : Difusao
    Verbo, c1982
```

But if the uniform title is included, the user will know why this record is displayed at this spot:

```
Wilde, Oscar, 1854-1900
    [The selfish giant. Portuguese]
    O gigante egoista / Oscar Wilde ; illustrado por Joana Isles . – Lisboa : Difusao
    Verbo, c1982 (examples taken from Yee and Layne, 1998)
```

Lack of mention of the uniform title can generate scattering of the information like in the following example:

If a searcher is interested in finding all the editions of the famous medieval epic *'Tristan and Isold'* and the books on it, he or she will start searching with *'Tristan'* as a word from title. The result is 16 titles in the alphabetical order of titles including this word. But, as a matter of facts the alphabetic list of titles includes also books about Tristan Tzara, the Romanian avant-garde poet (3 titles) and about Tristan Bernard, the French writer (2 titles). Therefore, out of 16 titles only 11 will match his query. Still, this result does not mean that these 11 titles are all the library collection has as editions of and books about the medieval epic. The same information can be hidden under titles that do not contain the word 'Tristan' in them. And these titles are lost. They can not be found unless there is a uniform title to gather them all.

By contrast, when there is a uniform title mentioned and used to search with, loss or scattering of information is hardly possible. Consider this example of a search performed in the online public access catalogue of the Central University Library of Bucharest (BCUB)[4]. Using the word *'biblia'* (Romanian for Bible) – in this case also a uniform title – as a search key we get the result of 68 titles displayed alphabetically on the screen.

If we go on and modify this result restricting it by using a word from title, this time the English (and French word, as they are inter-lingual homographs), *'bible'*, the result will be 11 titles with this word in them. Out of this set of 11, 2 titles are French and 9 are English. But this is not reflecting reality, as there are presumably more documents of our interest in the database (i.e. English and French documents out of the initial set of 68). Therefore, we modify again the initial result restricting it by language of the document. We do that because it is not mandatory that the word 'bible' be found in the title. When we restrict the result by using the English language as a restriction criterion, we get 20 titles displayed on the screen. One of these has a title including the very common words used for the Bible, 'the Holy Scriptures' (i.e. 'The New World translation of the *Holy Scriptures* rendered from the original languages by…'). As for the French language as a restriction criterion, the result was 3 titles.

2.2 Personal authors used as search key

The case of personal authors can be discussed from two different perspectives:

1. when there is an authority file for names (not problematic as most if not all the possible name variants will be given there) and
2. when there is no such device to make things easier.

If we search for the works of Goethe in the same database as above, for example, and there is no authority record for the author's name, we have as search result 40 titles under 'Goethe, Johann Wolfgang von' and 4 titles under 'Goethe, Johann Wolfgang'.

Besides the name variants proper (Dickens, Charles vs. Dickens, Charles John Huffam) which are of no multilingual relevance, there are situations with proper names where the character sets have a lot to say about the quality of the search result. The transliterated name of Peter Ilich Tchaikovsky can be found as: Tchaikovskij, Tchaikowsky, Chaikovski or Ciaikovsky.

Transliteration of non-Latin script such as Cyrillic, if not used according to international standards, can give strange search results. If we make a search for an author and use the truncated *'veli$'* form as search key we get the following display:

1. *VELI2CKOVSKI*
2. VELICA
3. VELICAN
4. VELICANU
5. VELICESCU
6. VELICHI
7. *VELICKOVSKIJ*
8. VELICU
9. VELIHOV
10. VELIKORECKIJ, etc.

If we select line 1, we shall find out that the author, Veli2ckovski Paisij, Arhimandrit[5], is a well-known Romanian clergyman, the abbot of a monastery in Moldavia in the 17th century.

[4] The search was performed in the BCUB database on February 26, 1999

[5] Arhimandrit = abbot

The author mentioned at line 7, i.e. Velickovskij, Vladimir, is a Macedonian sculptor. They have both the same family names but each is differently transliterated. If the same search key is used to search for personal names as subjects we get the following display:

1. *VELICIKOVSKIJ*
2. *VELICKOVSKIJ*
3. VELIKIJ
4. VELIKOVSKY

If we select again line 1, we shall find the complete personal name as VELICIKOVSKIJ, Paisie. If we select line 2, the same personal author is under a differently spelled name, i.e. VELICKOVSKIJ, Paisij. Therefore, there are four forms of the same name in the database and the documents are scattered in the catalogue accordingly, since there is no authority file to authorise a uniform name heading to gather them all. This situation is created in the first place because different transliteration standards were used and secondly, because there is no authority record to offer spelling consistency. An authority record for such a name should include:

Heading:	*Name variants:*
VELICKOVSKIJ, Paisij, Arhimandrit	Veli2ckovski, Paisij, Arhimandrit Velicikovskij, Paisie Velickovskij, Paisij Paisie cel Mare

For consistency purposes such names should be treated similarly in both the formal and the subject catalogues.

In order to explain the uncommon presence of digits in the middle of a word transliterated from Cyrillic in bibliographic records we give below an example of such a record taken from the BCUB catalogue. The purpose of this strange marking was a graphical way of identifying particular Cyrillic characters in order to make them easier retrievable when their replacement was possible *(Figure 8)*.

```
+ 11769 / 28481 ---------------------------------------- Format: BCUBT --+
¦TIT:  1Samanizam, i arhajske tehnike ekstaze / Mir2ca Elijade, preveos  ¦
¦      francuskog Zoran Stojanovi1c. -  Sremski Karlovici, Izdavacka     ¦
¦      Knjizarnica Zorana Stojanovica, 1990. - 365p., 25cm. - Biblioteka ¦
¦      Theoria. - Tit. orig. în lb. fr: Le chamanisme: et les techniques ¦
¦      archaiques de l'extase. - ISBN 86-7543-010-8                      ¦
¦UDC:  821.135.1-96=163.41                                               ¦
¦701:  =163.41 Serbian.                                                  ¦
¦709:  821.135.1-96 Romanian - Works of science and philosophy as        ¦
¦      literature # Literature                                           ¦
¦MFN: 38288                                                              ¦
+------------------------------------------------------------------------+
```

Figure 8. Example of graphical marking of Cyrillic characters in bibliographic records

Critical problems pose the Greek and Latin personal names when translated, as much as the use of diacritics. But well-maintained authority records (syndetic structures) can solve them all. There are some examples in *Figure 9*, which give both Greek names and Latin names of authors as they can be found in authority records. Mention should be made here on the different language variants of each name and also on the translated names such as: *Ioan Zlataust, Ioan Gura de Aur, Sfântul Toma*. Not only the names consisting of attributes attached to the name proper but also simple names should be given both the original and the translated variant: e.g. Louis XIV may appear as Ludovic XIV in a Romanian catalogue and Ludwig XIV in a German one. The important thing is that each of these variants is cross referenced from one another, e.g.

LOUIS XIV UF: Ludovic XIV	Ludovic XIV Use: LOUIS XIV
LOUIS XIV UF: Ludwig XIV	Ludwig XIV Use: LOUIS XIV

Names with diacritics in them should be cross referenced too, so that all forms will be retrieved (e.g. Müller, Mueller, Muller). For reasons of more effective search possibilities some ‘warning’ displays should be inserted to suggest the users other variants that might be of interest. This can improve the catalogue’s predictability (Yee, 1998, 82-83).

Special East European diacritics can give even bigger problems as partly shown earlier. That is why the character set is important in information retrieval and even more so in information exchange procedures. Consistency in transliteration is meant to eliminate loss or scattering of information and hence increase the quality of the online catalogues as a whole.

Headings	JOHANNES CHRYSOSTOMUS (ca. 345 – 407)	THOMAS AQUINAS (1224 - 1274)	ARISTOPHANES (ca. 445-388 BC)	ARISTOTELES (ca. 384-322 BC)
Name variants	Giovanni Crisostomo	Aquinas, Thomas	Aristofan	Aristote
	Ioan Zlataust	Aquino, Thomas van	Aristophane	Aristotel
	Jean Chrysostome	Pseudo-Thomas Aquinas	Aristophanous	Aristotele
	Joannes Chrysostomos	Pseudo-Thomas von Aquin		Aristotle
	Joao Crisostomo	Thomas van Aquino		
	Johannes Chrysostomus	Thomas von Aquin		
	John Chrysostom	Thomas, Sanctus		
	Juan Crisostomo	Tomas de Aquuino		
	Pseudo-Chrysostomus	Tommasso d’Aquino		
	Ioan Gura de Aur	Sfântul Toma		

Figure 9. Examples of authority records for Greek and Latin names

A brief description is given in Chapter 4 about the way the Helsinki University Library coped with Russian but also non-Russian names in Cyrillic script (see **§4.4**). Transliteration in the latter case is preceded by phonetic transcription of the non-Russian names in Russian, the twofold process resulting in names that can hardly be recognised such as: ‘Šekspir’, ‘Vulf’ standing for Shakespeare and Woolf.

2.3 Corporate author and corporate author words used as search key

What we have in this case is a combination of problems of the same kind as in personal authors and those of the title words. Consistency will be given is these corporate author names by well-maintained authority files. An authority record for a corporate body has to include all name variants in all languages available and ‘see also’ references for any changes in names. We give in *Figure 10* an example of such an authority record according to AACR2 where we have in USMARC format:

110 fields for preferred term
410 fields for ‘see’ references to the term in the field 110
510 fields for ‘see also’ references.

```
110  20  $a American Institute of Electrical Engineers
410  20  $a AIEE
410  20  $a Instituto Americo de Ingenieros Electricistas
410  20  $a Amerikanskoe Obshchestvo Inzhenerov-elektrikov
510  20  $a Institute of Electrical and Electronics Engineers
```

Figure 10. Example of authority record for corporate bodies in USMARC

Acronyms are very much used for corporate bodies and in order to avoid confusion it is advisable that the abbreviated institution names be included in the authority record with a 'see' reference to their full names.

There is a question of translation to be discussed here too, since in many corporate names more than one language is used (e.g. Organisation for Economic Co-operation and Development - OECD and Organisation pour Co-operation et Développement Economique - OCDE or International Federation of Library Associations and Institutions - IFLA and Fédération Internationale des Associations des Bibliothèques et Institutions - FIAB). Some other corporate bodies have changed their official names (Fédération Internationale de Documentation vs. Fédération Internationale d'Information et de Documentation). As earlier shown in the case of uniform titles, multilingual authority data, especially for international corporate bodies, can fully satisfy the user's requirements. It is the task of well kept authority files to provide for and put together all name variants - including changes made in the official names and acronyms - in order to prevent information loss.

2.4 More issues about acronyms and some multilingual aspects

Acronyms can rise problems of multilingual access in yet another situation. In a classification system like the UDC alphabetical extensions to the enumerative class marks are permitted by the UDC grammar rules throughout the tables. The moment when natural language words are added to the numerical codes it is likely that some language restrictions appear. An example will clear this statement.

For the classification systems, the UDC has a general notation i.e. '025.4' described as 'Classification and indexing. Including: Indexing and retrieval languages. Classifications, thesauruses etc. and their construction'. For particular Classification systems i.e. '025.44/.47' some examples of combinations are given allowing an alphabetical extension which can be used to denote individual types of classification. Mostly this is represented in acronyms and the examples given in the UDC Medium Edition in English are obviously, in the English language. One may have for example: 025.45DDC and 025.45UDC to denote the Dewey Decimal Classification and the Universal Decimal Classification respectively. It is recommendable though, and it is very likely to happen so that each of the library catalogues have these acronyms in the language of their own. Therefore in a Romanian classified catalogue we will have 025.45CZD (standing for Clasificarea Zecimală Dewey – Dewey Decimal Classification) and 025.45CZU (standing for Clasificarea Zecimală Universală - the Universal Decimal Classification).

In the field of biology, we have the same circumstance, i.e. acronyms like 'DNA' and 'RNA' which most probably will be configured in the information language according to the language each catalogue is using with additional references to other language variants of the same concepts.

Another situation may however appear: a concept like 'The International Standard Bibliographic Description' which everybody in the library and information science field would know as ISBD is unlikely to appear in a catalogue or database otherwise but in this recognised form. Similarly, the well known thesauri like the ERIC (Educational Resources Information Center Clearinghouses) Thesaurus and the ROOT Thesaurus will be found under these acronyms and most probably with additional 'see also' references from their expanded forms.

The use of acronyms in information retrieval may have other kind of frustrating result. If a search is performed in the Internet with such acronyms as TREC and/or ASIS as search keys the result can bring a lot of irrelevant documents like those belonging to the **T**exas **R**eal **E**state **C**ommission and/or **A**merican **S**ociety for **I**ndustrial **S**ecurity along with the searched for and expected **T**ext **RE**trieval **C**onferences and The **A**merican **S**ociety for **I**nformation **S**cience respectively.

2.5. Subject representation used as search key: a comparative investigation

If a search is performed in a bibliographic database taking a *subject heading* in the language of the catalogue as a search key, e.g. *ACCIDENTE*, (we use the BCUB online catalogue again) we may have the following search results:

1	Accidente aeriene	3 hits
2	Accidente de munca	6 hits
See also:		
3	Primul ajutor	39 hits
4	Protectia muncii	45 hits
5	Medicina muncii	19 hits
6	Accidente industriale	2 hits
7	Accidente industriale	2 hits
8	Accidente maritime	1 hit
9	Accidente navale	1 hit
10	Accidente nucleare	11 hits
11	Accidente rutiere	6 hits
12	Accidente vasculare	3 hits

Had it not been for the "see also" reference that brings 105 hits more, our search result would have been really poor, consisting of only *33 hits*.

By comparison, if we repeat the query as a *word from title*, *ACCIDENT*, and consider some of its lexical variants: *ACCIDENTE, ACCIDENTELE, ACCIDENTELOR, ACCIDENTS, ACCIDENTUL, ACCIDENTULUI* we will have as result *78 hits*. By chance this word has the same stem in French and in English as well, so the result will include titles in these two languages too, along with the Romanian ones.

Going on with our comparison if we use the *UDC number* having this meaning i.e. *614.8 - Accidents. Risks. Hazards. Accident prevention. Personal protection. Safety. Public health and hygiene* we will get a number of *56 hits* as response to our query.

This situation may have several reasons:

1. not all the records in the bibliographic database have descriptors for subject representation hence the low number of hits in the first instance (33 hits)

2. the search result using the title word ACCIDENT as query will surely include works of fiction whose title contain this word hence the high recall rate (78 hits)

e.g.

Accident : a day's news / Christa Wolf
Accident banal : roman / Al. Simion
Accidentul : roman / Mihail Sebastian

3. the closest to reality, therefore the best response will be the third which is also a most complete and relevant one, given every record in the bibliographic database has a UDC number and in this set there's no way works of fiction might be included (56 hits).

2.6 Conclusions

To sum up all we said about the search methods used in the average information systems and their multilingual implications we can mention the two distinct situations:

- access to known items
- access to subjects

For the *known item* situation it is clear that the searcher will only look for documents in a language that he or she has a good knowledge of. This is unlike the second situation when a query representing a subject formulated in the language of the catalogue (that is presumably known to the searcher) can bring documents in as many languages as available in the catalogue as search result.

The *methods of title* and *title words* used as queries have the following critical aspects for the information retrieval:

- spelling or typing mistakes on both the cataloguer's side and the searcher's side;
- title in a different language from that of the contents (e.g. *Le ton beau de Marot*);
- metaphorical title or misleading title (e.g. *The SPICE book*);
- different editions of the same work published under different titles;
- different language variants of the same work (e.g. *The selfish giant*).

In the last two instances, the right solution is offered by uniform titles that gather all the title variants in the same place.

Personal authors used as search keys will be problematic in case of Greek and Latin names but also in transliterated names from non-Latin scripts (e.g. Cyrillic), names with diacritics, translated names. Well-maintained authority files give the appropriate solution for them all.

Corporate authors are likely to be troublesome in information retrieval if not controlled by means of authority records as well. Most of the international corporations have names in more than one language and acronyms and abbreviations can be troublesome if not included in authority records.

As far as *access to subjects* is concerned, experience has proved that uncontrolled information languages have more chances to give higher number of retrieved records than the controlled ones (see the examples given under point 2 in **§2.5**). It is also true that free text searching can bring about useful information as long as *subject headings* are predictable and serendipity is given a good chance. Yet, browsing long lists of retrieved documents can be rather time consuming. Otherwise, as proved in our 'accident' example, the *classification notations*, however unfriendly they might be, can get the less frustrating search results. The extent to which this 'unfriendliness' can be turned to our advantage it is for us to demonstrate in the coming chapters.

CHAPTER 3
COMPATIBILITY AND CONVERTIBILITY OF INFORMATION LANGUAGES

3.1 Definitions and types

Many of the present theories dedicated to compatibility between information languages have been formulated and presented on the occasion of the Research Seminar of the TIP/ISKO Meeting in Warsaw, 1995 gathered under the title "Compatibility and Integration of Order Systems" (Compatibility, 1996).

Highlighting the importance of compatibility issues for the information science in the line of the above mentioned seminar, Maniez (1997) makes a distinction between the convergence of indexing languages in which case we deal with inter-lingual compatibility and refer to 'the search for proximity or similarity' and the convergence of indexing formulas which can be reached by the classical device of translation. Further he cites Riesthuis (1996) who gives the most commonly accepted definition of the term 'compatibility' or 'convertibility':

> "Compatibility means that for each term A of an information language P there is a term A' in information language Q with the same meaning so that we can convert A into A' without changes in meaning" (p. 24).

Riesthuis (1996) mentions three forms of compatibility depending on the syntax level considered: term compatibility (e.g. Japan and Nippon), sentence compatibility (e.g. the conversion tables made by Scott (1993) between LCC-DDC and DDC-LCC) and subject compatibility. The definition he gives for this third type of compatibility reads:

> "An information language P is fully compatible with information language Q when a sentence that denotes correctly – using the vocabulary and syntax of P – the subject A of a document M can be translated, without re-indexing, to a sentence that denotes correctly – using the vocabulary and syntax of Q – as if subject A was indexed with language Q directly" (p. 25).

To illustrate this type of compatibility consider the way indexing is done at the Central University Library of Bucharest by simultaneous use of the UDC notations and descriptors from a controlled vocabulary stored in the library database (*Figure 11*):

Les relations hôtes-parasites dans le modèle Téléostéens-Métacercaires de Labratrema minimus (Trematoda bucephalide) / présenté par Elisabeth Faliex . – Grenoble : Atelier National de Reproduction des Thèses, 1991

UDC notations :	*Descriptors :*
578.23:597.5:**576.895**.122(043)	Parasitology
	Relations between virus and host cell
578.23:**576.89**5.122:597.5(043)	Teleostei (Fishes)
	Trematodes (Worms)
	Dissertation

Figure 11. Bibliographic record with fully compatible complementary indexing

This is the ideal situation when subject of the document indexed with the UDC is on a par with the subject denoted by the assigned descriptors. In such a case the information retrieval can be performed in the catalogue by using either of the two search methods (UDC numbers and subject headings/descriptors) with the same result. The subject of the document in the example is partly represented by UDC notations built in by parallel subdivisions according to the instructions in the UDC schedule (note the number for ‘Animal parasitology’ in bold letters in *Figure 11*). If we look at the captions, it is easy to notice that the subject as a whole denoted by the classification notations is the correct representation of the content of the document:

578.23	Relations between virus and host-cell
576.89	Animal parasitology
597.5	Teleostei
595.122	Trematodes

The concepts we have to represent here by classification notations are found in Classes 57 and 59 of the UDC. We need to connect them such a way that the subject is correctly denoted according to the UDC grammar. An indication given in the tables at 576.89 for Animal parasitology reads that 576.892/.899 is subdivided like 592/599. Therefore, these notations will combine subdivisions from Zoology with those from Biology in order to adequately represent the topic of the document (Frâncu, 1999b).

These notations found in the tables have to follow the instructions of use existing in any of the UDC editions in order to represent coherently the subject. The differences between the numbers in the tables and the classification notations built according to the rules may only surprise someone who is non-familiar with the grammar and syntax of the UDC. Likewise, the descriptors have a structure that is clearly stated and agreed on by the vocabulary makers.

Another example will make evidence of situations when full compatibility cannot be achieved. The solution is that the more flexible of the indexing languages considered will complete the more restrictive one:

Siebenburgisch –Sächsisches Wörterbuch: mit Benutzung der Sammlungen Johann Wolfs / Ausschuss des Vereins für Siebenburgische Landeskunde. – Berlin: Walter de Gruyter

UDC notations:	*Descriptors:*
811.112.2’28(498.4)(038)	Dialectology
	German language
	Saxons
	Transylvania
	Dictionary

Figure 12. Bibliographic record with partially compatible complementary indexing

The subject of the dictionary given as example in *Figure 12* is the German dialect used by the Saxon population in the historical province of Romania, Transylvania. It is a well-known fact that the German-speaking people in Transylvania are called Saxons but there is no UDC number for such a specific category of subjects. The lack of specificity in one of the indexing languages and hence lack of compatibility between the two is an evidence of the shortcomings generated by the social and cultural determination of any indexing language.

Long before the Warsaw seminar Glushkov et al. (1978) made distinction between two types of compatibility:

a) Semantic compatibility
b) Structural compatibility

The *semantic compatibility* takes into account the body of knowledge or the discipline the information languages refer to. More specifically this can be reduced to the lexical, paradigmatic and syntagmatic compatibility. In other words compatibility exists as a function in the representation of entities, activities, properties and attributes and the hierarchical and non-hierarchical relations recognised.

The *structural compatibility* is seen by the aforementioned authors as morphological compatibility (similarity in the structure of terms) and syntactic compatibility (similarity with respect to the structure of groups of terms or phrases).

If we compare these two types of compatibility with the three forms established by Riesthuis (1996) and the other two of Maniez (1997) we can sum up:

1. Compatibility issues are discussed in terms of both meaning and form;
2. Full compatibility can only be achieved when each of the indexing languages have the same level of specificity (see *Figure 11*);
3. Partial compatibility will not affect the meaning and coherence of indexing as long as a compromise is made towards the complementarity of the indexing languages and a set of rules established in order to provide for indexing consistency (see *Figure 12*).

3.2 Practical applications

The concept compatibility defined by authors like Maniez (1997) and Schmitz-Esser (1996) considers the two kinds of linguistic discourse working together in information retrieval: that of the indexer and that of the searcher. Each of the contributing parties in this process may have in mind the same concept, a single reality; but this can be mapped onto the indexing language in a way, which is not always identical with the representation the searcher has for that concept. M. Iivonen (1996) speaks about the selection of search terms as a "meeting place of different linguistic discourses". The ideal indexing language will be structured such a way that each term will point to only one concept and each concept will be represented by only one term.

Yet, as suggested above, the terms or vocabulary elements of an information language (except for the classification systems using alpha-numeric codes) are taken from natural languages. The natural language expressiveness is measured by the richness of its vocabulary and the capability of phrase-building. Therefore, each term or vocabulary element in an indexing language can be expressed in a variety of forms (synonyms) while referring to one and only one concept. On the other hand, a term can be referential for more than one concept (polysemic words) and then disambiguation is required for increased relevance. Consider for instance, the French word 'peinture' which has a first meaning as 'painting' but also a second one as 'colour' or 'paint'. A syndetic structure of 'see' or 'use' references will be in this case the solution to provide for consistency and control of indexing and to offer the searcher the authorised term for a particular concept.

As a rule, communication hence information transfer, is mediated and governed by language and specifically by semantics. But semantics, in its turn is largely dictated by social and cultural factors. One may have 'breakfast', 'lunch' and 'dinner' and if necessary, 'supper' in the Western Europe and the American continent but 'breakfast', 'lunch' (as the main and most consistent meal of the day) and 'supper' in the Eastern Europe. In Romanian language 'fruits' do not include 'grapes', which are always mentioned separately in an agricultural context.

A controlled vocabulary is a necessary evil, providing for correspondence between the target domain (the conceptual content of a document) and the modelling domain (the indexing language) while simultaneously enforcing a rigid, static and artificial environment unresponsive to the dynamism and heterogeneity that characterises both human knowledge and natural languages (Jacob & Priss, 1999, 93).

Speaking about the conceptual compatibility at the level of one and the same indexing language one has to consider the one-to-one relationship between a concept and the term it is represented by or the many-to-one relationship the natural languages are offering making access to information possible via natural language terms. The real problem of compatibility is yet more controversial when more than one information language is involved.

Preoccupations for the convertibility of indexing languages date from the 1960s and the research conducted by the Danish Dan Fink in 1964 resulted in a classification system used as a multilingual dictionary: the Abridged Building Classification for Architects, Builders and Civil Engineers (ABC). This is a specialised classification based on the UDC, translated into eleven languages and complemented by a Swedish designed faceted system (Cochrane, 1994, 12). Other proposals for automatic conversions and conversion tables have emerged but only few of them were used in practical applications.

The Unified Medical Language System (UMLS) is an impressive example of harmonizing different classification systems and thesauri having as immediate effect the convertibility of bibliographic records indexed in one system into another system.

The purpose of creating the UMLS (Hoppe, 1996) was to improve the availability of machine-readable information sources for both retrieval and integration of biomedical information. UMLS is a long-term research and development project of the National Library of Medicine (NLM) in Bethesda, MD, USA meant to unify the great variety of medical terminology existing in different information sources and to disseminate useful information among disparate databases and systems. To reach these purposes the NLM considered two things as necessary: a) new machine-readable knowledge sources and b) sophisticated user interface programs.

The UMLS Knowledge sources contain the Metathesaurus, the Semantic Network, the Information Sources Map and the SPECIALIST Lexicon.

The Metathesaurus is the main vocabulary component of the UMLS. It includes a number of 200,000 terms from more than 30 biomedical vocabularies integrated by means of lexical and semantic links. The Metathesaurus is organised in a three-level hierarchy (concept-term-string). Each name or string has a unique (string-) identifier and, for English language only, is linked to all its lexical variants by a common term identifier. The same string in different languages (e.g. English and Spanish) has a different string identifier for each language. For all strings linked to one term and all terms linked to one concept, respectively, the "preferred form" is stated.

The SPECIALIST lexicon consists of a set of lexical entries with one entry for each spelling or set of spelling variants of a particular part of speech. Entries that share their base forms and spelling variants, if any, are collected into a lexical record in the unit format.

The unit lexical record is a frame structure consisting of slots and fillers. Each lexical record has a basic slot whose filler indicates the base form and optionally a set of spelling variants= slots to indicate the lexical variants. Lexical entries are delimited by entry= slots filled by the entry unique identification number (EUI) of that entry. EUI numbers are seven digit numbers preceded by an "E". Each entry has a cat= slot indicating the part of speech. The lexical record is limited by braces ({…}).

The unit lexical record for "anaesthetic" illustrates some of the features of a SPECIALIST lexicon record:

```
{base=anaesthetic
     spelling_variant=anesthetic
     entry=E0008769
          cat=noun
          variants=reg
     entry=E0008770
          cat=adj
          variants=inv
          position=attrib(3)}
```

The base form "anaesthetic" and its spelling variant "anesthetic" determine a lexical record consisting of a noun and an adjective entry. The variants= slot contains a code indicating the inflectional morphology of the entry; the filler reg in the noun entry indicates that the noun "anaesthetic" is a count noun which undergoes regular English plural formation ("anaesthetics"); inv in the variants= slot of the adjective entry indicates that the adjective "anesthetic" does not form a comparative or superlative. The position= slot indicates that the adjective "anaesthetic" is attributive and appears after colour adjectives in the normal adjective order (UMLS, 1998).

3.3 Integration aspects

Given the strong points of the Universal Decimal Classification such as its logical structure and terminological richness, its universal coverage and lack of any particular bias, this classification system has been considered by many authors as a potential candidate to thesaurification (D'Haenens and Lorphèvre, 1974; Frâncu, 1997; Riesthuis and Bliedung, 1990; Riesthuis, 1997; Scibor, 1997; Frâncu, 1999). Many commentators assume the UDC to have the necessary attributes for an international exchange language, or switching language. But the issues of switching languages will be presented later in this thesis.

Some daring steps in the creation of thesauri based on separate classes of the UDC have been made at the Central University Library of Bucharest. Parts of Class 0 i.e. the subdivision 02 for Libraries (Dumitrăşconiu, 1999), parts of Class 1- Philosophy. Psychology (Drăgoi, 1999), parts of Class 2 - Religion[4]. Theology (Achiri, 1999), parts of Class 5 i.e. the subclasses 57/59 for Biological Sciences (Popescu, 1999) converted into monolingual thesauri, and the whole of Class 8 - Linguistics and Literature (Frâncu, 1999a), converted into a multilingual one in Romanian, English and French, have already been published. As they were built independently of each other the real problems like coherence of the whole and overlapping or homonymous concepts were not given much attention. These problems will be overwhelming and hardly manageable when the time comes to merge all the contributing parts.

So far, some conclusions regarding aspects of convertibility from one information language to another – i.e. from the UDC to domain specific thesauri – can be formulated.

1. In most of the cases the working principle was the semantic factoring of the UDC text

e.g.

027.63	Libraries according to the age or sex of the users

was factored into:
LIBRARIES ACCORDING TO AGE + LIBRARY USERS
LIBRARIES ACCORDING TO SEX + LIBRARY USERS

[4] Class 2 Religion has been completely revised since this thesaurus was built. The revised tables are published in Extensions and Corrections to the UDC, Vol. 22 (2000), pp. 83-143

2. Many present day concepts are missing in some classes of the UDC such as Class 0 - Generalities and Class 1 - Philosophy which have not been revised and updated for quite a long time now. In order to overcome this shortcoming *combinations of notations have been added for new concepts that do not have a notation as such in the tables*. In so doing the purpose was that the relation with the classification structure is still preserved.

3. *The logical model for hierarchical arrangement of the schedule was kept* as far as possible. A well-known exception from the hierarchical structure of the tables is Class 2 where some very complicated solutions were adopted in order to maintain all concepts in a hierarchy (for example the range of notations encompassed in the extension 23/28 for Christianity). But there are large sections of the schedule with ready made hierarchies to be mapped onto a thesaurus structure such as those in Class 582 Systematic botany (Popescu, 1999) and a great deal of Class 8 Linguistics and Literature (Frâncu, 1999a).

e.g.

582.632	Fagales
582.632.1	Betulaceae. Birches. Alders. Hornbeam
582.632.2	Fagaceae. Beeches. Copper beech. Sweet chestnut. Oaks

The notations above and their captions are converted to a thesaurus structure as follows:

FAGALES
UDC: 582.632
NT : Betulaceae
Fagaceae

BETULACEAE
UDC : 582.632.1
BT : Fagales
NT : Betula sp.

BETULA SP.
UDC : 582.632.1
UF : Birches
Alders
Hornbeam

FAGACEAE
UDC : 582.632.2
BT : Fagales
NT : Fagus sp.

FAGUS SP.
UDC : 582.632.2
UF : Beeches
Copper beech
Sweet chestnut
Oaks

4. The vocabulary of the UDC tables was enriched with new terms frequently used in different disciplines.

e.g.

GENETIC ENGINEERING is a new term that is mapped to a synthesis '577.21.08' derived from two different numbers which do exist in the tables, i.e. 577.21 Molecular genetics + 57.08 Biological techniques. Therefore we will have in the thesaurus:

Genetic engineering
See: MOLECULAR GENETICS + BIOLOGICAL TECHNIQUES

According to the same rule now in the field of literature we will have:

Narrative art
See: ART OF WRITING + PROSE

Poetic art
See: ART OF WRITING + POETRY

One might argue that both ‘Narrative art’ and ‘Poetic art’ would rather be considered as descriptors. Apparently they are more currently used in this form than in the form suggested by the ‘Use’ reference. The reason why we take them as non-descriptors is that in both cases, each element of the combination of preferred terms has its corresponding UDC notation:

808.1	Art of writing
82-3	Prose
82-1	Poetry

5. *Plural forms* have been preferred in concrete entities expressed by countable nouns. Abstract nouns are given in the singular.

6. *Distinction was made between singular and plural forms* to denote species and forms. Possible ambiguity of terms is eliminated by scope notes and semantic relations (either hierarchical or associative)
   e.g.

   PRAYER
   UDC: 241.611
   SN : Used to denote the act of praying
   BT : Religious virtues
   RT : Prayers

   PRAYERS
   UDC: 243
   SN : Used for the text of prayers, books of prayers
   RT : Prayer

7. Scope notes can be taken from the UDC text in many occasions:
   e.g.

   CALCULUS OF PREDICATES
   UDC: 164.21
   SN : Used for the determination of a predicate according to the content

8. Upward posting for narrower concepts given in the UDC text after ‘including’:
   e.g.
   025.9 Administration of library buildings.
   Including: Maintenance. Cleaning. Removals

   Such a caption becomes in a thesaurus structure:

   ADMINISTRATION OF LIBRARY BUILDINGS
   UDC: 025.9
   BT : Administrative departments
   UF : Cleaning
   Maintenance
   Removals

3.4 Side effects

The concepts denoted by the UDC numbers may be expressed in a variety of forms ranging from perfect equivalence with the expression that concept has in the description, to partial equivalence and sometimes to hardly recognisable terms derived from the corresponding numbers. This speaks once again about the degree of compatibility between the information languages in question.

In order to make this statement more clear let us have a closer look at the way some parts of the classification system have responded to the requirements of conversion from class numbers to thesaurus terms. As a matter of fact the class which has proved as most suitable to this approach is Class 8 which was relatively recently revised and changed into a faceted structure. The hierarchies and the synonyms are given in the tables ensuring as a result the correct representation of the thesaurus semantic relations. Obviously, a faceted structure has great advantages for the thesaurus builder as far as the rules and guidelines are acknowledged and consistently followed.

Things will not go so easy and extensive difficulties will appear when dealing with non-faceted structures, lack of hierarchical configuration and more importantly, when the rules and guidelines are not entirely followed. Consistency is then highly damaged, the reciprocal relations (BTs, NTs and RTs) are not correctly available. This can result in confusion, misunderstandings on the use of terms on either side (indexer and searcher included) and hence in low effectiveness of the information language as a whole. Coordination of efforts and a unified set of methodological principles are extremely necessary in order to prevent expensive and time-consuming intellectual work with hardly acceptable results.

Going back to the thesauri based on the mentioned classes of the UDC tables one general conclusion can be formulated: the degree of equivalence between the two languages is descending from perfect equivalence, and that is mostly the case of Class 8 whose structure is highly convertible to a thesaurus and in a fairly good part of Classes 57/59, to partial equivalence in most of the broader concepts/terms corresponding to the subdivisions of Class 0, Class 1 and Class 2 and no equivalence at all as it was the case of extensive parts of Classes 1 and 2.

The main controversial remark about the configuration of Class 1 – Philosophy with serious consequences on the establishment of concordances between the UDC numbers and thesaurus terms is the lack of assigned numbers for philosophical systems (Class 14) from a historical perspective. A scope note was made for this particular class number that it should be used as a basis for all the philosophical systems that do not have an assigned number in the tables.

The same remark was made on the ethical doctrines (Class 17.03) that are just partly mentioned in the tables as well as on some other sections of Class 17. In all these instances the missing concepts have been added to the thesaurus structure, the corresponding UDC notations being either repeated for each separate concept (such as the terms defining philosophical systems) or modified by means of coordination devices like extension – graphically marked by stroke - and relation or connection between two separate numbers – graphically shown by colon. ‘Concessions’ of this sort of are nothing but restrictions imposed by the major requirement of keeping the structure of the UDC in permanent correspondence with the thesaurus terms. The first consequence of the restrictive way some parts of the schedules are translated into descriptors is that for certain concepts there will be no one-to-one correspondence between the UDC numbers as they are in the tables and their representation in words.

Some other parts of the classes under investigation are too specific in content (i.e. the subdivisions denoting details on ‘Prosody’ and on ‘Lexicography’ in Class 8). The device used in this kind of situation is the 'upward posting' of the very detailed terms to a broader one so that each of these concepts are represented in the thesaurus as entry terms and crossed referenced to the preferred one.

It is generally accepted that the main three conditions for compatibility among information languages are the coverage of the field of knowledge, the level of specificity and the level of pre-coordination (see p. 39). Concordance tables between vocabulary elements belonging to different indexing languages can to a large extent be made via computer-aided procedures. Despite the effective help these procedures can provide, dissemination between specific lexical units like homographs and polysemic words need human work.

The major problems with integrating or merging information languages refer to the amount of overlapping indexing terms and formulas and the manner used to distinguish among them. Distinction between homographs and polysemic words on one hand, the submission of a set of synonyms referring to the same concept on another hand and the identification of conceptually-related documents on the basis of associative or hierarchical relationships are strong devices to avoid situations when the search result will contain a large number of irrelevant documents compared with the search query.

The meaning of a natural language term is only understood within the context of the language game and the form of life with which it is associated (Blair, 1990). And the aforementioned examples of Schmitz-Esser - 'initiation' and 'positive' - are significant in this respect (see p. 15). The full meaning of an indexing language term can only be understood within the context of the conceptual relationships inherent within that indexing language. These relations will not be self evident unless explicitly shown in the indexing language structure. An important component of the effectiveness of such a tool for the information retrieval depends on the searcher also. Therefore the searcher's familiarity with a particular domain will have an impact on the ability to use the indexing language effectively (Jacob & Priss, 1999, 95).

At the moment when the decision is made to go from an information language based on numerical codes to one based on words from a natural language two requirements should necessarily be kept in mind:

1. The potential user, i.e. the public that particular language is designed for;
2. The richness of meanings of natural language words.

Indexing means information processing, in other words the knowledge included in the documents is interpreted and represented by the indexer through the documentary language used. The thinking of the author has a meaning that is first interpreted by the indexer, expressed in pieces of information via indexing terms (descriptors) and then again interpreted by the searcher. This communication process is approached differently at its two ends although it has always a language as its vehicle.

Therefore what the searcher or user gets in the end of the information retrieval procedure is a secondary picture of the contents of the documents retrieved. At the same time, the user will find more than one related documents grouped together based on the chosen characteristics. According to Chan (1994, 260) subject is the predominant characteristic for grouping all the works of a kind together. This is what the author considers to be a fundamental assumption in the classification theory. Before her, other theorists expressed more or less the same idea that in the process of classification 'the like things are put together' (Richardson, 1935, 1, Buchanan, 1979, 9). Since we deal with language and definitely consider the information retrieval process as a communication process (see **§1.1**) the representation of the subject matter of a document can factually reflect a false interpretation on the indexer's side. The same can be true for the user that may have a different image of the way the subject of a document he knows is represented by the indexer. There are several contributing factors at work in the act of organising knowledge and among them some of the most important are: knowledge of the classification system, knowledge of the discipline/domain the documents belong to, knowledge of the language that particular domain is using and last but not less important, the cultural and social context. Mai (2000, 26) makes a thorough analysis of the concept of likeness and of the way it is applied in classification theory his conclusion being is that the determination of the subject of documents is so fundamentally interpretative that the same document can be classified in different ways by two different classifiers.

As earlier stated, the natural language words are so rich in meanings that they can produce misinterpretations. When words from natural languages are used freely (i.e. without control) in information retrieval search failures may occur like the noise of too many irrelevant documents. The retrieval power of a controlled indexing language resides in its capability to disambiguate the meanings of terms that can be misleading. It also consist in its capacity to make connections between the word(s) likely to be used by the searcher in information retrieval and the term(s) used in the indexing language itself to define the same concept(s).

Polysemy and homonymy need a careful attention and for this reason the clear and discrete definition of the subject field is necessary to work as a filter. Such a filter is meant to screen the ambiguous terms. This desideratum can be achieved by either of these devices:

1. a bracket qualifier, e.g. "Acoustics' (Physics) and 'Acoustics' (Phonetics);
2. a scope note (see the 'Sonnet' example at page 13);
3. an additional term ('Religious symbolism' see 'SYMBOLISM' + 'RELIGION').

The relations between synonyms and near-synonyms and their corresponding preferential terms are made through reciprocating references of 'USE' and 'UF' type.

In a monolingual thesaurus these relations and disambiguation problems can relatively easy be overcome. In a multilingual thesaurus the relations between such terms have to be considered separately within each language and between languages. And that is so because not all the terms in a language are 'universals' and therefore some of them might need a special treatment for reasons of asymmetrical polysemy or asymmetrical homonymy. Quite often there can be words that define some particular entities, activities, properties or attributes in a certain language, and they have no corresponding terms in another language (e.g. the words 'dor'[1] and 'doina'[2] in Romanian and the word 'fado'[3] in Portuguese). These kind of words provide good examples of language barriers that can hardly be overcome because natural languages are culturally biased and thus circumscribed to different realities (Hudon, 1997). According to Schmitz-Esser (1999) the exact definitions of such terms are needed no matter how many words may be necessary for this.

In order to illustrate what we understand by asymmetrical polysemy let us consider the English word *bank (Figure 13)*. Even though all three languages share at least one meaning of this word, i.e. *credit institution,* there is a particular meaning that only exists in English, i.e. *riverside*. Hence the lack of symmetry in the equivalents of this polysemic term with direct consequences on the principle of equal treatment of the participating languages (Hudon, 1997).

[1] dor = Romanian word to express a feeling of longing and desire for the loved one

[2] doina = Romanian word to denote a lyrical poem specific to the Romanian folklore and expressing feelings of longing, love, revolt or lamenting, often accompanied by a melody according to its content

[3] fado = typical Portuguese song about love and death, death from the loss of love, destiny and almost always the singers – dressed in black – are talking about *saudade* (longing)

Romanian	English	French
BANCĂ	*Bank*	BANQUE
	use: CREDIT INSTITUTION	
	RIVERSIDE	
	CREDIT INSTITUTION	
	UF: *Bank*	
	RIVERSIDE	
	UF: *Bank*	

Figure 13. Example of asymmetrical polysemy in English

In addition to this, some more examples of different meanings of the word *bank* as they appear in the online catalogue of The Central University Library of Bucharest will make much clearer the impact of polysemy on information retrieval. For this purpose we formulated the query by using the words '*bank'* and '*banks'* as title words. The search result was as follows:

BANK

1. Instructor's manual with test *bank* to accompany computers and data processing : concepts and applications : with BASIC / Steven L. Mandell. - 3rd ed. - St. Paul [etc.] : West Publishing, 1985. - VIII, 456, B-200p.

2. Battle of Jutland *bank*. Russian offensive. Kut-El-Amara East Africa. Verdun. The great summer drive. United States and belligerents. Summary of two years' war. - 1916. - 509, [4] p. : h. (The story of the Great War; Vol. 5)

3. Text *bank* / Douglas W. Copeland. - Boston ; Dallas ; Geneva [etc.] : Houghton Mifflin, 1986. - VIII, 654p. : fig. ; 28cm. - Se utilizeaza împreuna cu: Economics / McKenzie. - ISBN 0-395-35531-1

4. Women of the left *bank* : Paris, 1900-1940 / Shari Benstock. - London : Virago Press, 1987. - IX, 531p., [16]f. : fotogr. ; 20cm. – Contine bibliogr. si index. - ISBN 0-86068-925-5

BANKS

5. Multidimensional filter *banks* and wavelets : research developments and applications / edited by Sankar Basu and Bernard Levy. - Boston [etc.] : Kluwer Academic, cop. 1997. - 238 p. : fotogr., tab., graf. ; 25 cm. - Contine bibliogr. - ISBN 0-7923-9848-3 . - (Multidimensional systems and signal processing; Vol. 8, Nos. 1/2)

Consider some other examples, from French and Romanian this time, given below and meant to illustrate the same aspect of asymmetrical polysemy *(Figure 14)*. The devices used in these examples to disseminate terminological ambiguities are a combination of terms in the first case – Matière colorante – and a bracket qualifier in the second case – *Broască (Lăcătuserie)* and *Broască (Zoologie).* Post-coordination of terms cannot be avoidable in this kind of instances:

Romanian	English	French
PICTURĂ	PAINTING	*Peinture*
		Use:TABLEAU
		MATIERE COLORANTE
		TABLEAU
		UF: *Peinture*
		MATIERE COLORANTE
		UF: *Peinture*
Broască	FROG	GRENOUILLE
Use:BROASCĂ (ZOOLOGIE)		
BROASCĂ (LĂCĂTUSERIE)		
BROASCĂ (ZOOLOGIE)		
UF: *Broască*		
BROASCĂ (LĂCĂTUSERIE)		
UF: *Broască*		

Figure 14. Examples of asymmetrical polysemy in French and Romanian

Homonymy or rather homography, as we deal with only written form of language, is subject of controversy among information scientists. And indeed the control of homographs deserves much of the attention of indexing language designers. In his book *Vocabulary control for information retrieval* Lancaster (1986, 7) argues in favour of vocabulary control, insisting on its necessity:

> "To promote the consistent representation of the subject matter… the control (merging) of synonymous and nearly synonymous expressions …distinguishing among homographs… to facilitate the conduct of a comprehensive search on some topic by linking together terms whose meanings are related…".

Lack of vocabulary control would scatter words of related meanings throughout the alphabetic list of subjects (p. 6), with immediate consequence on information loss. Likewise, identically spelled words will bring together documents with different subjects generating noise in the search result.

A slightly different opinion of the same author is presented in a later paragraph of the same book where we are told about the adequate search strategy working as compensation "for the lack of vocabulary control at input" (p. 162). Lancaster considers "the homograph problem" as "most trivial; it is more theoretical than actual. Homographs are usually only ambiguous when they stand alone. In information retrieval, however, one rarely uses words standing alone".

The example of the word '*seals'* Lancaster uses in order to make a sound argumentation for his statement is only relevant in case of specialised databases. The more restricted the subject/domain of a database, the less the probability of ambiguity of terms. He is perfectly right when he says that the context of the database considerably reduce the rate of ambiguity: the term "seals" would refer to an aquatic animal in a biology specialised database and to devices to close containers in an applied mechanics database. If both meanings occur in the same database, Lancaster argues and we agree, "possible ambiguity is reduced, if not eliminated entirely, though the context provided by the search strategy" (p. 162).

For all that, the success of a comprehensive search on a certain subject in an encyclopaedic database should be granted by some devices meant to guide the end user in performing it. For that purpose, the vocabulary builders should decide on using one or more of these devices:

- an additional term working as a qualifier (and thus increasing the precoordination) in order to express the context;
- an additional term used postcoordinately according to a search strategy the way Lancaster suggests;
- a scope note indicating different uses of the same term;
- a help message appearing on the screen at the search time suggesting the user to disambiguate the meaning of 'critical' terms (but this falls outside the responsibilities of vocabulary builders).

The conclusion of Lancaster's argumentation and his opinion on "the future of vocabulary control" read:

"It seems certain that natural language will become the norm in information retrieval and that use of conventional controlled vocabularies will decline. There are numerous reasons for this, including the escalating costs of human intellectual processing, the rapidly declining costs of computer storage, the increasing amount of text becoming accessible in machine-readable form" (p. 173).

3.5 Conclusions

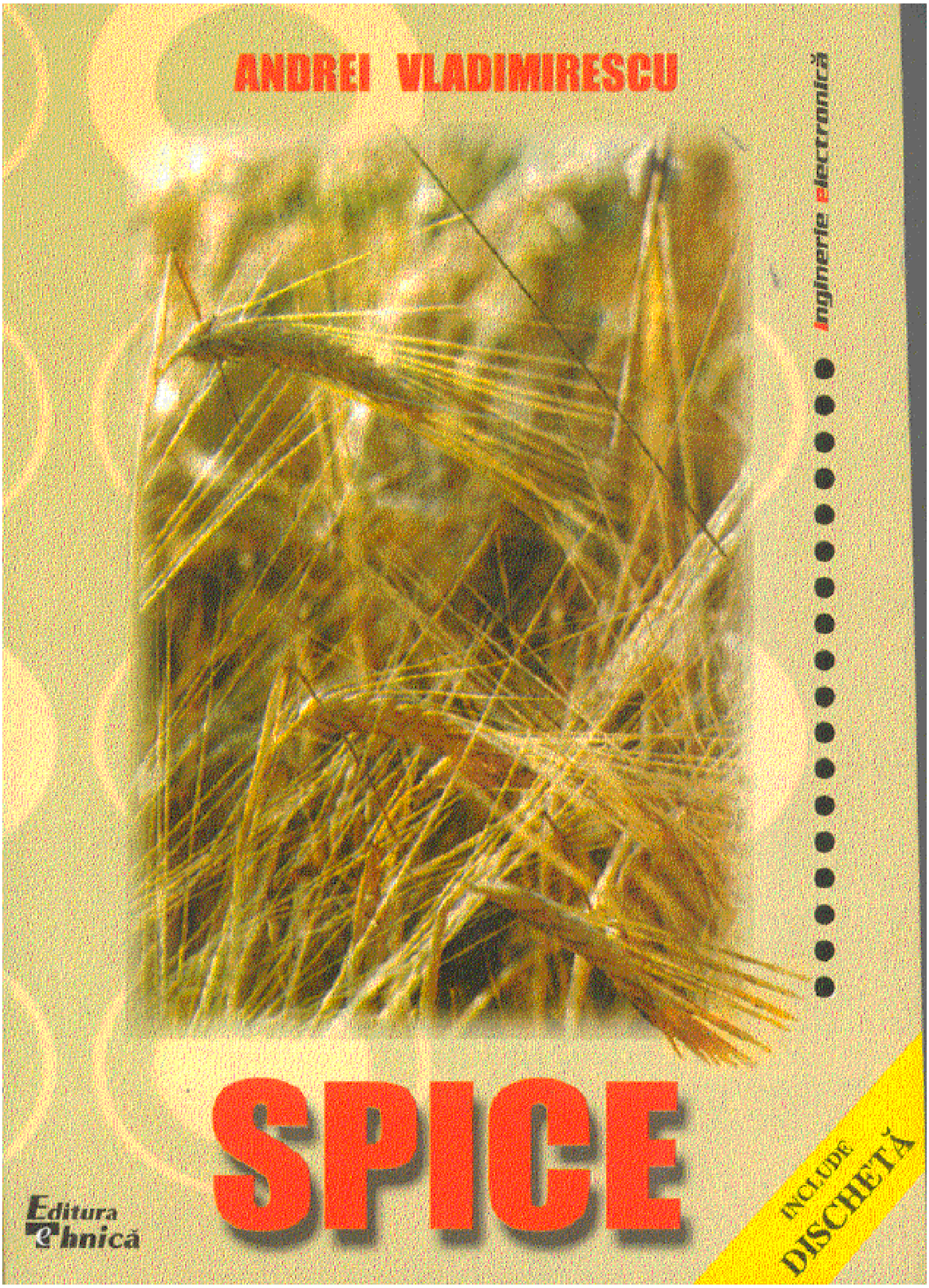
Even though each of the information languages has its own syntagmatic and paradigmatic structure it has been noticed that quite many of them are compatible with each other to a certain degree. It has to be so once they refer to the same reality and their functionalities are intended for the same purposes: organizing knowledge and therefore enable information retrieval. Various theoreticians formulated *theories of compatibility* of information languages thus opening wide possibilities for their application and generating a growing interest in the field. Some of these theories are mentioned in this chapter with the intention of applying them in our research.

Compatibility issues are argued in terms of structural and semantic particularities of information languages and it is on this basis that *full compatibility* and *partial compatibility* are considered. As long as full compatibility cannot be accomplished and the meaning and coherence of indexing are not affected, a compromise is made towards the *complementarity of the information languages*.

These basic theoretic outlines being given some practical applications are mentioned in order to show how compatibility can work towards a better retrievability and integration of information sources. Among them a famous one is the Unified Medical Language System (UMLS) that is briefly described here.

A larger space is given to *an application of compatibility and integration issues* in building of five domain specific thesauri based on some of the classes of the UDC. One of these – for Class 8. Linguistics. Literature – is multilingual. What makes this section important for the further development of our research is that here some individual characteristics of some classes are described and side effects and drawbacks of the working methodology are given account for. Additionally, the issues of ambiguity and disambiguation methods are dedicated a special attention.

Last, but not less important, *the problem of vocabulary control* as Lancaster sees it is put forward. The problem of ambiguity is more critical in case of encyclopaedic databases and in such a case it is for context to take over the disambiguation task. Mention should be made on the likeness of the disambiguating devices proposed by Lancaster and those applied in the UDC-based thesaurus that we introduce in the ongoing chapter.



Cover of the Romanian version of "The SPICE Book" as an example of misleading title

CHAPTER 4
CURRENT TRENDS IN MULTILINGUAL ACCESS

According to McIlwaine and Williamson (1999) thesauri are navigation tools for information retrieval used interactively in information systems. They are considered as components of expert systems, more and more likely to be tied to retrieval rather than to indexing. The authors reckon with 2 types of thesauri:

1. The "organiser" type, used for the systematic arrangement of indexes of any kind;
2. The retrieval aid which permits the searcher use his/her own words and connect to the term used by the database being searched in the manner of the original Roget's thesaurus.

As far as the multilingual thesauri are concerned the need is stressed for further research and revision of the existing standards, which, according to the above-cited authors, have seen no major changes since the early 80's.

4.1 CoBRA+ working group on Multilingual Subject Access (MACS)

From Newsletter No. 18 of December 1998 of IFLA's Division on Classification and Indexing (IFLA, 1998) we find out about a project between four European national libraries started with the aim to establish links between the different national authority files for subject headings. The project was initially called MUSE (MUltilingual Subject Entry), but this name was given up because some other projects were called the same. Therefore the name was changed into MACS (**M**ultilingual **AC**cess to **S**ubjects). This is a distributed project initiated by the CoBRA+ working group (**Co**mputerised **B**ibliographic **R**ecord **A**ctions) and its objective is the development of a prototype showing multilingual subject access to library catalogues by linking several existing subject heading languages (SHLs) i.e. RAMEAU, SWD/RSWK and LCSH (Clavel-Merrin, 1999). The initiative resulted from the efforts of the Schweizerische Landesbibliothek Bern with its collections in French as well as in German - for French publications the French subject authority file RAMEAU is used for indexing purposes in the Bibliographie de la France, whereas German publications are indexed with subject headings taken from the German Schlagwortnormdatei (SWD) in the Deutsche Nationalbibliographie.

Mention is made of the interest taken by libraries in both German and French speaking countries in the Library of Congress Subject Headings (LCSH) which are not used only for the indexing of American literature, but also by the British Library for the British National Bibliography and by many other national bibliographies throughout the world (see the example of the Finnish national bibliography in the forthcoming).

The working group was established with two members each from the British Library, the Schweizerische Landesbibliothek, the Bibliothèque nationale de France and the Deutsche Bibliothek.

The project is not intended to translate terms from one language to another or establish a homogenous thesaurus with terms in different languages. The aim is the effective linkage between already existing national subject heading lists, which remain in their own linguistic surroundings, but offer the opportunity of switching to the user of an OPAC or a national bibliographic database. Without even noticing the changing language access, the user should be able to extend the search for literature to publications in other databases. By the end of 1998, the pilot project was to be finished. The experiment was limited initially to a trilingual list of subject headings in the fields of sports and theatre plus a certain amount of very frequently used descriptors. Additionally the working group selected a number of publications by internationally well-known publishers, which were indexed in all three systems aiming to compare the quality and congruence of the subject heading strings used for the same title in LCSH, RAMEAU, and SWD.

Two different and rather specific fields of knowledge were selected to extract all relevant subject headings from the three authority files (LCSH, RAMEAU and SWD) and to establish links between these fields: *sports* and *theatre*. Though the systematic approaches especially between LCSH and SWD are rather far from each other (broad pre-coordinated subject headings in LCSH vs. specific subject headings with post-coordinated strings in SWD) it was possible in about 70 % of about 300 subject headings in each subject and authority file to establish links between the systems.

The mentioned newsletter states that for the growing international relations a multilingual thesaurus would also be very desirable, but professional advice is needed when establishing the links between very specific and not easily to translate descriptors in the different subject heading systems. Sometimes even terms which seem to be identical turn out to be completely different - "*Schwarzes Theater*" in the German authority file SWD as something absolutely different from "*Black theater*" in LCSH and "*Théâtre noir*" in RAMEAU. The working group members are aware of the difficulties in detail as well as in the project in general (especially as far as data processing is concerned). In fields of knowledge which are closely related to the cultural, historical, linguistic and administrative specifics of a country, e.g. in the humanities and partly in the social sciences (in the first step represented by theatre) it is likely to be less equivalent relations between the three systems than in science and technology or a field like sports. Yet, the first results were considered encouraging enough to make the working group proceed with the work.

Considering the three approaches to multilingual thesaurus construction recommended by ISO 5964 the members of the working group decided that the national libraries contributing to the project should investigate ways to offer multilingual access to their collections without having to abandon or translate their own subject headings. Therefore, they initiated a feasibility study on how to offer multilingual access using three different SHLs, by establishing links between headings in each language. A similar approach had been adopted in other two major terminology projects, the creation of equivalents between the Art and Architecture Thesaurus (AAT) and other controlled vocabularies in the same field, and in the Unified Medical Language System (UMLS). The approach adopted by the group does not entirely follow the guidelines in ISO 5964, but the solutions proposed by the standard were used partly in the establishment of the group's linking methodology.

The conclusion of Patrice Landry (2000), one of the members of the MACS working group, is that the analysis of the bibliographic records that contained LCSH, RAMEAU and SWD/RSWK indexing gives a favourable impression of convergence among the three indexing systems. At the authority level, there were a fairly high amount of headings found in the other indexing. These varied between 30% (SWD - RAMEAU) to 52% (LCSH - SWD). This is an indication that each of the subject lists contained headings that expressed very similarly the same concepts. At the indexing level (heading strings), it was also interesting to see the same variation of number of concordances. When the results of perfect and partial indexing concordances were added, the group found that there were between 29% to 55%

concordances at that level. As already indicated, these concordances appeared predominately between heading strings composed of two to three headings.

The attempt of comparing the three systems proved to be a difficult exercise. The methodology of using a trilingual list to find recurrences of headings was fairly efficient in judging LCSH and RAMEAU indexing practices. Since both systems use a fairly similar application principle and develop headings using a similar syntactic practice, it was not too difficult to analyse the results. The concordances of SWD headings and RSWK indexing rules with the other subject heading lists were more difficult to evaluate. It is only with additional studies that the high level of concordances between LCSH and SWD will be confirmed.

As this study focused principally on how the indexing done in one system could be useful to an indexer working in another indexing system, it would appear that there could be indeed some benefits for indexers in having a multilingual access to bibliographic records. To be able to access bibliographical records with the assistance of a multilingual thesaurus would give the indexers a primary insight on the headings used in a different indexing language. This access could be of help in determining the subject content of a document in a language not so familiar to an indexer. These benefits, as well as the potential assistance to library users and researchers should be further examined in the next stages of the Multilingual Subject Access Project.

The project MACS is brought again to the public attention on the occasion of the 68th IFLA Council and General Conference in 2002. During a meeting of the Professional Group on Classification and Indexing, Martin Kunz from the Deutsche Bibliothek Frankfurt am Main presented his view on the recent developments in mono and multilingual environments discussing in detail the usefulness of such an approach as MACS (Kunz, 2002). Given the circumstances of such a project the author expresses his doubt regarding the efficiency of multilingual thesauri built according to the existing international standards taking into account the considerable expenses involved in building this kind of indexing tools. He characterizes this approach – the multilingual thesaurus approach – as classical and no longer justifiable while giving full credit to the Internet type of approach based on links between authority files.

Once the equivalences are intellectually (and not automatically) established between the controlled vocabularies involved in the project, the only thing that remains to be solved is the user's problem, in other words, the user interface. He – the user – will prefer to search in his native language or a language that he is familiar with. The link from the term of a thesaurus in his language to its equivalent in another one might be useful to him. To illustrate this the author gives an example of a German user interested in *cycling,* going from SWD and the titles in DDB to the Bibliothèque Nationale de France by means of MACS.

Kunz is arguing once more in favour of structured vocabularies and their certain qualities in terms of precision in searching while opposing them to non-structured dictionaries that allow browsing with higher recall rates but at the expense of lack of precision. And again he stresses on the importance of the coordination of terms, the equivalences being not so problematic in case of comparable degrees of pre and/or post-coordination of terms of the controlled languages involved.

There are some future prospects connected with MACS. The possibility of creating a multilingual thesaurus to serve the purposes of MACS is excluded. The ultimate goal is the inclusion of all subject areas in it. According to Kunz, "the next logical step […] would be to open it up to other languages".

The real problem with the project MACS is that progress is very slowly made. As long as automatically established equivalences between terms are not admitted as alternative (or additional) procedure, the slowdown in accomplishing this stage in the development of the project, i.e. establishing equivalences between different vocabularies, will go beyond control.

4.2 EXPO 2000

A thoroughly new thinking in the way of thesauri as information retrieval aids is given by Schmitz-Esser (1998). His model was intended for a Visitors Information System at a world exhibition, i.e. EXPO 2000 in Hanover. The project was designed to bridge what the author calls "this world of real and virtual objects with the conceptual spaces encountered in the heads of some 40 million expected visitors". Schmitz-Esser argues that in order to achieve this goal two requirements are necessary:

- The content of the show should be conveyed in four official languages: English, French, Spanish, German
- The system interface should be very simple enabling quick and efficient access of any visitor.

The conceptual instrument is a four-language thesaurus with search texts consisting in English language thesaurus descriptors. Therefore English acts as an intermediate language while the other three languages are formally considered as target languages. Each search text represents a single topic of the EXPO 2000 put in a format that guarantees noise-free search results on all foreseeable questions, which might come from a visitor.

The thesaurus is largely based on the outline for a lexicographic, multilingual, computerised, universal thesaurus for linguistic engineering and information retrieval proposed in 1993 by the German Committee on Thesaurus and Classification Research in Jena, Germany.

The model presented is mainly centred on the often-controversial problem of relations in a thesaurus. Two classes of relationships are considered within the thesaurus terms:

Class 1 – referring to synonyms and polysemes as unique phenomena to each individual language;

Class 2 – referring to relations proper which are valid to all four languages, i.e. five types of relations: abstract/generic, partitive, beneficial, detrimental, and geographically partitive.

One of the innovations this thesaurus model brings to our knowledge is the missing of what the traditional thesauri labelled as Related Term (RT). This particularity emerged, according to the author, from the need for more refined and better-defined relations, which might be used not only in information retrieval but also in applied linguistics.

There is a serious problem though, and that is the problem of consistency, which cannot be guaranteed by this system. Yet, the author argues that consistency is not that necessary, as long as the thesaurus works well with the user.

Fugmann (1999) also considers consistency as an inadequate criterion of indexing quality. As far as the search file of an information system contains predictable expressions no matter how many they are for the same concept, the retrieval is made much easier. The alternative search statements can enhance the search result indeed, we should add, provided that all such terms are clustered appropriately in authority files along with clarifying syndetic structures of 'see' or 'use' references between them.

Let us now have a brief look at Schmitz-Esser's new approach to thesaurus design and construction (Schmitz-Esser, 1999). The author estimates that the new outlook he is introducing could have significant implications for change in the way multilingual thesauri are handled and integrated with each other. He proposes a new structure for a "machine-readable, linguistic, plurilingual, lexicographic, universal and domain-independent thesaurus". This model should prove its usefulness in classic areas of linguistic engineering such as machine-aided translation, abstracting and information retrieval. One of the consequences of its application would be the opening up of new frontiers towards new more animated, surprising, fun and playful approaches to information.

The two classes of relationships are reconsidered and redefined as Class I relations, or concept-term relations and Class II relations or concept-concept relations or Concept Relations Proper (CRP). The relations in first category are only valid for each language separately so they must be stipulated language by language. In a multilingual thesaurus all concepts can be made subject to the stipulation of conceptual interrelations according to one and the same set of different types of relations. They must be language universals, well defined and free from overlap or intersection.

An important issue working for the practical manageability of the thesaurus is the necessity of control on terms. Each term or sequence of terms must be unique in the system and clearly express the meaning of an object of thought and by no means any other object of thought in that language. The unique term is called Descriptor (D) and all other terms of equal meaning are called Additional Access Expressions (AAE). Descriptors together with the AAEs give the Access Expressions (AE). Looking comparatively at the configuration of this thesaurus model as against that of traditional ISO standards we have to admit that this view is more abstract and hence more comprehensive.

Inevitably the problem of the intermediate language is undertaken. While in the previously described model English was considered as intermediate or middle or source language working like a pivot for the other target languages, here the situation is changed. For reasons of political correctness and from a logical point of view, a numbering system was preferred and given the name of Meta Language Identification Number (MLIN). This number is to be assigned to each Equivalence Chain of Descriptors (ECD) and the Concept Relations Proper (CRP) should be formally applied among MLINs. A chain like: *airplane – avion – avion – Flugzeug* will be assigned a unique MLIN acting as an identifier in the system. All relationships to other ECDs, represented by their respective MLINs would be stipulated on the basis of their ECDs.

The multilingual thesaurus management software MTM3.1, the program that we used in building the thesauri for this study (see **§5.3** and **§5.5**) works the same way. Any relationship between two descriptors, be they broader, narrower or related terms, within a given language, will be stipulated similarly between equivalent descriptors in the other languages on the basis of the facet number, acting as a MLIN. Hence the possibilities for integrating multilingual thesauri provided that at least one of the languages of that particular thesaurus matches with the language pattern of the receiving multilingual thesaurus.

Schmitz-Esser (1999) argues that monolingual thesauri can be integrated with a multilingual thesaurus on condition that they follow the same format. The AAEs that have not been considered can be imported into the receiving thesaurus enriching in this way the total number of AAEs in that particular language. Expert vocabularies as well as taxonomies for animals or plants can also be integrated on condition of clear, unambiguous denomination of concepts. The latter category, i.e. taxonomies should in this case comply with the Abstract/Generic type of relationship.

Since the model imposes no restriction as to the length of the expression standing for Descriptor, it also recommends the use of the most specific expression taking into account the requirements for a multilingual set of descriptors. The access expressions may be either single terms (monems – one word with one semantic root, or synthems – one term with more than one semantic roots) or synthemised expressions. The one-to-many and many-to-one types of equivalence are included here, the example given being the French "pain cuit au four du bois" and its German equivalent "Holzofenbrot". Another example of this type of equivalence is the English word "siblings" which apparently has no one-word equivalent in other languages i.e. "frères et sœurs" in French, "fraţi şi surori" in Romanian "Brüder und Schwester" in German.

The semantic control in such a multilingual environment implies the necessity to clearly define both the meaning and the use of each individual expression. This is done through referring them to one or more classes of elements (columns) in the Basic Semantic Reference Structure (BSRS) (Schmitz-Esser, 1999) including a total number of 8 different types of use of Descriptors' Names *(Figure 15).*

What is it?	Who is it?	What an event is it?	Where is it?	Local extension	As seen from	When is it?	Extension in time
Concept	**Name**	**Event**	**Location 1**	**Location 2**	**Aspect**	**Time 1**	**Time 2**
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Tree planting			Peru		Technical aspect	1991	>2000
Canal profiles			France		Design	1600	1700
Portable heaters			World		Design	1960	>2000
Narrow gauge lines		Congo-Océan	Brazzaville	Pointe Noire	Construction	1921	1934
Civil wars		Spanish civil war	Spain		Process description	1936	1939

Figure 15. Application of the Basic Semantic Reference Structure as a general indexing scheme

The BSRS can serve as framework for various applications among which the author cites: a general reference tool, an explanatory instrument for instances, a general indexing scheme and a construction principle for encyclopaedias.

The undeniable originality in this paradigm of thesaurus construction is the 13 types of relation it proposes, namely: abstract/generic, partitive, part/whole, geographic-partitive, descendancy, instrumental, cause/effect, beneficial, detrimental, matter, form and appearance, process, state. For clarifying the boundaries between them, each type has a definition along with explanations and rules of application. These relations are determined by the type and by the direction given in the relational definition. Not all of these relations will be necessary in one or another of the applications of this thesaurus format, so each type is given a code number allowing classification of an individual type of thesaurus. Such coding would also facilitate data interchange in case of integration with other thesauri following the same format.

The table in *Figure 15* illustrates how an entry line in the BSRS can be used in indexing. Such an entry line represents a tuple e.g.

Narrow gauge lines (1), Congo-Océan (3), Brazzaville (4), Pointe Noire (5),
Construction (6), 1921 (7), 1934 (8)

This would read: construction of narrow gauge lines between Congo-Océan and Pointe Noire in the period of time 1921-1934. The model designers argue that a document would be represented in the database by a number of such tuples, which would enable search procedures to be more efficient than those used today. Additionally, equivalent entries from the authority files to bibliographic classification systems like UDC or DDC could be attached in a supplementary column to the BSRS. This could serve as a conceptual bridge between the proposed format and other ordering systems.

4.3. Multilingual access to information in Swiss libraries: the case of ETHICS

The problems of multilingual and multicharacter access to bibliographic data along with some solutions as practiced in libraries from Switzerland and Finland are described in a paper given by Geneviève Clavel-Merrin and Riitta Lehtinen (1995) in a meeting of the

Professional Group on Cataloguing at the 61st IFLA Council and General Conference. We shall have a look at the way problems of multilingual access to information are solved in Swiss libraries.

What makes Switzerland unique as a confederation of states is that it has four official languages (German, French, Italian and Romantsch) and a population of only 7 million people. English is added to these four languages being taught as an additional language in schools. The linguistic diversity is reflected by the printed production of the country: the German production goes up to 60% followed by the French 21% and the English 10%, while the Italian accounts for only 2%.

While discussing the meaning they give to the term "multilingual" the authors make the assumption that the search environment - like display and help screens and the user dialogue language - do not make a system multilingual. It is the access points that confer this quality to a system: authors, both personal names and corporate names and above all, subjects. Reference was made earlier, in a preceding chapter, to some of the controversial problems behind the appearing non-problematic linguistic aspects of personal authors and corporate authors. It is now the time to have a look at the way subject is dealt with in the system the authors describe.

Clavel-Merrin and Lehtinen (1995) argue that in an ideal system the indexers should be able to analyse documents and assign subject headings in their own language and users should be able to enter subject search terms in their own language, irrespective of the language of the document. In practice the number of languages available in such a system will be limited and again searchers will only be interested in literature whose language they have a good command of.

The languages considered for the author's purposes are German, French, Italian and English. These are actually the languages available for a user to formulate search requests in. The Swiss National Library (SNL) and VTLS Inc. agreed on a strategy capable to provide the necessary system structure in which a multilingual subject heading list be loaded and used as indexing and searching tool.

The first point discussed in the paper focuses on searching in the subject heading list. The options for the subject search language being displayed the user may have a choice of his own and follow the search strategy prompted by the system. The subject headings displayed in response to the user's search request will be in the chosen language. Languages cannot be changed in the middle of one search session, as they are not interfiled in the same list. If the search is started from a keyword or an author, the system will display the subject headings in a default language unless otherwise specified by the user in the beginning of the search. Boolean searching using subject terms is available in any of the languages. Keyword searches are possible from all the indexed fields in the database such as: author, title, subject and annotation.

It is convenient at this point to present some of the options of the SNL as they are described in the paper. A topic up for much debate for the system designers was the interfiling of the subject headings in all the participating languages. On the one hand this was regarded as simplifying the user's search procedure as it was no longer necessary for him to go for a search language. On the other hand this mixture of languages might just as well confuse the user who might have great difficulties with terms having different meanings in various different languages.

The examples chosen by the authors of the article to illustrate a mixed language subject sequence are taken from the ETHICS system. The ETHICS system (**E**idgenössischen **T**echnischen **H**ochschule **I**nformation **C**ontrol **S**ystem) is something that can hardly if at all be avoided when it comes to two topics: the multilingual access to subjects on one hand and the online applications of the Universal Decimal Classification on the other.

The first given example shows alphabetically displayed subject headings in English (E), French (F) and German (D) as they are interfiled in the ETHICS system. The dialogue language is French:

```
REGISTRE-MATIERES ALPHABETIQUE :                         LANGUE REG-MAT.: A
   1  ADAPTATION/BEWEGUNGSADAPTATION (ANATOMIE U.PHYSIOLOGIE)      D O,Q
   2  ADAPTATION/BOTANY                                            E O,Q
   3  ADAPTATION/BRIGHT TO DARK ADAPTAION (VISION)                 E O,Q
   4  ADAPTATION/CELULAIRE A L'ENVIRONMENT (CYTOLOGIE)             F O,Q
   5  ADAPTATION/CELULAR ADAPTATION                                E O,Q
   6  ADAPTATION/CLIMAT (ANATOMIE ET PHYSIOL.)                     F O,Q
   7  ADAPTATION/CLIMATIQUE ET EDAPHIQUE (PHYTOGENETIQUE)          F O
   8  ADAPTATION/COLORATION (ANIMAL ETHOLOGY)                      E O,Q
   9  ADAPTATION/CULTIVATED PLANTS                                 E O
  10  ADAPTATION/CULTURAL                                          E O
  11  ADAPTATION/DARK ADAPTATION (VISION)                          E O,Q
  12  ADAPTATION/DE L'AGRICULTURE                                  F O
  13  ADAPTATION/DUNKELADAPTATION (PHYSIOLOGISCHE OPTIK)           D O,Q
  14  ADAPTATION/ECOLOGIE VEGETALE                                 F,Q,E,U
  15  ADAPTATION/EVOLUTIONARY FACTORS (BIOLOGICAL EVOLUTION)       E O,Q,U
```

The same subject sequence is shown after this with alphabetically displayed subjects only in French:

```
CLE D'ACCES: ADAPTATION

REGISTRE-MATIERES ALPHABETIQUE :                        LANGUE REG-MAT.: F
 1    ADAPTATION (BIOLOGIE)                                         O,Q,U
 2    ADAPTATION (EVOLUTION BIOL.)                                  O,Q,U
 3    ADAPTATION/ANIMAUX TERRESTRES                                 O,Q,U
 4    ADAPTATION/AU TYPE D'EXPLOITATION (ECONOMIE D'ENTREPRISE)     O
 5    ADAPTATION/CELULAIRE A L'ENVIRONMENT (CYTOLOGIE)              O,Q
 6    ADAPTATION/CLIMAT (ANATOMIE ET PHYSIOL.)                      0,Q
 7    ADAPTATION/CLIMATIQUE ET EDAPHIQUE (PHYTOGENETIQUE)           O
 8    ADAPTATION/DE L'AGRICULTURE                                   O
 9    ADAPTATION/ECOLOGIE VEGETALE                                  O,Q,E,U
10    ADAPTATION/PHYSIOLOGIE GENERALE                               0,Q
11    ADAPTATION/PHYSIOLOGIE VEGETALE                               O,Q
12    ADAPTATION/PHYSIOLOGIQUE (ECOLOGIE ANIMALE)                   O,Q
13    ADAPTATION/QUADRIPOLES D' (TECHN.OSCILLAT.ELECTR.)            O,U
14    ADAPTATION/ZOOLOGIE                                           O,Q
15    ADAPTATION/CINEMATOGRAPHIQUES D'OEUVRES LITTERAIRES             Q
```

The equal treatment of languages in such a multilingual subject access system may suffer so it can happen that one of the languages has a more developed vocabulary and relational structure than the other. The user might in such a case want to see cross-references in other languages than the selected one. Considering this might lead to the user's confusion the SNL decided not to display cross references in a language other than the chosen one. For that reason the designers of the system should warn the user by help messages that choosing one or another of the languages may significantly reduce search results.

The technical aspects of the multilingual access to subjects make the substance of the second part of the mentioned paper. The technical support best fitting the SNL requirements was the MARC authority record extended by additional 1xx tags. The preferred form of a heading in each participating language is put in a 1xx field, the non-preferred forms in 4xx fields, each of them coded by language e.g.:

150	Bibliothèque, magasins	(fre)
150	Bibliotheksmagazine	(ger)
150	Library stacks	(eng)
450	Bibliothèque, rayonnage	(fre)
450	Rayonnage de bibliothèque	(fre)
450	Magazine, Bibliothek	(ger)
450	Library shelving	(eng)
550	Signature (Bibliothéconomie)	(fre)
550	Signatur (Bibliothekswessen)	(ger)

The authority records will display all the headings, scope notes and references in all languages but only for professionals and not for the end-users. For the user of the system as much as for import and export of authority records only the appropriate 1xx, 4xx and 5xx headings and scope notes according to the language code specified will be accessible.

Although it may be regarded with some amount of caution because of its not so user-friendly character the ETHICS system has to be given the attention it deserves by all means. Although it is no longer used we give here some opinions on the system as different authors formulated them at different times along with some details on the system's functionalities.

Marcella and Newton (1994, 233-236) describe it as a system that uses a separate file of verbal descriptors linked to the document file through UDC numbers. After a reasonably detailed presentation in their manual of classification the two authors come to the conclusion that the system supplies "no systematic testing to get user feedback on efficiency".

Buxton (1993, 114) considers ETHICS as "the best example of a sophisticated subject retrieval system based on the UDC".

Grünewald (1994) argues that the ETHICS starts from the UDC tables but not as they are in the original version. ETHICS has accommodated the ETH Library's own way of using the UDC having as target the usefulness of the classification structure for the library's purposes. About 35-40% of the class numbers in ETHICS are no longer found in the UDC tables. This may have bad consequences for the information transfer. In order to solve the problem a conversion table should exist between the original UDC and the class numbers used in ETHICS.

Basically the ETH Library in Zurich is using the conceptual and structural configuration of the UDC. Between 1960-1983 they currently used a Schlagwortkatalog and since 1987 they started the Online Katalog ETHICS. Some adjustments to the UDC numbers were made in order to fit their needs: they use separately the main numbers and the common auxiliaries in order to make conversion into verbal equivalents possible. There are two exceptions to this rule: the common auxiliaries of language and the common auxiliaries of time. The special auxiliaries are always combined with the main numbers as recommended by the UDC tables. They do not appear in the subject register separately.

The verbal equivalents including descriptors and their synonyms are used to expand the table of concepts and are given in three different languages: English, French and German. The subject register is displayed according to the dialogue language selected by the user.

The user's dialogue with the system can be initiated either by UDC numbers or by words or phrases. In the first case the systematic subject register will appear on screen. If the user types a UDC number on the subject search screen the systematic display from the subject register is shown (Buxton, 1993, 116):

Schlüssel: 656

SYSTEMATISCHES SACHREGISTER VERKEHRSWESEN

1. 656	VERKEHRSWESEN
2. 656%912	VERKEHRSKARTEN
3. 656*1	GESCHICHTE/ VERKEHR
verwende 9R%656	VERKEHRSKARTEN
4. 656”38”	FERIENVERKEHR + REISEVERKEHR
5. 656-61	VERKEHRSINGENIEUR

For the UDC number *538.9* the English terms in the subject register are:

Condensed matter physics
Matter / Condensed matter physics
Physics / Solid state physics
Solid state physics

as illustrated by Buxton (1993, 115).

In the second case, the result will be the alphabetical display of the subject register e.g.

Schlüssel: LOCOMOTIVES

ALPHABETISCHES SACHREGISTER

1. LOCOMOTIVES	E O, U
2. LOCOMOTIVES (VEHICULES SUR RAILS)	F O, U
3. LOCOMOTIVES A CREMALLIERE (VEHICULES SUR RAILS)	F O
4. LOCOMOTIVES A TURBINES A GAZ	F O
5. LOCOMOTIVES A VAPEUR	F O, Q, U

In both the above given examples the system responds with the number of records corresponding to the selected term. As a rule the display of the search results gives:

- the total number of titles;
- the number of titles including the search term as a single concept;
- the number of titles in which the search term is combined with other search terms i.e. concepts related by colon onto the basic UDC number.

The citation order of the descriptors is fixed by numbers mentioned under them. The example below is cited by Buxton (McIlwaine, 1993, 116) and it translates into “an annotated check list and selected bibliography of South African fungi for the period 1946-1977”:

1. Fungi. Eumycota (Mycologie)	582.28
2. Südafrika, Republik (Südl. Afrika)	(680)
3. Fachbibliographien	016
4.	“1946/1977”

1::2::3::4

Inasmuch as coordination of terms is concerned both precoordination (in the subject representation) and postcoordination (for the document’s form, language and time) are used. The system allows for the free combination of terms and the use of Boolean operators AND, OR, NOT.

About mono and polyhierarchies Grünewald (1994) argues that while the UDC tables provide references to related topics, ETHICS gives possibilities to polyhierarchies. On the average there are 60,000 UDC numbers and about 400,000 corresponding verbal equivalents. The UDC is working like a switching language. According to Grünewald there is no "see" or "see also" reference supplied because of the one-to-one concordance between the UDC numbers and the concepts they represent. Yet, the difference in numbers between the two categories of access (60,000 vs. 400,000) has to be covered somehow. Among the verbal terms assigned to each UDC main number or common auxiliary only one is known as descriptor and used in the systematic display. The system is very well equipped with synonyms.

For the availability of broader and narrower terms in the hierarchy the user has indications on each line of the alphabetical display: "O" stands for Oberbegriffe (higher terms) and "U" for Unterbegriffe (lower terms).

Described here more as an online application of the UDC, the ETHICS example is, despite its rather tedious handling, suggestive for a system allowing the link between a separate thesaurus file of descriptors and the UDC numbers in the bibliographic records. We shall see in the following some of the capabilities of the ETHICS as a system offering multilingual subject access.

4.4. Multilingual and multicharacter set data in library systems from Finland

While the first case presented in the previously cited paper deals with problems of multilingual access in libraries of Switzerland, the second case focuses on libraries in Finland as multilingual and multicharacter environment (Clavel-Merrin and Lehtinen, 1995).

The two official languages in Finland are Finnish and Swedish. Other languages taught in schools are English, the most commonly used language in international cooperation, French, German and Russian. The printed production of the country in 1993 was: 78% in Finnish, 5% in Swedish, 15% in English and 2% in other languages as the author reports.

The most important libraries in Finland are part of the LINNEA network and they all use the VTLS system. There is a union catalogue database called LINDA created from the local catalogues of each participating library. In addition to these there are other sources for the union catalogue like the Swedish National Catalogue and the Library of Congress CD-ROM.

Placing all these sources together in one big database in order to make the maximum benefit out of it means a great deal of "tuning" or harmonizing activities particularly in case of classical Greek or Latin author names and of the transliteration of Greek and other non-Latin characters. As for the subject headings those in the library's primary language are preferred and the Swedish or English headings are kept as such. Most of the libraries keep all the headings but they decide the tags for uncontrolled terms depending on whether they want them to be indexed or not. Subject headings are mostly in Finnish with some exceptions in Swedish and the medical subject headings in English from MeSH. The author's opinion is that although languages are interfiled in the subject index since they are so different they do not confuse the reader that much.

The authority files for Finnish personal and corporate authors are maintained by the Finnish University Library. For the corporate authors the authority records contain 'see also' references for other forms and official translations of the names in other languages like Swedish, English, French and German.

The Finnish general thesaurus of about 14,000 terms is maintained by the same institution. It contains 'see' references for non-preferred terms and 'see also' references for broader, narrower and parallel terms. In addition to that there are 20 specialised lists of subject headings on various domains like history, linguistics, education, sports and physical education, law, social sciences, library and information science; these are maintained by the specialised libraries themselves. For those disciplines/domains where no thesaurus or list of subject headings exists, uncontrolled terms are used.

The Finnish academic libraries use subject headings mainly from the 1980's. Until then almost all of them only used the UDC as subject description method. Nowadays they still use the UDC codes as an alternative indexing and search method this being the only search key to retrieve any of the records in the database.

The problem of multilingual subject access in Finnish libraries is tackled along with three possible solutions:

1. to load the Finnish and Swedish versions of the general Finnish thesaurus, add language codes and link the two forms; this model could also be applied to the other existing authority record files for subject headings;
2. to create one authority record per concept per language and link them together via separate link records;
3. to combine the UDC numbers with matching subject headings which are friendlier to the user.

Generated by the large amount of Russian literature in the Finnish library collections the problem of Cyrillic script in non-Cyrillic environment had to be coped with. In the old card catalogues of the Slavonic library in the Helsinki University Library the Russian books were catalogued using Cyrillic characters. With the advent of the OPACs these bibliographic records had to go through transliteration into Latin characters in order to be retrieved given the incapability of computers at that time to display Cyrillic characters.

It may happen that transliteration standards are different from one institution to another. A union catalogue is likely to copy records from different sources. For that purpose a clear statement on which is the international standard used in transliteration of Cyrillic characters must be made. For different transliteration standards, as the author gives in here ISO 9 and ISO R9, a sort of conversion table is necessary to make mappings from one standard to another possible e.g.:

Cyrillic	ISO 9	ISO R9 (National form)
Я	â	ja
Ю	û	ju
щ	ŝ	šč, štš

An authority record is provided to illustrate the different transliterated forms of a Russian author name and the way they are entered in the MARC record fields:

```
100   Чехов $h Ahtoh
400   Chekhov $h Anton
400   Chekhov $h A. P.
400   Tjechov $h Anton
400   Tchekhov $h Anton
500   Čehov $h Anton Pavlovič
500   Tšehov $h Anton
```

It makes a difference whether the terminal permits the use of Cyrillic characters or not. The example above gives the preferred form in Cyrillic but when a non-Cyrillic terminal type is used two entries will give the same form of the heading which may result in confusion. Therefore the Cyrillic form was decided to be marked in the index in order to prevent confusion. A search for a title beginning with “sovetskaâ” will retrieve the following when a non-Cyrillic terminal type is used:

```
1>      1 Sovetskaâ arhitektura
2.      1 Sovetskaâ arhitektura
3.      1 Sovetskaja justicija
4.      1 Sovetski entsiklopeditseski slovar
5.      1 Sovetskoe gosudartsvo I pravo
6.      1 Soviet, East European and Slavonic studies in Britain
7.      1 Soviet education
8.      1 Soviet geography
9.      1 Soviet law and government
```

When a Cyrillic terminal type is used the search result will differ a little for the Cyrillic characters in the first line:

```
1>      1 Советская архитектура
2.      1 Sovetskaâ arhitektura
3.      1 Sovetskaja justicija
4.      1 Sovetski entsiklopeditseski slovar
5.      1 Sovetskoe gosudartsvo I pravo
6.      1 Soviet, East European and Slavonic studies in Britain
7.      1 Soviet education
8.      1 Soviet geography
9.      1 Soviet law and government
```

Names of non-Russian origin in Russian documents can also give troubles. These names go through a very intricate process of transliteration that is done in two steps: first they are phonetically transcribed in Russian and then transliterated in Latin characters that do not respect the names’ original spelling in Latin script. When compared with the original spelling they look rather different, e.g.

Cyrillic	ISO 9	In original language
Щекпсир	Šekspir	Shakespeare
Джованоли	Džovan’oli	Giovannoli
Олдриж	Oldridz	Aldridge
Вулф	Vulf	Woolf

In such cases ‘see also’ references must be provided in order to assist searching. According to the Finnish author of the study the planned multilingual access will not be used for Cyrillic data given the multitude of problems generated by the different transliterations of the Cyrillic character set.

The general conclusion is that the problematic access to multilingual and multicharacter set databases can only be solved through co-operation and that in both cases presented the technical solution, while complex, is feasible although some data management questions are still to be solved.

4.5 Multilingual and multiscript subject access in Israel

The case of Israel as a multicultural, multilingual and multiscript environment is a perfect example of how Israeli libraries have to cope with more than one language and script in order to make information contained in bibliographic databases available to users. The problems at issue in the different approaches the libraries undertake to enable access to their materials in different languages and scripts were presented by Elhanan Adler (Adler, 2000) in a meeting of the same Professional Group on Classification and Indexing at the 66th IFLA Council and

General Conference. The author considers the use of subject headings and word searching, primarily in English, as prevalent trend in academic libraries. The public libraries though are just evolving from the classified catalogue.

The difficulties emerging from the already existing two official languages, Hebrew and Arabic, are added newer ones by the use of English, a third common yet unofficial language. Not only the script is different in these languages but the directionality also, Hebrew and Arabic being written from right to left. Additionally, the representation of only consonants in most texts written in the two official languages makes romanization a complicated matter. What is indeed unique in Israel is that Hebrew, a language used for about 2000 years only for writing and prayer, was reborn as a spoken language during the last century. Hence the complications of creating in a relatively short time a large amount of new terminology according to the necessities of modern life. Moreover, since the country encourages the Jewish immigration from all countries of the world, the mother tongue of these people bearing influences from the places they come is different from the official languages spoken in Israel.

After introducing all the characteristics of the Israeli multilingual and multiscript environment Adler goes further presenting the solutions the bibliographic community adopted in order to solve these critical issues. As far as the descriptive cataloguing is concerned, according to the Israeli cataloguing practice, separate catalogues are maintained for three scripts: Hebrew, Arabic and Latin. In addition to those, some of the libraries maintain also a fourth catalogue in Cyrillic (this being an easier script to romanize) and they do so given Russian is the mother tongue of a large amount of the Israeli public.

Furthermore, another Israeli author (Seymour, 2000) made a pertinent analysis of the Israeli cataloguing activities in the same context. He states that many Israeli libraries use four character sets: European, Hebrew, Arabic and Cyrillic. If the computer is not equipped to display Arabic or Cyrillic characters the Arabic is displayed in Hebrew transliteration and the Cyrillic is automatically romanized. The display problem is thus solved but the search problem is still pending. Searching has to be performed using either Arabic or Cyrillic characters. The system is bi-directional. Seymour cites Lazinger and Adler (1998, 183) about the provision of the system with a code at the beginning of each line by which the computer is informed about the dominant script.

Adler (2000) appreciates subject cataloguing as far more troublesome than descriptive cataloguing in such an environment and therefore it requires much more and careful attention. Traditionally, the Jewish National and University Library imposed their practice over the Israeli library community, the subject access being made via the Dewey Decimal Classification (DDC). The DDC codes were found to be an ideal solution for the subject approach to library catalogues particularly because of its language independence.

The University of Haifa Library made an important turn from this tradition by being the first to adopt the Library of Congress Classification for their open shelf collection. The next step from this point was their decision to use the Library of Congress Subject Headings (LCSH), the main reason for that being the possibility to use LC copy cataloguing and classification. The outcome of this switch was that several university libraries followed the policy of the University of Haifa Library on the account that people in the academic community would have good reading knowledge of English to enable them handle the scientific terminology used in subject headings.

Subsequently, the University of Haifa created a thesaurus of Hebrew indexing terms for a project started in 1977 called Index to Hebrew Periodicals. This thesaurus served as a basis for a list of Hebrew subject headings used in public libraries.

The Library of Bar-Ilan University has a rather intricate subject system. They created a subject heading list to be used on one hand with Hebrew language publications (and this include Judaic and Israeli topics for which Hebrew terminology is better qualified) and on the other with Latin character publications (for which the terminology is translated from the LCSH). This system has the disadvantage of split subject files difficult to handle when a given subject is in both Hebrew and Latin publications.

Recently, several university libraries, while maintaining their classified catalogue have tried to enrich subject access by textual retrieval elements. Consequently the UDC and DDC numbers were added English language cross-references working as additional subject access (searchable as headings and by word).

From the subject access point of view the Israeli public libraries are still confronted with shortcomings for little expectation of English literacy from their users. However dissatisfied with the existing classified catalogues their Hebrew subject headings recently received by copy-cataloguing and the effectiveness of their usage are still to be evaluated.

4.6 “Crossing the language barrier” by way of the Cross-Language Information Retrieval (CLIR) tracks in Text Retrieval Conferences (TREC)

The globalisation of the information society is challenging today’s information user with the difficult task of querying multilingual document collections. Irrespective of his knowledge of foreign languages, any user is likely to “feel at home” when using his mother tongue to approach the information needed. To put it in a different way, it can happen that a user is sure to find the required information in a library catalogue using a language out of his reach: he can speak English and French as second and third languages but the catalogue’s language is Dutch. He has a retrieval problem that can be solved by:

- searching via author and/or title if what he is searching for is a known item;
- searching via classification codes if the classification system is one he is familiar with;
- translating the query (one or more search terms) by means of a machine readable dictionary that includes Dutch.

The retrieval difficulties can remarkably grow in case of free text searching. It imposes so many restrictions on the search result and causes as much confusion as only synonymy and homonymy can generate. The side effects of using free text searching have been to a large extent described in the preceding chapters and we shall not insist on them now.

The Cross-Language Information Retrieval (CLIR) tracks of the Text Retrieval Conferences focus on retrieval when documents are written in a language different from the language of the queries. According to Voorhees and Harman (1998) TREC-6 used documents in English French and German and queries in English French, German, Spanish and Dutch.

Therefore, the main task of the CLIR tracks is to find solutions to the problem of matching the query and the documents across different languages. For this purpose research groups have to use queries written in single language in order to retrieve documents in many different languages. Over the following two CLIR tracks in TREC-7 (1998) and TREC-8 (1999), Italian has been added as document language to English, German and French, already used in their research.

In an overview of the CLIR track in TREC-6 Schäuble and Sheridan (1998) expose the possible applications for finding information written in a language other than the user’s native or preferred language:

- the user may want to find all possible relevant information in a multilingual text database, irrespective of the language of the relevant information (e.g. patent or legal information)

- the user may have some language comprehension ability in the language of the documents (passive vocabulary) but not enough active vocabulary to confidently specify queries in those languages; in this case cross-language search permits the user to specify native language queries and retrieve documents in their original language.

The major advantage of cross-language retrieval therefore is that it requires only one query addressed to a multilingual text collection rather than having the user submit individual queries in each of the languages of interest.

The participants in the experiments done in the CLIR tracks produce sets of queries from the topic statements given to them and run those queries against the documents. According to the proceedings CLIR research started with experiments using translation of the queries and documents into a controlled, language independent indexing vocabulary (e.g. WordNet synsets). Nowadays free text searching is most common and depending on the resources used to cross the language barrier the approaches to it can be: machine translation, machine readable dictionaries or corpus-based resources.

Machine translation was experimented in CLIR by a number of groups. The conclusion however was that machine translation alone does not solve a difficult problem in CLIR: queries entered by users into a retrieval system are hardly if ever complete sentences therefore context is not provided for disambiguation of meaning.

Corpus-based approaches are very much used in the CLIR tracks of TREC. Basically they are facilitated by the comparable nature of the Schweizerische Depeschenagentur (SDA) collections issued both in German and in French. SDA is the major source of document collection that consist in newswire stories edited by the Swiss agency for the German-speaking and the French-speaking parts of the country. The corpus of several tens of thousands of documents in the German and French SDA collections are assigned descriptors manually by SDA reporters (dates of the stories and common cognates in their texts). Search queries are formulated in one language to retrieve documents in the other. The results of these retrieval experiments are evaluated by standard ad-hoc TREC evaluation measures. Participating groups were free to experiment searches with various query length and using both automatic and manual procedures according to the main TREC ad-hoc task.

The sources for the document collections used in the CLIR tracks are:

▪	English	Associated Press (AP) news covering three years (1988 to 1990)
▪	German	Schweizerische Depeschenagentur (SDA) news from the same period
▪		Neuer Zürcher Zeitung (NZZ) articles from 1994
▪	French	Schweizerische Depeschenagentur (SDA) news from 1988 to 1990
▪	Italian	Schweizerische Depeschenagentur (SDA) news from 1989 to 1990

Unlike in TREC-6 where the topics were developed centrally at National Institute for Science and Technology (NIST), since it was proved difficult to produce topics in all languages in a single place, the topics for CLIR in TREC-7 and 8 were created on a distributed basis in four different sites, each located in an area where one of the topic languages is natively spoken, such as:

- English: National Institute for Science and Technology (NIST), Gaithersburg, MD, USA
- German: IZ Sozialwissenschaften, Germany
- French: EPFL Lausanne (TREC-7) and University of Zurich (TREC-8) Switzerland
- Italian: CNR, Pisa, Italy

From each site seven topics were chosen to be included in the topic set. The other 21 queries were translated. This led to a pool of 28 topics, each available in all four languages. Participants experimented these topics with both automatic and manual runs. English was the most used topic language followed by German. Every language was used by at least one group. As to the dictionary resources, while there were plenty of dictionaries of English and the other languages, a shortage of non-English language pair dictionaries was noticed (e.g. German to Italian).

Relevance assessments were produced for the evaluation of these runs at the same sites were the topics were created. On the average precision was evaluated to have improved as of the preceding CLIR track (TREC-6). The most used approaches were: statistical translation models, dictionary-based translation with fuzzy query expansion terms, query translation using bilingual dictionaries and online machine translation.

Topic translation raised the typical problem involved in any translation: there has to be a perfect understanding of the source in order to achieve a perfect understanding of the target. What is still controversial, according to the report, is how far the target version can deviate from the source in terms of style, vocabulary and authenticity in order to get an acceptable balance between precision as to the source and naturalness as to the target language.

At TREC-8 it was decided that cross-language system evaluation activities should move to Europe since the languages involved are traditionally considered as European and much of the work was done in Europe. In the year 2000, TREC offered a CLIR track using English and Mandarin documents and English topics.

More importantly, the range of issues was enlarged with the launching of an independent activity known as CLEF (Cross-Language Evaluation Forum) involving a greater variety of tasks such as: multilingual information retrieval, bilingual information retrieval and monolingual (non-English) retrieval. The main task in CLEF (2000) was searching for relevant documents in a multilingual document collection and listing the results in a merged ranked list. With the hope that additional languages would join in, problems of word order, morphology, diacritics and language variants were given special attention.

4.7 Conclusions

We have presented this chapter several systems that provide access to multilingual document collections. This was meant to give an idea of the state-of-the-art in the field of multilingual information retrieval. Some of them are just projects which are still going under evaluation procedures in order to be put into practice (like MACS), others have been abandoned for lack of financial resources meant to enable the information scientists carry on their research (EXPO 2000). While confronted with many impediments there are systems that have been working for quite a long period of time and that is the case of a multilingual country as Switzerland with the well known ETHICS system at the ETH Library in Zurich, despite some denigrating voices. Problems of multilingual and multicharacter subject access are given particular attention in Switzerland, Finland and Israel and the painstaking efforts made to overcome them are really encouraging. Finally, the CLIR tracks of the Text Retrieval Conferences take a firm step forward by applying language engineering procedures to multilingual information retrieval.

The trends and experiments presented in this chapter are each characterised by a functional principle that makes them work such as:

1. mapping of terms and heading strings via links between the participating subject heading systems (RAMEAU, SWD/RSWK and LCSH, respectively) in the case of the MACS project;

2. multilingual access having as source the structure of the UDC tables slightly adapted to best serve the purposes of the information retrieval system in ETHICS;
3. development of an ontology based on the ability of modern languages to express new and unknown concepts on the basis of acknowledged universals; one of the applications of the Basic Semantic Reference Structure developed for EXPO 2000 which is of our interest is that of a general indexing scheme able to accommodate a thesaurus format and facilitate data interchange;
4. controversial aspects of subject access in multilingual and multiscript environments and the efforts made to overcome them primarily by carefully designed and maintained authority records and in this respect we looked at the situation of multilingual countries like Switzerland, Finland and Israel;
5. the query is formulated in a particular language other than the language(s) of the document collection and is automatically translated into various languages within the cross-language information retrieval project in the TREC experiments.

The common aspect found in all the systems described is the requirement for cooperation and distributed work. This asks for further coordination activities but such experiments cannot work otherwise. Another necessary demand is that considering the complexity of the translation activities with respect to such aspects like authenticity and cultural differences, people with very good knowledge of the target language should do the translation work if native speakers are not available. Last but not least financial support is crucially important in such undertakings. As we saw in the last paragraph of this chapter it is essential to get more languages in such projects in order to provide equal opportunities to information access regardless of the language barriers.

CHAPTER 5
BUILDING UDC-BASED MULTILINGUAL THESAURI

5.1 Introductory notes on the UDC as an intermediate language

As aforementioned, the perspective of using the UDC as an intermediate language has been regarded with confidence by authors like McIlwaine and Williamson (1995, 1997), Riesthuis (1997), Scibor (1997), Frâncu (1996, 1999b), and earlier by Riesthuis and Bliedung (1990).

Formulated in simple words the *definition of an intermediate language* would be: an intermediate language L_X is an information language that permits information retrieval in a system using a given language L_A by means of subject notations (subject headings or classification codes) belonging to a different language L_B. The only condition necessary and sufficient is that the subjects represented in either of the two different languages, L_A and L_B respectively, should be reciprocally translatable in the intermediate, or switching language L_X. The number of contributing information languages can be expanded as long as there is a conversion table or a table of equivalence between each additional language(s) and the switching language L_X (*Figure 16*).

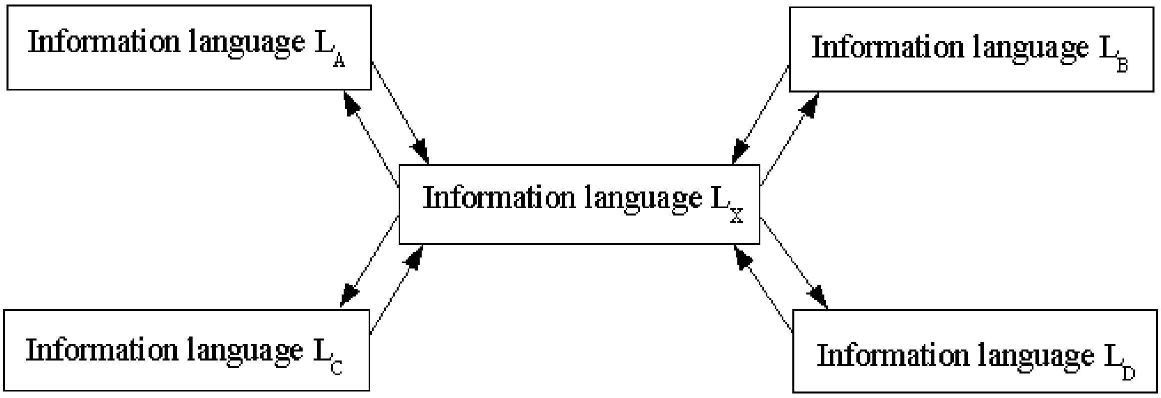

Information language L_A
Information language L_B
Information language L_X
Information language L_C
Information language L_D

Figure 16. Diagram showing how the switching language works

As they announce the goal of the restructuring of Class 61 – Medical sciences of the UDC, McIlwaine and Williamson (1995, 11) propose an update of the tables making them more open to online manipulation through

a. improving retrieval by identifying each concept uniquely;
b. providing the means to express complex concepts consistently throughout the classification;
c. providing the means to make future revision work easier.

The whole restructuring work is based on the facet framework supplied by Class H for Health Sciences, Biomedical Sciences from the Bliss Bibliographic Classification, second edition – BC2 (McIlwaine & Williamson, 1997, 39). Not only the notation is taken into account for revision but the ‘see’ and ‘see also’ references too. The final step intended by the authors is the creation of an index to the tables by deriving a thesaurus from the restructured configuration.

In very much the same spirit of adapting the UDC with a view to using it in information retrieval systems Scibor (1997, 201) reinforces the qualities of the tables recommending one of the following methods:

a) use of full (compound or complex) UDC numbers;
b) retrieval according to a subset (or subsets) of characters contained in a full UDC number;
c) use of verbal equivalents to UDC numbers.

Scibor stresses on the imprudent and not very well-founded dispose of the UDC in favour of subject headings alone in indexing activities with the arrival of computers in libraries and other institutions. Re-indexing work requires huge amounts of efforts and time and the productivity of such an attempt can only be disappointing. Keeping the UDC-based subject descriptions and assigning subject headings according to their meanings can effectively improve the information retrieval power of the two information languages taken as a whole. This statement fully justifies our attempt and undertaking described henceforward.

According to Scibor (1997, 202) the use of verbal equivalents in information retrieval can be done following two different procedures:

- first, documents are indexed at the same time by UDC numbers and by verbal equivalents added to them, those equivalents acting like a direct retrieval tool without translating them into UDC numbers and
- second, verbal equivalents (as much as their synonyms and quasi-synonyms) serve only as a kind of interface and are automatically translated by the computer program into UDC numbers used by the computer to search with without the user knowing it.

In a remarkable endeavour to fix the major stages of development in the history of the UDC as much as its role and future potential, Michele Santoro (1995) mentions the possibility of interaction between the UDC and the thesauri. He supports his statement by illustrating it with the case of the *EJC/TEST thesaurus of engineering* having exact semantic equivalents of the UDC codes in 80% of its descriptors.

Furthermore, Santoro (1996) refers to the UDC as an international exchange language or intermediate language given its potentiality to switch between different documentary languages.

And indeed the International Federation for Documentation (FID) repeatedly proposed that the UDC be adopted as the most suitable candidate for the role of switching language (Lloyd, 1972). A special interest group created for this purpose and a project was started to elaborate a *Standard Reference Code* (Lloyd, 1972) meant to coherently revise the structure of the UDC as switching language. The Standard Reference Code consisted of a synoptic map of knowledge condensed in 5000 subject fields. These were further organized in two parallel tables, one for disciplines and the other for entities, categories or facets that could contribute to the formulation of concepts.

Among the qualities of the UDC that give good reason for it assuming the role of a switching language, information scientists like Vickery (1961), Perreault (1969) and Dahlberg (1975) enumerate the following:

- disciplinary and terminological adequacy,
- possibility of being transformed into a totally faceted scheme,
- compatibility with thesauri and other documentary languages.

All these characteristics of the UDC will be largely employed in our case study.

One of the common conclusions of all these authors is that faceted structures are more likely to permit conversion from the UDC to a word system structure. This is essentially the trend in latest revisions of the UDC tables beginning with those in volume 14 of "Extensions and Corrections to the UDC"[5]. The actual version of Class 8 for Linguistics and Literature is definitely the most faceted class of the schedule so far. It is for this reason that a multilingual thesaurus in Romanian, English and French was created without difficulty, based on the structure of Class 8 of the UDC (Frâncu, 1999a) bringing evidence for the feasibility of such a conversion.

As repeatedly underlined the subsequent revision work of the UDC is going in the same direction. The restructuring of Class 61 for Medicine thoroughly demonstrates it. This is, in as much as the authors admit, just an intermediary step making possible the easier conversion from classificatory to thesaurus structure.

Why does a faceted structure permit easier conversion from bibliographic classification systems to thesauri? And why do they best fit the online information retrieval purposes?

To begin with, a bibliographic classification subdivides the world of knowledge into smaller parts (classes and subclasses) according to various principles of division in order to facilitate information retrieval. The name of a class characterises a whole unit of analysis, whatever its hierarchical level (Beghtol, 2000). The examples provided by Beghtol (p. 314) are taken from Botany, i.e. "Deciduous trees" describe a class having as superordinate 'Trees' and as subordinate 'Oak trees'. Each of these categories is considered in turn as a whole class. They are also capable of being subdivided into types and parts as taxonomic or patronomic subdivisions. Subdivisions are presented in hierarchical displays in classification systems and are usually thought to exhibit relations among concepts. In practice, Beghtol argues, we need to know whether a concept may occur in a taxonomy for one kind of thing (concept, event, etc.) or in a patronomy for another thing (concept, event, etc.). Failure to specify a clear distinction between "kind of" and "part of" facets can generate ambiguity. From the examples provided (Beghtol, 2000, 316) we select and consider the following:

Kinds of words	***Parts of sentences***
Adjectives	Adjectives
Nouns	Nouns
Verbs	Verbs

In these examples the same set of foci is applicable for both parts and kinds. It is the name of the facet that differentiate them providing what Lansing (1995) calls the "vantage" point from which the observations are made. Consequently, an accurate class name denoting the whole provides the needed context.

Going back to the previously asked questions we may argue that the synthesis present throughout the UDC Class 8, for instance, makes it better adaptable to the requirements of converting class numbers into vocabulary terms to be used postcoordinately in information retrieval. In so doing the main advantage is that facets like morphology (81'366), syntax (81'367), etc. can be considered as vantage points for whatever individual language that makes the subject of a document. Initial grouping of numbers e.g. 811.135.1 *Romanian language* or 811.111 *English Language* will effectively provide context for the linguistic aspects dealt with in the documents studied. Conversely, vocabulary elements of an indexing language based on such components of a complex UDC number can be combined without any difficulty in an online retrieval system in order to provide information access.

[5] This volume of Extensions and Corrections to the UDC contains the revision of Class 8 – Linguistics. Literature into a faceted structure

In the next chapter we will introduce our findings as far as they resulted from a research on harmonising a Romanian abridged UDC edition with an interdisciplinary multilingual thesaurus. This research was a second stage of a project conducted at the initiative of The Romanian Institute for Standardisation (IRS) with the support of The National Institute for Information and Documentation (INID) in 1998. In the first stage, a working group established a list of descriptors that should be used in indexing by public libraries along with the corresponding notations of an abridged UDC edition. In 1999 IRS decided that the existing list of descriptors should be built into a multilingual thesaurus. We shall call this thesaurus LTHES.

The main purpose of creating this indexing tool was to enable easier access to information contained in Romanian public library catalogues. Thesaurus terms once available for the search offer improved search possibilities given the number of entry terms likely to be predictable or familiar to the user. Since one classification notation is the counterpart of only one descriptor, both the indexer and the searcher may switch between the UDC numbers and the thesaurus terms at their own choice. Secondly, the chance that bigger public libraries will be accessible via Internet is growing therefore multilingual access facilities will be advantageous to a larger variety of users.

It has often been said that a very appropriate way to get a better perception on something is by contrast. Therefore a more detailed multilingual thesaurus has been built taking the UDC Pocket Edition (1999) as a basis. We shall call this thesaurus PTHES. This thesaurus will certainly have a different influence on information retrieval than the first one as we shall demonstrate.

5.2 About relationships within indexing languages

In designing an interdisciplinary multilingual thesaurus with a broad coverage but a relatively low level of specificity, many of the difficulties emerged from the broad coverage, particularly from the existence of overlapping concepts, homographs and polysemic words on one hand, and from problems of translatability, on the other.

In most of the information systems of today, the users primarily go for the currently available title word index as a search method. The result may consist in large sets of retrieved records with a relatively low rate of precision. We have seen the '*SPICE'* example and the '*BANK'* example in this respect (see **§2.1**). To give another example: a book with the title "*The English language: its beauty and use*"[6], despite the evidence in the title does not deal with linguistics but with topics like creative writing, literary aesthetics and the like. The searcher may get a false representation of the subject of a document if the query is not formulated with the maximum of precision avoiding as much the possibility of confusion as possible. For the same topic request a searcher may get far better results if classification codes are used to search with instead of title words. In terms of precision and regarded from the user's point of view, classification codes exceed the performance of an indexing language in information retrieval on the well-known condition: the user needs expertise in handling them.

Apart from title words and classification codes, subject terms from an indexing language are also possible as alternative search method. Presumably, this gives improved results in information retrieval, situated between the high recall of the title word index (in spite of low precision) and the higher precision of the classified catalogue. Hunter (1994) appreciates the role of classification as a commonly used method to find things easier so we cannot afford to ignore the use of classification in an OPAC. There is a sound remark in the introduction to the pocket edition of the UDC (BSI, 1999, 5) about the effectiveness of organising information by

[6] The English language: its beauty and use. - London : Odhams Press, cop. 1947. - 384 p.

classifying it given the fact that classification codes are not limited in use by language impediments.

> “The problem for an information seeker is to find what is relevant and access what is needed – finding a way through the overwhelming volume of irrelevant material. There are various aids to doing this – some for virtual information (search engines for the Web) and some in either print or electronic form (bibliographies, catalogues, directories). They vary in effectiveness, and when relying on natural language can be limited by problems with words. (Did you use the right term? Are you searching in a single language? Are you missing relevant items in other languages?) Aside from sheer luck, search strategies are more effective if they can draw on information organised into patterns that correspond to the needs of most users – or are at least familiar with them – with related items brought together, and unrelated ones excluded – in other words, information that has been classified.”

The qualities of the UDC as a strong knowledge organiser have been proved by its use in organising the material of encyclopaedias. One of the important French encyclopaedias, *Bordas Encyclopédie*[7] is systematically divided into disciplines according to the UDC codes (their first two digits) throughout its contents[8]:

Vol. no.	*Volume Title*	*UDC No.*
vol. 1:	La vie animale	59
vol. 2:	Astronomie / préf. de Paul Courdec	52
vol. 3:	Philosophie ; Religions / préf. de Georges Pascal	10/19 and 20/29
vol. 4:	Histoire universelle	93
vol. 5:	Histoire universelle	94/99
vol. 6:	Visage de la Terre / préf. de Maurice Le Lannou	90/92
vol. 7:	Les lois de la nature / pref. de Jean Teilhac	53/54
vol. 8:	L'aventure littéraire de l'humanité / préf. de Roland Barthes	80/81
vol. 9:	L'aventure littéraire de l'humanité / préf. de Roland Barthes	82/89
vol. 10:	La vie des plantes / préf. de M. Guinochet	58
vol. 11:	Les nombres et l'espace	50/51
vol. 12:	Sciences sociales : part. I et II / préf. de J. Duvignaud	30/39 and 40/49
vol. 13 :	Beaux-Arts	70/71
vol. 14:	Beaux-Arts	72/78
vol. 15:	Matière inerte, matière vivante / [préf. de Roger Caratini]	55/57
vol. 16:	Jeux, divertissements, sport / [préf. de Roger Caratini]	79
vol. 17:	Médecine	60/61
vol. 18:	Art de l’ingénieur	62
vol. 19:	Techniques et métiers	63/69
vol. 20:	Techniques et métiers	63/69
vol. 21:	Généralités, bibliographie ; Index général	01/09

Figure 17. Example of systematically organised encyclopaedia according to the UDC structure

However, it is a well-known fact that classification notations are employed in information storage and retrieval only by experienced indexers and trained searchers (McIlwaine, 1993, 104). The average searcher in a bibliographic database will hardly if at all make use of the classification codes, finding the verbal expressions far more attractive for searching than the strings of digits, regardless of the latter accuracy. That is why we consider it very important

[7] Bordas Encyclopedie / par Roger Caratini. - Paris : Bordas, 1968-1975
[8] Some of the UDC numbers are out of use nowadays after several revisions of the tables

that designers of information systems should base their conception on information retrieval procedures on serious studies of user behaviour and not only take into consideration what they intuitively think it is appropriate for that system. When they approach an IR session the users look for something they do not know and the first thing that comes to their mind is a word from the title of a book. The extent to which that word is significant enough to give satisfactory search result is something we have already discussed above.

The experienced users of an information system will take advantage of and explore the alternative search methods the OPAC is offering and thereby will be able to compare and evaluate search results. Consequently, the system designers should bear in mind the effectiveness of clear and straightforward suggestions of alternative search methods in order to improve the information accessibility. The search methods in an OPAC can be context free (like the currently used 'word from the title') but also context dependent (like the descriptors belonging to a controlled vocabulary where the relational network between terms provide them with different meanings).

The rationale of the above-mentioned project (LTHES) was to build an interdisciplinary thesaurus based on the UDC logical structure. In making his point of view, the above-cited Hunter (1994) also insists upon the advantage of a thesaurus that adds a second dimension to an indexing language i.e. the interterm relationship. The hierarchical relations are given through notations within each class of the UDC scheme. Sometimes associative relations between concepts belonging to different though topically-related classes are also given. By its discipline-oriented character a classification system will hardly provide all the references to related facets of a given subject.

Unlike a classification scheme, in a thesaurus the interterm relationships are explicitly shown. This way the navigation through the conceptual links inherent in a thesaurus structure is made possible while those relations are openly indicated to the user who might not even be aware of them at first sight. We could add here again the usefulness of the subject index as a relative index, which is likely to show the context of a particular subject, thus providing an ideal search facility.

According to Rowley (1998, 208) a relative index contains at least one entry for each subject in the scheme and, by means of the alphabetical sequence of the scheme, brings together all aspects of a particular subject which are likely to be scattered by the discipline-oriented classification structure.

Marcella and Newton (1994) argue that the alphabetical index to a classification scheme is a necessary guide for the indexer to find the appropriate section of the schedule where a particular subject may be found. Additionally the relative index not only gives the subjects, but also locates them. It may sometimes include synonyms, showing related aspects of those subjects. Regardless of the classification scheme itself, the relative index brings together various aspects of a particular subject beyond one or another of the classes.

The index of the UDC International Medium Edition in English (UDC, 1985) gives six locations of the concept 'acoustics' in the scheme, i.e.:

Acoustics
  Applications 534.8
  Music 781.1
  Physics 534
  Seawater 551.463.2
  Stereo 681.84.087.7
  Technical 681.8

One more concept, *'Phonetics 81'342.1'*, which unquestionably fits into the same paradigm, can be included here. Therefore, the relative index goes behind the restrictions of the classificatory structure itself, offering a more flexible guide to both the indexer and the user of the classified catalogue.

Going as far back as Cutter's days, we learn about his syndetic approach, one in which subjects are linked together in an underlying classificatory structure (Cutter, 1904). He recommends that related subjects should be linked by a network of references to give a syndetic – from general to particular – catalogue (e.g. 'Literature' *see also* 'Drama' but not 'Drama' *see also* 'Literature').

For all that, a thesaurus is a more advantageous instrument compared with a relative index given its added value from relational point of view. The presence of the RTs in the thesaurus structure along with the BTs and NTs improve the effectiveness in use of the latter. Another advantage of a thesaurus is the existence of the '*see*' and '*see also*' references largely clarifying the meaning and use of terms. In addition to those, the scope notes do not permit misinterpretation or inappropriate usage of the thesaurus terms.

The semantic relations between the terms of a thesaurus largely clarify their meanings in a thesaurus. These relations, relevant for the thesaurus, are available as such in the classificatory structure. The hierarchies are given in the succession of the notations and the synonyms are summed up in the captions. What is left and is really critical to sort out is the issue of homonymy and polysemy. It is true that intelligent computer-aided procedures are able to establish concordances between vocabulary elements of more than one indexing language. With the existing online dictionaries the trouble of mapping vocabulary elements in several different languages is made far easier. But dissemination of such problematic lexical units i.e. homonyms and polysemes, requires human work especially when the different language variants are inter-filed and not considered separately. Consider for example the definite article "*the*" in English and the noun "*thé*" in French, or the noun "*rudiments*[9]" (see also *Figure 7*) which has in English a completely different meaning from French and Romanian (Frâncu, 1997). For such instances the scope note (SN) eliminates ambiguities defining the meaning and use of a term.

The effectiveness of an indexing language in terms of its retrieval power, resides, generally speaking, in three main characteristics:

- its capacity to distinguish among overlapping terms or formulas;
- its availability to give many entry terms – synonyms and quasi-synonyms – anticipating the users formulations;
- its ways to avoid situations in which search results contain too many irrelevant documents compared with the query.

To accomplish the general purpose of mapping classification notations and indexing terms, the correspondence or conversion table established between them, plus the cross-references between preferred and non-preferred terms are meant to diminish the weaknesses of each of the two information languages. In so doing, such a concordance works as an aid to their complementarity. Thus, the control on terms is kept and consistency in indexing is ensured.

But, it is also true that multiple language versions of the word-based information language enhance the indexing and the searching possibilities alike. To give a simple example: 'soft drinks' is a common compound term used to define 'non-alcoholic drinks' in English. Strange enough in Romanian there is no perfect equivalent to match this term. Therefore, several

[9] rudiment: 1. Biol. Organ which can hardly be seen, is growing or under-developed; beginning. 2. Figurative, usually in the plural. First elements of a theory, of an art, etc. (Romanian);

1. Plur. premières notions d'une science, d'un art. 2. Biol. Forme ébauchée ou atrophiée d'un organe (French)

formulas were thought of as approximate counterparts but most helpful in this situation was the French match, 'boissons refraîchissantes', that has a real equivalent in Romanian, 'băuturi răcoritoare' (see **§5.3e**). In a multilingual environment the UDC has the position of a switching language between the classification as such and each of the language variants of the thesaurus.

It has been said and is taken for granted that different language versions of the UDC tables will give the same result in classifying. As previously argued (Frâncu, 1997), this statement may only be true under two necessary conditions, namely:

- the level of specificity of those language versions is the same, and in our case the UDC editions considered should be at the same revision level (i.e. according to the same updating level imposed by *Extensions and Corrections to the UDC*);
- the wording and translation problems created by the equivalence of terms in the descriptions of notations should be given semantic solutions.

A multilingual thesaurus-based system gives the searcher the opportunity to select the dialogue language according to his knowledge or preference in order to access the information contained in the catalogue. The principle of equal treatment of the languages included in a multilingual thesaurus will provide the user with the same search result regardless of the chosen language (Hudon, 1997). Although as we saw in the preceding chapter, it is not always possible to treat the languages equally, there are ways to go around such impediments and find solutions to critical issues like the asymmetrical polysemy.

The requirement of equal treatment is often ensured by the way the thesaurus management software is designed (see point 4 below). The multilingual thesaurus management software used in building both thesauri is MTM3.1. The program is structured on CDS/ISIS support and has strong functionalities such as:

1. Control on the vocabulary terms within each language, i.e. the software doesn't permit the existence of a descriptor more than once in the thesaurus;
2. Automatic generation of reciprocating relationships whenever relationships are added i.e. once a BT was added to a descriptor automatically that descriptor becomes the NT of that one;
3. Automatic deleting of reciprocating relationships whenever relationships are deleted;
4. Mandatory provision for equivalent descriptors in each contributing language, i.e. the program does not allow leaving the entry menu before the fields for each of the mandatory languages are filled in;
5. Global modifying facilities by force of which one can replace, delete, split, merge old terms and add new terms;
6. Customised display formats can be created if the default ones are not satisfactory;
7. Import and export of files;
8. KWIC report generation;
9. Automatic translation of local descriptors when the replacing descriptor is mentioned in only one of the contributing languages;
10. Generation of alphabetical, systematic and hierarchical lists of terms;
11. Thesaurus-aided searching in bibliographic databases indexed with thesaurus descriptors;
12. Translation and validation of index terms in an application database;
13. Any updates of the thesaurus terms are recorded in a log database providing thus a means to keep track of all the changes made to records in the thesaurus database;
14. Checking on the validity of the thesaurus terms and relations and fix the possible failures.

5.3 Methodological issues in harmonizing the UDC structure with a thesaurus structure: the case of LTHES

In a comprehensive study, Clare Beghtol (1998) remarks the obvious tendency generated by a legitimate current need to move from traditional discipline-oriented classification systems to more flexible subject heading or thesaural systems allowing for the representation of multidisciplinary topics. The UDC unlike the DDC was not intended for shelving in the libraries therefore there is no limitation as to emphasise only one of aspects of the subject field of an individual document. Multidisciplinarity can be expressed in different forms by UDC codes, from multiple notations as access points to various synthetic devices like parallel subdivision, common or special auxiliaries provided by the tables and relation between two or more classes by colon.

Consider the following example:

Etnomedicina lingvistică/ Maria Purdela Sitaru . - Timişoara: Amacord, 1999

UDC notations:	*Descriptors:*
811.135.1’373:615.89	Linguistics
	Romanian language
	Lexicology
	Traditional medicine

The interdisciplinary concept in the title of this book is relatively easy to be represented in UDC numbers and equally easy to be indexed by descriptors derived from them. Each segment of the complex UDC number has its counterpart in terms from the controlled vocabulary. In addition to that, the syntax of the descriptors is following the syntax of the classification system. The alternative indexing by means of descriptors increases the user-friendliness of the information retrieval procedure. Irrespective of its complexity, the subject of the document in the example above is fully covered by the combination of the two UDC numbers and to the same extent by the descriptors assigned as indexing terms. The advantage of the subject headings over the precoordinated UDC notation is that the former permit the postcoordinated use of each of the descriptors in information retrieval. Each of the descriptors can stand alone for the representation of the subject of documents dealing in turns with Linguistics, Romanian language, Lexicology or Traditional medicine.

But what happens when the level of specificity is below the user’s expectations? What if, as in our case, the number of descriptors is limited and has to be restricted to the selection of UDC numbers in the abridged version of the UDC? To overcome this shortcoming, the device used is *upward posting* as recommended in the manual for thesaurus construction by Aitchison and Gilchrist (1987). Consequently, the number of non-descriptors will exceed by far the related number of descriptors thus providing a greater number of entry terms.

LTHES, the thesaurus described here, is an interdisciplinary multilingual thesaurus in English, French and Romanian of descriptors derived from a list of indexing terms based on a Romanian abridged version of the UDC. It is intended for public libraries of a large coverage and for a fairly low level of specificity.

In order to overcome the restrictions imposed by the limited number of descriptors we basically enriched the thesaurus terms by providing a great number of non-descriptors. By definition, a non-descriptor or non-preferred term is a ‘synonym or quasi-synonym of a preferred term’ (Aitchison and Gilchrist, 1987, 12). The analogy relation is implicit in this definition and for our purposes (Frâncu, 2000, 203) we expanded it to cover more aspects such as:

- Part to whole expressing, as in the example below, parts of an entity, concept or process e.g.

  GENERAL MECHANICS
  UF: Dynamics
  Gravitation
  Kinematics
  Kinetics
  Solid mechanics
  Statics

- Lack of it as part of the concept definition e.g.

  ARMAMENT
  UF: Disarmament

- Type of e.g.

  DECORATIVE ARTS
  UF: Artistic leatherwork
  Artistic paperwork
  Artistic textile work
  Artistic woodwork
  Floral Arts

- Place for e.g.

  PORTS
  UF: Docks
  Quays

- Agent of e.g.

  HEAT TREATMENT OPERATIONS
  UF: Heat treatment equipment

- Avoidance of e.g.

  ACCIDENTS
  UF: Accident prevention
  FIRE HAZARDS
  UF: Fire prevention

Acording to the rule, all these UF references are necessarily correlated with the reciprocal USE indications, according to the existing standards.

On the whole, the LTHES thesaurus arises not so dramatic problems of dissemination between homonyms and polysemes given its relatively small size, i.e. more than 1250 descriptors and double as many non-descriptors in each of the contributing languages. The degree of ambiguity is directly connected with the size of the vocabulary, i.e. the bigger the number of descriptors hence the specificity of the indexing language, the greater the possibility of overlapping terms.

Polysemy had to be treated within each individual language (e.g. the English 'Drawing' is a concept used in Technology and in Art but this is not true for its French equivalents 'Etirage' and 'Dessin', nor for the Romanian ones 'Tragere' and 'Desen'). Therefore, the English descriptors were attached a qualifier for each of the instances: Drawing (Technology) and Drawing (Art).

Another example might be the English term 'Drilling', which is found in the UDC both as a process in Machining (621.95) and as an operation in Mining (622.24). The solution adopted to disambiguate the term was to provide context to those terms that are prone to generate confusion, thus disambiguating them. The task here is carried out on the one hand by the combination of the ambiguous polysemous term with another word ('Drilling' and

‘Drilling technique’) and on the other hand by the neighbouring terms – context – used as semantic disambiguating factors (*Figure 18*).

English	French	Romanian
DRILLING	*MACHINES À PERCER*	*GAURIRE*
BT: Machining	*TG: Usinage*	*TG: Prelucrare mecanică*
DRILLING TECHNIQUE	*UP: Technique du sondage*	*FORARE*
UF: Drilling rig equipment	*SONDAGE*	*UP: Instalații de foraj*
BT: Mining operations	*TG: Exploitation des gissements*	*TG: Lucrări miniere*

Figure 18. Example of context as disambiguating device

The hybrid indexing instrument created by harmonising a classification scheme with a multilingual thesaurus allows the subjects of documents to be consistently indexed i.e. by always giving the same corresponding term to a UDC notation, thereby enabling alternative searching. Should the information system be appropriately designed, at the moment when a classification notation is selected to index with, the corresponding descriptor should be displayed on the screen. Conversely, any term used in searching, either descriptor or non-descriptor should lead to one and only one classification notation. Should this requirement be satisfied the controversial issue of consistency in indexing is no longer problematic. From this point on, the search can be enlarged or restricted according to the information need.

There is a prevailing trend nowadays among the library system developers to create systems that are capable of disseminating among various meanings of polysemous terms. Intelligent computer programs will ask the user which is the point of view a particular term is operating and the only thing the searcher has to do is select the context in which the term operates. For example, if ‘Mercury’ were the search term, given it may have various meanings, the computer would ask the searcher whether the context is ‘Physics’, or ‘Astronomy’ or ‘Car manufacturing’.

As previously stated the thesaurus design was limited in size by the number of terms in the descriptor list based on classification notations. Nonetheless, the total number of entry terms was considerably enlarged by the amount of non-descriptors. Once implemented in the information system it is for the users to determine the extent to which such an indexing tool responds to their requirements. Some of them might consider different terms in the position of descriptors or suggest more entry terms to access the information. With a view to enhance the quality of the thesaurus content and develop it in compliance with the users needs, it is highly recommendable that a field be provided for suggestions of additional terms or changes in the already existing ones. In keeping with the requirements of the vocabulary control, the proposals made by the users should be taken into account and evaluated on a regular basis by the thesaurus managers thus achieving the necessary feedback from the users (Yancey & Clarke, 1999).

The basic meaning of the concept of information retrieval has been through such a change until now that one can often hear about ‘discovering information’ rather than ‘finding information’ in a system. And it is so because of the interactive nature of many IR procedures that enable the user to navigate through conceptually related documents by means of the indexing language structure in order to get increased relevance of the search result. For that reason the presence of as many as possible synonyms, near-synonyms and also related terms is considerably practical for the purposes of a reliable information language. Navigation rather than browsing is a functionality that permits an expanded scope of and hence a larger view on the relevant information existing in an information system.

5.4 Remarks on the feasibility of LTHES, an interdisciplinary multilingual thesaurus based on an abridged UDC edition

As promissed in the beginning of this chapter we introduce here our findings as far as they resulted from a research on harmonising a Romanian abridged UDC edition with an interdisciplinary multilingual thesaurus. The purpose of creating this indexing tool was to enable easier access to information contained in Romanian public library catalogues.

This brings us to some conclusions to our approach. We can mention some of the problems that we identified in going from an abridged classification table to a thesaurus in three languages via an existing list of descriptors:

1. In a multilingual collection it is relevant to mention the language of a particular document as this is presumed to be of interest to the users of that collection; therefore the auxiliaries of language in the UDC tables were assigned 3 types of descriptors i.e.
    - for the language of a document (dictionaries or conversation guides always need a mention about their language or languages);
    - for linguistic topics associated with a particular language (a descriptor like 'grammar' should always be associated with a particular language);
    - for literature or fiction in a particular language (fiction cannot be read but in a language that a reader has very good command of).
2. There was some overlap in the terms assigned to classification numbers belonging to different classes of the UDC, therefore a disambiguation strategy had to be tackled (as in the English '*drawing*' example above); it has also been the case that only one of the three languages was problematic because of ambiguous terms (e.g. in Romanian the same word, '*locuință*'[10] was entered in the systematically built list of descriptors to designate different instances in its use: once for the living space from its functionality point of view and a second time to indicate different types of dwellings from architectural point of view). The strategy adopted in such situations was singular-plural distinction plus scope notes to distinguish between meanings *(Figures 19, 20, 21*).

Aitchison and Gilchrist (1987, 14) argue that there are differences in the use of singular and/or plural forms of terms according to the traditions of the language communities of the thesaurus compilers. Those in the French and German communities tend to prefer the singular with some exceptions where singular and plural forms have different meanings. Those in English-speaking communities while preferring the plural, use either plural or singular forms according to a standard set of rules. For reasons of respect for the principle of equal treatment of all participating languages we preferred plural forms for the languages used in our thesauri.

There are two basic categories of terms set out by the International Organisation for Standardisation ISO 2788 (1986): concrete entities and abstract concepts. In the first category 'count nouns' (e.g. names of countable objects subject to the question "how many") and 'non-count nouns' (e.g. names of materials and substances subject to the question "how much") are included. The former are given in the plural and the later in the singular. Abstract concepts comprise abstract entities and phenomena, properties, activities and disciplines and are given in the singular.

To give some examples let us have a look at the way a multilingual thesaurus is organised. For that purpose we shall consider the *Thesaurus de l'éducation Unesco: BIE* (1984), a thesaurus designed for indexing and retrieving information contained in documents on education with French terms and their equivalents in English and Spanish.

[10] locuință - Romanian for dwelling

ACTIVITE SCOLAIRE
SCHOOL ACTIVITIES
ACTIVIDAD ESCOLAR
BT ACTIVITE
NT ACTIVITE DE LA CLASSE
RT 357

ACTIVITE SPORTIVE
ATHLETIC ACTIVITIES
ACTIVIDAD DEPORTIVA
BT ACTIVITE PHYSIQUE
RT 357

ACTIVITE VERBALE
SPEAKING ACTIVITIES
ACTIVIDAD DISCURSIVA
RT 537

ADAPTATION
ADJUSTMENT
ADAPTACION
SN A l'environnement
NT ADAPTATION DE L'ELEVE
ADAPTATION EMOTIONELLE
ADAPTATION PROFESIONNELLE
ADAPTATION SOCIALE
RT 551

Some remarks are necessary to be made. The numbers given under each of the descriptors refer to the semantic fields of the facets and they are: 357 Activités, 537 Activités d'apprentissage and 551 Adaptation. The French descriptors and their Spanish equivalents are given in the singular. The English equivalents are in the plural with one exception, i.e. the term 'Adjustment' that is an abstract concept.

Special mention should be made on the English equivalent of the French term '*Activité sportive*' that has not exactly the same meaning, '*Athletic activities*' being more specific and narrower in meaning, to be exact, is a subordinate term of the mentioned descriptor, hierarchically speaking.

LOCUINŢĂ
E: Dwelling
F: Maison
CZU: 643/645
NE : Utilizat pentru spaţiul de locuit din punct de vedere al funcţionalităţii
UP : Bucătărie
Casa
Dormitor
Sufragerie
TG : Economie casnică
TA : Locuinţe

LOCUINŢE
E: Residential buildings
F: Constructions résidentielles
CZU: 728
NE : Utilizat pentru diverse tipuri de locuinţe din punct de vedere arhitectonic
TG : Arhitectură
TS : Arhitectură rurală
Blocuri cu apartamente
Castele şi conace
Cămine
Dependinţe
Locuinţe familiale
Locuinţe provizorii şi mobile
TA : Locuinţă

Figure 19. Example of singular/plural distinction as disambiguating device

Let us now have a look at the way the singular - plural distinction was used for disambiguation purposes. The two thesauri dealt with here were built according to the rules set out by the two well-known existing standards: ISO 2788 (1986) and ISO 5964 (1985) and have preference for the use of terms in the plural if not otherwise required (see *Figure 19*). In the following examples – and we speak about THES in particular – the singular plural distinction was made on purpose and underlined by the scope notes meant to clarify the meaning and use of each of the two variants

```
MAISON
E: Dwelling
R: Locuință
   NE : Employé pour l'espace
        habitable du point de vue
        de sa fonctionalité
   CDU: 643/645
   EP : Chambre à coucher
        Cuisine
        Domicile
        Salle à manger
   TG : Economie domestique
   TA : Constructions
        residentielles
```

```
CONSTRUCTIONS RESIDENTIELLES
E: Residential buildings
R: Locuințe
   CDU: 728
   EP : Maisons d'habitation
        Quartiers résidentiels
   TG : Architecture
   TS : Architecture domestique
        rurale
        Blocs d'appartements
        Châteaux et manoirs
        Dépendances domestiques
        Habitations unifamiliales
        Hôtels résidentiels
        Logements occasionnels et
        mobiles
   TA : Maison
```

Figure 20. Alphabetical display of French descriptors

```
DWELLING
F: Maison
R: Locuință
   SN : Used for the living
        space from its
        functionality point
        of view
   UDC: 643/645
   UF : Bedroom
        Dining room
        Home
        Kitchen
   BT : Home economics
   RT : Residential buildings
```

```
RESIDENTIAL BUILDINGS
F: Constructions résidentielles
R: Locuințe
UDC: 728
UF : Domestic architecture
     Dwellings
BT : Architecture
NT : Castles and manor houses
     Domestic dependencies
     Multi-family dwellings
     Occasional and mobile
        dwellings
     Residential hostels
     Rural domestic
        architecture
      Single-family dwellings
RT : Dwelling
```

Figure 21. Alphabetical display of English descriptors

For the other two languages, English and French, these terms are not problematic at all therefore they don't need to be disambiguated. The French corresponding terms are different from each other and so are the English ones (*Figures 20-21*).

3. Some more remarks on the feasibility of LTHES:
   a) quite often though the problem of overlapping terms within the 3 languages involved has been reasonably easy overcome given the low degree of specificity. We give below an example. A word like *"tour"* has several meanings in French among which:

(1) bâtiment très élevé de forme généralement ronde ou carrée,
(2) machine-outil servant à façonner une pièce monté sur un arbre animé d’un mouvement de rotation,
(3) rotation.

In order to differentiate the meanings of the word the following solution was adopted:

English	French	Romanian
LATHEWORK	FAÇONNAGE AU TOUR	STRUNJIRE
(621.941)	(621.941)	(621.941)
UF: Lathes	EP: Alésage	UP: Alezare
Turning	Tours	Strunguri
BT: Machining	TG: Usinage	TG: Prelucrare mecanică
SUPERSTRUCTURES	SUPERSTRUCTURES	SUPRASTRUCTURI
(624.9)	(624.9)	(624.9)
UF: Chimneys	EP: Cheminées	UP: Castele de apă
Masts	Mâts	Catarge
Water towers	Tours d'eau	Turnuri
BT: Civil engineering	TG: Constructions du génie civil	TM: Construcţii civile

Reciprocally, the UF relation of the French terms will then be:

Tours	Tours d'eau
EP : FAÇONNAGE AU TOUR	EP : SUPERSTRUCTURES

b) some of the descriptors, when compared with the other entry terms included in the thesaurus structure were found, formally speaking, unsuitable to the status of authorised headings, so they became non-descriptors; the difference mostly consisted in the preference for the scientific form of the term as compared to the popular one used for it;

c) during the process of assigning English and French equivalents to the existing Romanian descriptors, even though the English and French editions of the UDC proved very helpful, additional difficulties came to light such as:

- a number of names of peoples and language names in French are homonyms (e.g. *Danois, Suédois*, vs. *danois, suédois*), graphically distinguished from one another by the capitalised initial so an additional word was supplied in order to make distinction between such terms (e.g. *Peuple danois, Peuple suedois*); the same treatment was applied to English terms that denote both the language and the people (e.g. *Danish, Dutch, French*), the difference in natural language being made by the definite article;

- the English term *Soft drinks*, for instance, has more than one matching term in Romanian without any of those being a perfect equivalent of the source term. The French equivalent is semantically related to the English term and identical in form and meaning with one of the Romanian matching terms. Since none of them is a perfect equivalent of the English concept, the choice was made for the best match in the two target languages, i.e.: *Băuturi răcoritoare* in Romanian and *Boissons refraîchissantes* in French:

English	French	Romanian
Soft drinks	*Boissons refraîchissantes*	*Băuturi răcoritoare* Băuturi nealcoolice Băuturi uşoare[11]

- upward posting (*Figure 22*) that we already mentioned several times before, is largely used to provide as many access terms as possible for a thesaurus of this size and restricted to a previously established number of classification notations:

English	French	Romanian
FIELD CROPS	PLANTES DE CULTURE	PLANTE DE CULTURĂ
UF:Aromatic plants	EP:Plantes aromatiques	UP:Plante aromatice
Beverage plants	Plantes à boisson	?
Cereals	Céréales	Cereale
Condiment plants	?	?
Edible roots and tubers	Racines comestibles et tubercules	Rădăcini comestibile şi tuberculi
Forage grasses	Plantes fourragères	Plante furajere
Industrial plants	Plantes industrielles	Plante industriale
Leguminosae	Leguminosae	Leguminosae
Medicinal plants	Plantes médicinales	Plante medicinale
Oleaginous plants	Plantes oléagineuses	Plante oleaginoase
Plants yielding stimulants	Plantes stimulantes	Plante stimulante
Sugar plants	Plantes sucrières	Plante de zahăr
Tanning plants	Plantes à tanin	?
Textile plants	Plantes textiles	Plante textile
		Plante de câmp

Figure 22. The use of upward posting for lead-in term provision

What would happen if all these UF terms were not non-descriptors but just preferred terms? What would be the consequences of this change in the status of so many entry terms? First and foremost the coverage of the thesaurus would grow with immediate consequences on its specificity. While being very broad the thesaurus would cover the specific needs of a larger scale of users. While addressing a larger variety of users its flexibility needs to be enhanced so additional number of non-descriptors will be needed and more intellectual work as well. The descriptor 'Field crops' and its equivalents 'Plantes de culture' and 'Plante de cultură' will then become BT's for all the subsumed terms that change to NTs. More detailed subdivision of hierarchies will be required and equivalence of terms as authorised descriptors will be mandatory. What will then happen with such terms as 'Beverage plants' and 'Condiment plants' or 'Tanning plants' in English that have no equivalent in one or both of the other languages included in the thesaurus? With the growth in the number of descriptors the occurrence of overlap will be greater by far. The need for efficient disambiguating devices will increase and the control on terms will be harder to maintain. To sum up, the growth in number of thesaurus terms will increase the retrieval power by an increased precision of the thesaurus, it will increase its addressability but that would happen on the account of an increased amount of intellectual work and far greater maintenance difficulties than a smaller-size thesaurus will ever ask for.

Going back now to our thesaurus what is worth to be discussed about the above example is that not all the terms in one language have a satisfactory match in the other two languages. Some equivalents may have a form that is not always very much used in natural language: e.g.

[11] Literal translation of 'Soft drinks' in Romanian

‘*Plante stimulante*’ in Romanian. English being in our case the language with the greatest number of access terms, ‘*Condiment plants’* are not found in either French or Romanian, and ‘*Beverage plants’* and ‘*Tanning plants’* are not found in Romanian, as previously pointed out. We deal here again with asymmetrical structures but his time they are of little consequence to us. Since these terms are not descriptors, equivalence of terms is not mandatory, which works to our advantage. The cultural differences have an influence in this case too. What is remarkable here again is the resemblance in both lexicological and syntactical aspects of the two Romance languages considered. Compare:

Racines comestibles et tubercules	with*:*	*Rădăcini comestibile şi tuberculi*
Plantes fourragères	with*:*	*Plante furajere*
Plantes industrielles	with*:*	*Plante industriale*

The high degree of lexical and syntactical similarity between languages that belong to the same language family make the issues of wording and translation far easier for thesaurus designers than in the case of languages belonging to different language families.

```
643/645
   MAISON
   E: Dwelling
   R: Locuinţă
   NE : Employé pour l'espace
        habitable du point de vue
        de sa fonctionnalité
   EP : Chambre à coucher
        Cuisine
        Domicile
        Salle à manger
   TG : Economie domestique
   TA : Constructions résidentielles
728
   CONSTRUCTIONS RESIDENTIELLES
   E: Residential buildings
   R: Locuinţe
   EP : Maisons d'habitation
        Quartiers résidentiels
   TG : Architecture
   TS : Architecture domestique
        rurale
        Blocs d'appartements
        Châteaux et manoirs
        Dépendances domestiques
        Habitations unifamiliales
        Hôtels résidentiels
        Logements occasionnels et
        mobiles
   TA : Maison
```

```
ARCHITECTURE
ARCHITECTURE DOMESTIQUE RURALE
BLOCS D'APPARTEMENTS
Chambre à coucher
   EM : MAISON
CHATEAUX ET MANOIRS
CONSTRUCTIONS RESIDENTIELLES
   EP : Maisons d'habitation
        Quartiers résidentiels
   TA : MAISON
Cuisine
   EM : MAISON
DEPENDANCES DOMESTIQUES
Domicile
   EM : MAISON
ECONOMIE DOMESTIQUE
HABITATIONS UNIFAMILIALES
HOTELS RESIDENTIELS
MAISON
   NE : Employé pour l'espace
        habitable du point de vue
        de sa fonctionnalité
   EP : Chambre à coucher
        Cuisine
        Domicile
        Salle à manger
   TA : CONSTRUCTIONS RESIDENTIELLES
Maisons d'habitation
   EM : CONSTRUCTIONS RESIDENTIELLES
Quartiers résidentiels
   EM : CONSTRUCTIONS RESIDENTIELLES
Salle à manger
   EM : MAISON
```

Figure 23. Systematic display of descriptors with French as filing language and French alphabetical index

d) the display format is following the recommendations of the Guidelines for the Establishment and Development of Multilingual Thesauri (ISO 5964, 1985); there is an alphabetical arrangement of terms (both preferred and non-preferred) having each contributing languages as filing language in turns (see *Figures 20* and *21* for examples); there is a systematic display having the UDC number as entry element; additionally, each systematic section has its own alphabetical index in order to point out the non-preferred terms for each language as shown in *Figures 23, 24* and *25*.

This is the right point to underline that the languages are not interfiled in the alphabetical display in order to avoid confusion. In addition to that, the alphabetical index associated with the systematic section for each language includes some relational information such as synonymous and related terms. We give below an example of a sequence of the systematic display of descriptors with English as filing language and one of the English alphabetical index:

```
643/645
   DWELLING
   F: Maison
   R: Locuinţă
   SN : Used for the living space
        from its functionality point
        of view
   UF : Bedroom
        Dining room
        Home
        Kitchen
   BT : Home economics
   RT : Residential buildings
728
   RESIDENTIAL BUILDINGS
   F: Constructions résidentielles
   R: Locuinţe
   UF : Domestic architecture
        Dwellings
   BT : Architecture
   NT : Castles and manor houses
        Domestic dependencies
        Multi-family dwellings
        Occasional and mobile
        dwellings
        Residential hostels
        Rural domestic architecture
        Single-family dwellings
   RT : Dwelling
```

```
ARCHITECTURE
Bedroom
   USE: DWELLING
CASTLES AND MANOR HOUSES
Dining room
   USE: DWELLING
Domestic architecture
   USE: RESIDENTIAL BUILDINGS
DOMESTIC DEPENDENCIES
DWELLING
   SN : Used for the living space
        from its functionality point
        of view
   UF : Bedroom
        Dining room
        Home
        Kitchen
   RT : RESIDENTIAL BUILDINGS
Dwellings
   USE: RESIDENTIAL BUILD
Home
   USE: DWELLING
HOME ECONOMICS
Kitchen
   USE: DWELLING
MULTI-FAMILY DWELLINGS
RESIDENTIAL BUILDINGS
RESIDENTIAL HOSTELS
RURAL DOMESTIC ARCHITECTURE
SINGLE-FAMILY DWELLINGS
```

Figure 24. Systematic display of descriptors with English as filing language and English alphabetical index

Most of these remarks apply to the case when the thesaurus is used in printed form rather than in machine-readable form. The latter is more difficult to be installed and exploited at all its potentialities. As we shall see in the next chapter there have been attempts and research projects have been conducted to prove the feasibility of using classification systems and particularly the UDC in a machine-readable format in order to permit automated classification and indexing.

5.5 Remarks on the multilingual UDC thesaurus based on the Pocket Edition (PTHES)

The second thesaurus, for reasons explained in the following paragraphs, does not cover all the classes of the UDC. Our second thesaurus has a total number of 7553 terms of which 2033 are descriptors, the ratio between them amounting at 1 in 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, a part of 633 and Class 8 of the UDC Pocket Edition.

The primary reason for not building a thesaurus on the whole of the UDC Pocket Edition (BSI, 1999) was that the remaining classes would behave more or less the same as the already existing ones, both in building the descriptors based on the UDC captions and in using them in information retrieval. Classes 5, 6 (except for 61), 7 and 9 are partly enumerative, partly faceted, and so are Classes 3 and 8 that have descriptors in both thesauri. Therefore most of the ongoing remarks will apply more or less the same to the classes that were left out. We shall focus in our considerations on the classes that have descriptors in both thesauri and compare the results.

```
643/645
   LOCUINŢĂ
   E: Dwelling
   F: Maison
   NE : Utilizat pentru
        spaţiul de locuit din
        punct de vedere al
        funcţionalităţii
   UP : Bucătărie
        Casă
        Dormitor
        Sufragerie
   TG : Economie casnică
   TA : Locuinţe
728
   LOCUINŢE
   E: Residential buildings
   F: Constructions résidentielles
   NE : Utilizat pentru
        Diverse tipuri de
        locuinţe din punct de
        vedere arhitectonic
   TG : Arhitectură
   TS : Arhitectură rurală
        Blocuri cu apartamente
        Castele şi conace
        Cămine
        Dependinţe
        Locuinţe familiale
        Locuinţe provizorii şi
        mobile
   TA : Locuinţă
```

```
ARHITECTURĂ
ARHITECTURĂ RURALĂ
BLOCURI CU APARTAMENTE
Bucătărie
   VEZI:LOCUINŢĂ
Casă
   VEZI:LOCUINŢĂ
CASTELE SI CONACE
CAMINE
DEPENDINŢE
Dormitor
   VEZI:LOCUINŢĂ
LOCUINŢĂ
   NE : Utilizat pentru
        spaţiul de locuit din
        punct de vedere al
        funcţionalităţii
   UP : Bucătărie
        Casă
        Dormitor
        Sufragerie
   TA : Locuinţe
LOCUINŢE
   NE : Utilizat pentru
        Diverse tipuri de
        locuinţe din punct de
        vedere arhitectonic
   TA : Locuinţă
LOCUINŢE FAMILIALE
LOCUINŢE PROVIZORII SI MOBILE
Sufragerie
   VEZI:LOCUINŢĂ
```

Figure 25. Systematic display of descriptors with Romanian as filing language and Romanian alphabetical index

To begin with, some comments on the thesaurus building process are following. They generally give account on the difficulties encountered in making up a multilingual thesaurus based on a selection of the UDC schedule, the Pocket Edition. They also present some solutions adopted in one or another of the situations. The selection of the UDC notations having been made on some unknown criteria, a general remark that we can derive from our experience is that the UDC Pocket Edition was hardly issued for information retrieval purposes as such. It has rather been created for organizing collections of documents by providing instructions – particularly in the case of faceted classes – on how to combine numbers, how to build them up according to the UDC rules. Likewise, this edition clearly gives an account of the logical structure of the UDC as a classification system. Consider as example the indications given under 811 Languages:

> *"The subdivisions of 811 are derived from =1/=9 (Table 1c) by substituting a point for the equals sign, e.g. 811.111 'English' derives from =111. Only a few examples are given here. Denote the linguistic details by hyphen and/or apostrophe auxiliaries from 81"(p. 176)*

This kind of indication may be found with more or less the same content in Class 821 of as much as elsewhere in the tables (Class 3 and Class 9 for instance). Building a thesaurus on such structuring indications demands the creation of extensive lists of possible combinations of notations prior to converting the notations into thesaurus terms.

At this point we can state that a list of descriptors previously made and expressing with a fair degree of exactness the average needs of the average user of the document collection is preferable for our purposes. This was our hypothesis in building LTHES. The troublesome side of such an approach is that not all the possible combinations of notations are predictable by the thesaurus compiler so the problem of coverage remains unsolved. A combination of descriptors that can effectively be used in multiple variants at the moment of searching seems to be the appropriate solution e.g.

Romanian linguistics (811.135.1)
Use: ROMANIAN + LINGUISTICS

Romanian literature (821.135.1)
Use: ROMANIAN + LITERATURE, etc.

But before formulating other conclusions let us examine some particularities of the newly made thesaurus as they came up while building it:

1. Many of the language names in the auxiliary tables are cognates in all 3 languages e.g.: Pidgin English, Plattdeutsch, Koine, Hindi.

2. As previously argued documentary languages are considered to be artificial languages (see §1.1, p. 7) which implies that certain terms cease to have the same meaning in a documentary language as it has in the natural language it is derived from. We can then have the situation of false friends, as it is the case of the French equivalent of the English language name Welsh: Velche ou welche 1. Français ignorant, montrant des prejuges. – Par ext. Homme ignorant, naïf et lourd. 2. pour les All., terme de mepris appliqué à ce qui est français, belge, suisse romand ou (parfois) italien. (cf. Dictionnaire de notre temps. Hachette, 1988). The right French equivalent for the English word Welsh is gallois.

3. For some UDC notations e.g. (0.027.5), the captions in the French medium edition (1990) gives approximate equivalents like: *Paperback editions = Editions de poche*.

Given the structural, semantic and lexical similarities between French and Romanian, the Romanian equivalent translates the French caption (*Ediții de buzunar*) and not the English one.

4. Spelling variants can avoid duplicate terms e.g. Aethiopia (as place of the Ancient Africa) and Ethiopia (as place of the modern world). Alternatively, each of the terms would get a qualifier functioning like contextual disambiguator: Aethiopia (antiquity) and Ethiopia (modern world).

5. There is a reciprocal impact of the auxiliaries of language and the main numbers of Class 8. Linguistics that is to say languages as common auxiliaries are homonyms to languages as main numbers for individual languages from linguistics point of view. The addition of the word 'language' (and its corresponding terms in the other two participating languages) for the Class 8 range of descriptors would differentiate those from the descriptors corresponding to the auxiliaries of language e.g.:

   English is the corresponding descriptor for *'=111'* in the notation:
   54(038)=*111* meaning: Dictionary of chemistry in English

   English language is the descriptor for the same *'=111'* changed according to the UDC grammar in
   811.*111*`36 meaning: Grammar of the English language

6. Likewise, there is a reciprocal impact between numbers belonging to the same class, in our case Class 0 and particularly between many of the auxiliaries of bibliographic form and main numbers such as:

   (01) Bibliographies and *011/016 Bibliographies* or
   (05) Serial publications and *050 Serial publications*.

Solutions: *(01) Bibliography (form)* and *011/016 Bibliographies*.
(05) Serial publications (form) and *050 Serial publications and periodicals*

7. The situation is even more critical when the same captions appear in the same class, but not as subdivisions of the same hierarchy as in the previous example. Here also, the solution is an alternative expression for the more or less the same concept (see the Bibliography examples). In such cases the scope notes will clarify the meaning of the terms. Example:

Study of organization
Etude de l'organisation /
Studiul organizării
005
SN: Study of the theory and
principles of classification
and taxonomy
UF: Systematisation in general
BT: Prolegomena + Generalities
RT: Methodology

Methodology
Méthodologie / Metodologie
001.8
UF: Analysis and synthesis
General study of method
BT: Science and knowledge
++ Generalities
NT: Case studies (as subject)
Organization of science and
scientific work
RT: Study of organization

8. It is hardly avoidable not to mention the difficulties of finding French equivalents to the English-oriented terminology used in computer science when we consider Class 004 of the UDC. Although most of the natural languages adopted loan words to define concepts belonging to computer science and technology, this is not the case of the French language that employs sometimes quite strange terms and formulas for commonly acknowledged and literary warranted English words in this field. Here are just a few but illustrative examples:

   - *Messagerie électronique* for *e-mail*
   - *En ligne* for *online*
   - *Logiciel* for *Software*
   - *Science et technologie de l'informatique* for *Computer science and technology*

9. The hierarchies in Class 2 under the subdivision 225 New Testament in general in the Pocket Edition are completely erroneous. The Gospels are not a subdivision of the New Testament but a subdivision of the Bible itself (226). Likewise, The Epistles (227) are a new subdivision of the Bible instead of being a subdivision of the New Testament. This situation gives problems in establishing the hierarchies of the thesaurus[12].

10. There is also a problem of grouping several sub-categories under a higher one: e.g. the Pocket Edition has under 295 Parseeism. Zoroastrism. Mazdaism. Mithraism, without placing them under a category that sums up them all. The French Medium Edition[13] has under the notation 295, Religions perses, which gives the solution to the problem.

11. With Philosophical viewpoints (the subdivisions of 141) the problem is of a different nature. The expressions given in the captions are very long and by converting them into descriptors predictability is affected. We have under 141.1, for example, a number of four philosophical viewpoints according to number and quality of principles: monism, dualism, platonism, neoplatonism. That is because our purpose and basic rule is to give each UDC number one descriptor and only one, and consider the subdivisions of the entity, concept or process it represents as non-descriptors pointing to the descriptor as preferred term. Therefore, all the four philosophical viewpoints will be non-descriptors because they do not have a number in the table. The descriptor, though, has to have a form predictable enough to be accepted by a potential user.

12. As a rule, the upper subdivisions of a class number collect the meanings of many other lower ones e.g. *330.1 Science of economics. Basic economic concepts, theory. Value. Capital. Funds*. Almost all concepts are listed as entries, with their own number, further in the tables e.g. *330.13 Profitability. Economic principle. Utility. Value. Value principle; 330.14 Capital. Funds for material production*. The manual indexing that permits the use of any of the two possibilities (the upper or the lower subdivision of the UDC class) and hence one or more (different) descriptors for the same kind of subjects, will have as result of such inconsistency the scattering of information in several sections of the subject catalogue (see the examples from Class 159.9 – Psychology given in **§6.5.2**, p. 114).

[12] In the meantime the structure of Class 2 Religion has been completely changed into a faceted one (see Extensions & Corrections to the UDC Vol. 22, The Hague, UDCC, 2000)

[13] The French Medium Edition was used for the French descriptors in the thesaurus

13. Sometimes the upper subdivision of a class includes categories from the lower hierarchies without grouping them under a collective name whatsoever e.g. Cavalry. Mounted troops. Motorised troops (357) with subdivisions Cavalry (357.1) and Motorised troops (357.5) (see also point 12 above). There are two options available to solve the inconvenience:

    - to create a *dummy term* or *node label* e.g. *Cavalry and motorised troops* and keep the UDC notation for the sake of the logical hierarchy or

    - to go further down and consider the hierarchical subdivisions, in our case 357.1 and 357.5, and skip the upper UDC level which gathers them (see the rule in our previous example of *Political parties*).

14. Quite often in the tables we can find identical captions under different UDC numbers as subdivisions of the same hierarchy, e.g. *Power of the state* (342.1) *and Power of the state* (342.5). In such cases an alternative caption has to be formulated as descriptor while a clarifying scope note brings the necessary specification on the meaning of the term. The groups of descriptors have the following form:

    342.1
    E: Power of the state
    F: Pouvoir de l'Etat
    R: Putere de stat
    UF: Nation
    People
    State
    BT: Public law
    Law
    NT: Structure of states
    RT: System and function of organs of government

    342.5
    E: System and function of organs of government
    F: Système et fonction des organes de gouvernement
    R: Sistemul şi funcţia organelor de guvernământ
    SN: Used to denote the political organization of the state
    BT: Public law
    Law
    RT: Power of the state

15. Special auxiliaries in each class are difficult to manage when it comes to converting from the UDC into thesaurus terms. We gave a single example of their use and not all the possible combinations between main numbers within that class and these auxiliaries. Examples: *314.044 Voluntary change* and 314.045 *Forced change* and their possible use in *314.7.044/.45 Voluntary/forced migration*. Likewise, the combinations between main numbers such as *311.3 Official statistics* with other class numbers to generate new concepts e.g. *311.3:331.56 Official unemployment statistics* or *311.3(492) Official statistics in the Netherlands*. A possible solution to this problem could be the application of the same algorithms in the case of the UDC-based descriptors as to the decomposed UDC notations. Using prescribed combinations of

descriptors *Official statistics* and *Unemployment* or *Official statistics and the Netherlands* would not hinder by any means either the indexing or the retrieval. As a matter of facts this is how 'individual literatures' and 'individual languages' were treated (see **§7.2**).

16. The subdivisions of *329 Political parties and movements* enumerate different political attitudes and outlooks with little regard for the levels of division: *329.1/.6 Political parties and movements according to general political outlook and aims* in principle covers all the various categories but in practice the last number found in the table is *329.4 Parties and movements with predominantly ethnic, racial, linguistic aims*. This is, of course, due to the selection of numbers from the MRF. Since its meaning brings nothing new in the class structure, we skipped it altogether. Furthermore, subdivisions like 329.27 and 329.28 are not even in the MRF itself, therefore, the immediate higher hierarchical level being absent, the BT for the subsequent subdivisions of those i.e. 329.271/.3 and 329.285/.286 is the same as for the rest in the class, namely 329 Political parties. A rule can be drawn from this example: whenever the immediate lower subdivision of a UDC number brings nothing new to the hierarchy of the class but just puts together a range of numbers, the highest level in the hierarchy should be taken as broader term in the thesaurus structure and the range of numbers underneath should be skipped. Each of the lower subdivisions become then narrower terms. Should any of the lower subdivisions miss their super-ordinate level in the hierarchy, they go under the same broader term as the other terms, irrespective of the level of subdivision.

Classificatory structure

329	Political parties and movements
329.1/.6	Political parties and movements according to general political outlook and aims
329.11	Conservative attitude
329.12	Liberal attitude
329.13	Progressive attitude. Revolutionary attitude
329.14	Socialist attitude. Social-democrat attitude
329.15	Communist attitude
329.17	Nationalist attitude
329.18	Fascist attitude
329.21	Monarchist attitude
329.23	Republican attitude
329.271	Adherents of a unitary state (unitarianists)
329.272	Adherents of a federal state (federalists)
329.273	Separatist outlook. Secessionist outlook
329.285	Anarchist outlook
329.286	Nihilist outlook
329.29	Advocates of a corporate state (corporatists)
329.3	Parties and movements with religious outlook
329.36	Parties and movements with antireligious, atheist, anticlerical outlook
329.4	Parties and movements with predominantly ethnic, racial, linguistic aims

Thesaurus structure

Political parties
SN Used for political parties and movements according to general political outlook and aims
UF Political movements
BT Politics
NT Anarchist outlook
Communist attitude
Conservative attitude
Corporatists
Fascist attitude
Federalists
Liberal attitude
Monarchist attitude
Nationalist attitude
Nihilist outlook
Parties and movements with ethnic, racial, linguistic aims
Parties and movements with religious outlook
Progressive attitude
Republican attitude
Separatists
Socialist attitude
Unitarianists

17. Precoordination by bracket qualifier was used in a relatively small number of cases and that was mostly when one and the same concept occurred in several different disciplines such as: *Marriage (Ethics), Marriage (Religion), Marriage (Demography), Marriage (Ethnography)* or *Cosmogony (Metaphysics), Cosmogony (Religion).* As a rule, the term used as a qualifier is not the immediate higher number in the hierarchy but rather denotes the class or subclass, as seen in the examples.

18. *Synonymy* is language-dependent as already stated in the foregoing. In French and Romanian for instance, different words are used to express concepts that in English can only be expressed by one word. To give an example: under *Military administration* (355.6) the term *Pay* is identical with the same term under *Salaries* (331.2) in English. Therefore they both need and have a qualifier in brackets. Not in French and Romanian where there is a special word to denote salaries in the army: *Solde* in French and *Soldă* in Romanian,
    e.g.

    E*: Pay (Military affairs)* to be distinguished from *Pay (Labour)*
    F: Soldes
    R: Soldă

19. The Supernatural (398.4) makes 'par excellence' an example of culturally-biased range of subdivisions. Such categories as good and evil spirits are characteristic to each of the languages, as they do belong to each of those language speaking peoples, separately. Therefore, what we deal with here is not a mere translation of terms; hardly can we even talk about equivalents, but rather have to do with an enumeration of supernatural beings, expressions of the supernatural world, in each of the three languages strongly related to the folklore popular traditions and beliefs of each of the three nationalities considered
    e.g.

English	**French**	**Romanian**
Supernatural	*Le Monde surnaturel*	*Supranatural*
UF Bugbears	UF Bons et mauvais spirits	UF Balauri
UF Demons (folklore)	UF Démons (folklore)	UF Căpcăuni
UF Dragons	UF Diables	UF Demoni (folclor)
UF Elves	UF Esprits de la nature	UF Himere
UF Fairies	UF Fantômes	UF Năluci
UF Ghosts	UF Fées	UF Spirite bune şi rele
UF Giants	UF Gnômes	UF Stafii
UF Goblins	BT Folklore	UF Zmei
UF Good and evil spirits	RT Démons (Religion)	BT Folclor
UF Gremlins		RT Demoni (Religie)
UF Witches		
BT Folklore		
RT Demons (Religion)		

20. Going back to the structure of LTHES and the restrictions imposed by the pre-existing list of descriptors that limited the number of indexing terms, let us examine the situation of PTHES that has no such restrictions or, if they exist, they do not affect that much the hierarchical structure of the thesaurus. Many or better said almost all the non-descriptors in LTHES changed their status into that of descriptors with beneficial

consequences for the precision rate in information retrieval in PTHES. Furthermore, each descriptor in the subdivisions of 633 – to take an example – has its own non-descriptors thus increasing the number of access terms considerably (see Appendices 2 and 3). In the forthcoming chapter on the impact of specificity of the indexing language on information retrieval we shall show what are the consequences of this change in the status of thesaurus terms. We give hereunder the configuration of the descriptor group for 'Field crops' in PTHES (compare this with *Figure 22*):

English	**French**	**Romanian**
Field crops	*Plantes de culture*	*Plante de cultură*
UDC: 633	CDU : 633	CZU : 633
NT Aromatic plants	NT Céréales	UP Plante de câmp
Cereals	Herbes fourragères	NT Cereale
Edible roots and tubers	Plantes aromatiques	Ierburi furajere
Forage grasses	Plantes fourragères	Plante aromatice
Forage plants	Plantes industrielles	Plante de zahăr
Industrial plants	Plantes stimulantes	Plante furajere
Plants yielding stimulants	Plantes sucrières	Plante industriale
Sugar plants	Plantes textiles	Plante stimulante
Textile plants	Racines comestibles et tubercules	Plante textile
BT Agriculture		Rădăcini comestibile şi tuberculi

5.6 Methodological issues related with building the described UDC-based thesauri

The question may legitimately arise: why derive a thesaurus from the UDC? The simple answer is that the UDC provides the basic logical structure necessary to start building on it a more user-friendly and more easy-to-use tool to search with. The consistency and control of notations connecting related topics within a discipline and across disciplines are also among the major advantages of using this particular numeric system and not a random one based on simple running numbers.

Once the descriptor list has been established the added relationships characteristic to the thesaurus structure enhance the effectiveness in use of such an indexing and searching tool. The thesaurus thus established gives the user the chance to enlarge and restrict the search according to his needs.

Another advantage of the thesaurus presented is that it provides context that functions as major disambiguating device. The given examples of "drilling" as a process in Machining (621.95) and at the same time as an operation in Mining (622.24) are clarifying this assumption (p. 79).

There are certain requirements that the thesaurus builder has to be aware of when approaching the creation of such an indexing and retrieval tool. In the first place the thesaurus building guidelines and rules have to be accurately followed. Should any restrictions be imposed on the structure or usage of the thesaurus they have to be dealt with such a way that on the whole, the thesaurus remains within the limits of acceptance of the existing international standards. The two sides of such an approach should be kept in mind:

1. the thesaurus building side with all its particularities
2. the thesaurus usage side giving account on its effectiveness

Another requirement is one of a totally different nature, i.e. the thesaurus builder should know what kind of a bibliographic database the thesaurus is meant for (in other words, the document collection), the growing rate of the database, how much detail the users need and what category of users the collection is addressing.

While a descriptor is assigned to a document up for indexing it will automatically be associated with its corresponding UDC number and reverse. This functionality may have far going consequences in terms of indexing consistency.

If the thesaurus is implemented in a database whose bibliographic records are indexed with UDC numbers alone and those numbers are included in the thesaurus structure, there are two methods the automatic assignment of descriptors can be made possible:

1. based on links between bibliographic databases and authority files containing both UDC numbers and thesaurus terms (provided that software specialists define the correct approach for that);
2. within the same database with fields specially designed to hold different indexing languages or indexing methods.

Having the descriptors as alternative indexing and retrieval device the friendliness of the retrieval system is much enhanced and, which is of utmost importance, the reliability of the information system itself is growing.

Yet, if we consider the LTHES, shortcomings may come up connected with the level of specificity, taking into account the number of authorised terms. In spite of that, the multitude of lead-in terms supplied as a consequence of the use of upward posting as the main building device is rewarding. Alternatively, the low level of specificity diminishes the occurrence of ambiguity and the amount of overlapping terms allowing for a larger yet more compact number of retrieved records that can be restricted by other search methods (words from title, author, etc.). This is much in the line of Gillman (1997, 116-117) who argues that large general thesauri such as the "Root thesaurus" and the "ILO/BIT Macrothesaurus" are commonly used as sources of indexing terminology, not to drive the retrieval process.

However, the extent to which the first multilingual thesaurus presented (LTHES) can be efficiently used in information retrieval is still to be confirmed. The multilingual aspects of this interdisciplinary thesaurus they have largely been treated in the preceding paragraphs of this chapter (see **§5.3** and **5.4**).

The second thesaurus that was developed (PTHES) created more or less the same type of problems but difficulties appeared more acutely in the following situations:

a. the degree of specificity being higher more concepts have homonymous expressions therefore disambiguation was necessary; the main disambiguating device used was the addition of a qualifier in brackets, specifying the discipline that concept belongs to (e.g. 'marriage' and 'cosmogony');
b. many times the hierarchical structure of the UDC is troublesome which makes the thesaurus building difficult in some parts of it (e.g. Class 2 Religion);
c. for concepts that are not expressed in the tables but can be built up by means of existing numbers (as for example those expressed by parallel subdivisions or special auxiliaries) the solution used in PTHES is different from LTHES; in the former we adopted the combination of descriptors (e.g. Romanian linguistics and Romanian literature) whereas in the later, the list of descriptors the thesaurus is based on includes a number of predictable combinations in an enumerative sequence (e.g. the languages and literatures part of the thesaurus);
d. *node labels* were created aiming at preserving the hierarchical structure PTHES is based on (e.g. *Cavalry and motorised troops – 357* having two terms as subdivisions: *Cavalry* – 357.1 and *Mounted troops – 357.5*); such terms are not supposed to be used

in indexing but just mentioned in the thesaurus to make the hierarchical relation possible and preserve the selection of UDC numbers in the edition used as a base;

e. for geographical locations found under different UDC numbers (one for *Places of the ancient world* and another for *Places and countries of the modern world*) the disambiguating device used was of lexical nature – *lexical variants* – but again the combination of two descriptors would be an alternative (e.g. *Syria* (569.1) and *Ancient Syria (394)* was preferred to *Syria* and *Syria + Antiquity*).

Some questions may come up after all the topics discussed here. Why still preserve the UDC numbers in a thesaurus based on them? Why use alternative indexing and hence searching procedures instead of using only one, i.e. the friendlier one based on natural language words? Imagine this scenario. A user starts a search session in a classified catalogue bearing in mind his information need: books on organic chemistry. He does not speak the language of the catalogue therefore he cannot formulate his query in that language. The only possibility for him to find something is to start searching with an author's name that he is sure of or try out different variants of title words using truncation.

Such an approach may give some results but the recall rate will be fairly low. From the few records that he might get he should derive the UDC number likely to represent the concept of organic chemistry. Knowing that all the bibliographic records have classification codes for subject indexing he will get improved search results as compared with what he got in the first instance. Additionally, if the subjects of bibliographic records in that catalogue are indexed with descriptors too, he may go further and call those records using descriptors as alternative search method, irrespective of the language of those descriptors.

Therefore, classification notations, no matter whether they be known to the potential user or not, should be kept in the information system, even though from a certain point on, the decision is made to use natural language words in indexing. Sooner or later, those classification notations might prove useful in situations like the one just described. This is even more so, as we shall see in an forthcoming chapter, when classification codes given for subject indexing constitute the pre-requisite of a more sophisticated development: adding text to the subject notations in order to make searching with words possible and automatically assign descriptors based on classification numbers.

5.7 Conclusions

What is worth mentioning in the first place about this chapter is that it addresses once more *the issue of the UDC being in the position of an intermediate language or switching language*. Many of the attributes of the UDC that stand for this quality have been advantageously used in this research. The possibility it provides to an information system to switch between the classification notations and the thesaurus terms on one hand and between one and another of the three languages on the other, fully demonstrate this.

The multilingual thesaurus we have designed and built and that we introduced in this chapter proved to be the right 'missing piece' for our study case; therefore it was implemented and extensively used to demonstrate our statements in the sixth chapter. Based on our previous research in the field of *harmonizing the UDC structure with a thesaurus structure*, the multilingual thesaurus covering all the classes of the UDC demonstrated its feasibility. Despite the restrictions imposed by the pre-existent list of descriptors and the relatively low degree of specificity of LTHES, the number of entry terms (both descriptors and non-descriptors) authorizes its functionality, as our experiment will provide evidence.

A brief account on the capabilities of the *UDC as a powerful knowledge organiser* creates the background for its relational configuration as compared with that of a thesaurus.

Translation difficulties are also pointed out and ways and manners to solve them are discussed. Once again, the lexical and syntactical resemblance between two of the languages involved in the project belonging to the same language family, Romanian and French are outlined. Translatability and wording are far easier problems within the same language family than in the case of different language families. *Issues of homonymy and polysemy across the three languages* and some *problems of equivalence of terms* between them are focussed on. Some examples of *systematic displays and alphabetical indexes* are given for each participating language. Moreover, the two appendices at the end of the thesis give samples of the alphabetical and systematic displays of both LTHES and PTHES.

We insist on the problem of lack of specificity in the descriptors of one of the above-described thesauri that may be critical to some extent. Among the drawbacks of the approach we are going to use in our case study we can mention the necessity of a close relationship between the specificity of the classified catalogue of the given database and the specificity of the UDC-based thesaurus for a better evaluation of the system. If this requirement is ignored then we have a problem of compatibility between the classified catalogue and the selection of UDC numbers the thesauri are based on. The direct consequence of this situation is the existence of different rates of recall and precision but to some extent also information loss.

With a view to enhance the quality of the thesaurus content and develop it in compliance with the users needs, it is highly recommendable that an additional field is provided in the system for suggestions of new entry terms or changes in the already existing ones. By users here we mean both indexers and searchers that should nominate new candidate terms (see *Figure 4*). This has to be done with care in order to always keep the rule of only one descriptor (or prescribed combination of descriptors) for one classification number or else the principle behind our approach is not working any more. A history note of the terms recording the changes those terms undergone would be recommendable.

CHAPTER 6
ONLINE APPLICATIONS OF THE UDC-BASED MULTILINGUAL THESAURI

One of if not *the* earliest project ever elaborated and meant to explore the online capabilities of a classification system was the experimental system called AUDACIOUS, with an acronym that stands for **Au**tomatic **D**irect **Ac**cess to **I**nformation with the **O**nline **U**DC **S**ystem (Freeman & Cochrane, 1968).

The experiment used a database containing references from a single issue of Nuclear Science Abstracts and the Special Subject Edition of UDC for Nuclear Science and Technology and it was developed to allow information retrieval. The UDC was considered as a translation tool from the user queries (formulated in natural language words) into logical statements containing UDC numbers.

One of the reasons why the UDC and no other indexing language was used for this experiment was that outside the United States this was the most extensively used tool. Another reason was that UDC was best suited for handling scientific information, providing an internationally accepted controlled vocabulary to work with. With the perspective of a continuously growing number of online users of international information networks and the awareness of large amounts of files with UDC-indexed documents being continuously created the experiment was started long before the Internet was initiated[14].

The experiment was a command-driven interactive system which provided remote direct access to files containing 2330 items from Nuclear Science Abstracts indexed by UDC. AUDACIOUS was a part of a larger project providing machine-readable files of UDC, automatic typesetting, composition of UDC schedules and statistical evaluation of the UDC as a retrieval tool (McIlwaine, 2000).

In the conclusions of the detailed report the authors recommend solutions for possible indexing methods that "would serve adequately for users who do not share a common natural language". The three suggested solutions were:

- To use the language in which the largest volume of literature is written, i.e. English
- To use a form of indexing that is not dependent on natural language, i.e. the UDC
- To use a system that would permit indexing and searching using a controlled natural language vocabulary of local choice but having also a table of equivalences between the UDC and the natural language vocabulary; this system would take advantage of the logical structure and hierarchical notation of the UDC without the users being aware of this; the UDC being the internal form of indexing, the users of the system would address their queries in natural language words without regard to the original indexing language used.

These conclusions had a great influence on the future applications of the UDC in automated systems.

[14] In 1973, the U.S. Defense Advanced Research Projects Agency (DARPA) initiated a research program to investigate techniques and technologies for interlinking packet networks of various kinds. The objective was to develop communication protocols that would allow networked computers to communicate transparently across multiple, linked packet networks. This was called the Internetting project and the system of networks that emerged from the research was known as the "Internet"(http://www.isoc.org/internet/history/cerf.shtml)

6.1 Searching with words derived from the UDC text as online application of the UDC

As previously shown the UDC has some beneficial online applications already in use at the ETH Library in Zurich (see **§4.3**). Although not straightforwardly said, the CoBRA+ Forum consider using for the creation of their prototype the DDC numbers in the LCSH file available at the British Library (Clavel-Merrin, 1999):

> "It would also be productive to study to what degree selection of headings in a field could be automated e.g. using the hierarchical relationships (narrower terms), or by using classification numbers already assigned to headings, according to each SHL classification scheme (for example DDC numbers in the LCSH file used at the BL)".

The universal classification systems, as stated above, despite their respectable age, have some qualities that, if adequately explored, permit significant online developments and thus contribute to improving information access.

Another application of the same type as the ETHICS system is called GERHARD (German Harvest Automated Retrieval and Directory) (Möller et al., 1999). This is an automatic classification and indexing system for the World Wide Web that allows integrated searching and browsing. The automatic classification is based on the UDC authority file used by the ETH Library in Zurich. As stated above, an additional advantage of this authority file is that the index terms attached to classification numbers are in English, German and French (see **§4.3**). The first step towards the automatic classification method was the extraction of the vocabulary called "UDCZ-Lexicon" from the ETH system. The underlying principle of GERHARD is text analysis of German Web pages harvested into a database and then matching the words they contain to UDCZ-Lexicon and assigning notations to each of them. The resulting clusters of documents harvested under these notations undergo several subsequent operations using algorithms and other statistical and weighting methods the result of which is a user-friendly system that permits both browsing and searching.

The characteristic feature of the UDC that determined its application in online environment preferable to any of the other largely used classification systems is that individual elements constituting a compound notation are clearly marked and delimited by symbols and punctuation. This feature was used much to the advantage of designers of online applications of the UDC as the ETHICS in Zurich and the research described below.

This was indeed the rationale of the approach taken by Riesthuis (1999, 24-32) in his research on the possibility of searching with words as they derive from the text added to decomposed UDC notations. Having as starting point his findings as resulted from his doctoral thesis (Riesthuis, 1998) his study is conceived as an attempt to answer two questions (p. 24):

1. Is it possible to convert precoordinated UDC notations into postcoordinated UDC notations in such a way that searching with individual components of complex notations or arbitrary combinations of parts – using Boolean algebra and truncation – becomes feasible?
2. Is it possible to add text to the subject notations in order to make searching with words possible?

His approach has two distinct stages in keeping with the announced purposes. In the first stage, in going from precoordinated to postcoordinated UDC notations the elements of the compound UDC notations have to be isolated or decomposed according to seven groups of algorithms (p. 25). These groups of algorithms are developed in close connection with the symbols and punctuation used by the UDC as facet indicators i.e.

- form facets: *language* =… and *bibliographic form* of a work (0…)
- content facets: *place* (1/9), *race* (=…), *time* "…" and *general subject*.

The second stage of adding text to the notations consists in attaching textual meaning to each separate part of the UDC notations as they resulted from the decomposition procedures.

What is noteworthy about this particular stage in the progress of the study is that the need was felt to provide context to the separated parts derived from the decomposition work. For this reason a subfield **^x** was included in the bibliographic format to accommodate the necessary context. Specifically this was done by extracting 175 high level notations from the main table of the MRF, representing each of the disciplines/classes of the UDC by the first two digits (except for classes where more than two digits are necessary to make clear what the number represents, e.g. 008 *Civilization. Culture. Progress*, *159.9 Psychology, 616 Pathology. Clinical medicine*). This subfield once added to the subject notations of each bibliographic record in the database[15] has a direct outcome i.e. the possibility of creating special subject/domain bibliographies at the push of a button.

Consider this example given by the author of the study (Riesthuis, 1999, 31) to illustrate the manner in which text was added to the decomposed UDC notations. If a book in the catalogue has as UDC notations:

821.135.1-1
821.135.1Hasdeu, B. P.1.01
099.5Rebreanu, P. F.
099.3Ilin, S.

the meanings of these UDC notations, according to the decomposition and textual definition procedures applied, will be as follows:

Hasdeu, B. P. – Complete works or sets – in original language
Rebreanu, P. F.
Ilin, S.
821.135.1-1^e Romanian – Poetry, Poems, Verse ^x Literature
821.135.1 ^e Romanian – [Literature] ^x Literature
099.3 ^e Works with autographic dedication ^x Manuscripts. Rare and remarkable works
099.5 ^e Works of value because of their provenance. Books belonging to historic, royal, literary personages, great collections
^x Manuscripts. Rare and remarkable works

This particular functionality has far going consequences that we consider beneficial to both indexers and searchers alike (see also the first paragraph of this chapter). But they will be presented in detail further in this thesis. The main advantages defined by the author of this research if its results were put into practice would be the following (p. 32):

1. the search possibilities can be enhanced and indeed they can be tremendously enlarged by the detailed information otherwise hidden in complex notations, for instance the ranges with a stroke "/"; searching postcoordinately by means of words representing parts of the precoordinated UDC notation expands the area of searching increasing the recall rate;
2. the control of complex notations can be done automatically if they were built according to the grammatical rules of the UDC;
3. the use of the algorithm procedures makes easier the keep up of the bibliographic database with the changes in the information language.

[15] The example given and all the examples in the article of Riesthuis (1999) are taken from the online catalogue of the Central University Library of Bucharest (BCUB) - Romania

6.2 A short historical background of subject indexing in the BCUB

The BCUB catalogue has a long tradition in using the Universal Decimal Classification for subject cataloguing.

Founded in the year when the first international bibliographic conference was held in Brussels, 1895, the then "University Foundation Carol I"[16] adopted the latest practices of the time in all library activities. For the systematic catalogue (Tzigara-Samurcas, 1933, 75-79) the University Foundation adopted as early as 1914 the "decimal system" for reasons of 1) its logical structure and 2) its notation in Arabic numbers conferring language independence.

While being aware of the advantages of the new decimal system and considering it superior to both Cutter's system and Dewey's system Tzigara-Samurcas (1933, 79), the head librarian of the institution, clearly states these advantages:

a. it is *international*: in all libraries, although in different countries, the classes are the same;
b. it is *encyclopaedic*: it offers a complete and simple structural frame in which every document can be classed;
c. it is *practical*: the notation is as simple as possible under the appearance of complexity;
d. it is *mnemonic*: it gives the possibility, despite its coverage, to be easily mastered, given its basic principles. It is sufficient for one to memorise its main subdivisions and they will guide him throughout the variety of knowledge;
e. it is *expandable* continuously, without imposing troublesome changes;
f. it is *rational and scientific* as it follows the basic lines of scientific classification, keeping an equal distance from the numberless classifications of sciences which are not always concordant.

Once the direction was established in the way the collections of the library were reflected in the subject catalogues the institution's policy was to keep the pace with every new development in the structure of the UDC for the coming years.

Towards the end of the 1980s, despite the Romania's isolation, the BCUB recorded a growing interest for topics like: library automation, online catalogues, subject cataloguing using descriptors. However the automated activities were rather scarce and inefficient and the efforts made here and there could hardly meet with the expected results.

In spite of that in 1986 one project conducted at the initiative of the Centre Européen pour l'Enseignement Supérieur (CEPES) of UNESCO brought together the intellectual force and professional abilities of librarians for a completely new indexing activity. A working group of librarians with good knowledge of foreign languages was established and their task was to index a number of foreign periodicals on education in English, French, Italian, Czech and Russian. They had to follow the rules of indexing via a different indexing language but what they knew. The indexing tool used was the BIE (Bureau International d'Education) Thesaurus of UNESCO (BIE, 1991), a multilingual thesaurus with a large coverage in education and related topics. The descriptors were manually entered in a printed work format that was subsequently used as a source for a database.

Automation became a reality in the BCUB in the following years. Started in the beginning of the 1990s automation meant a tremendous change in the library's activities particularly for the cataloguing and classification routines. Initially 'accepted' with a lot of reserve or even opposition, automation was considered as an authentic challenge by many of the librarians. New skills like typing and editing abilities were mere necessities to start with. Poor typing abilities (like lack of words spacing or carelessness in typing the sign of repeatability in a field) would lead to information scattering or information loss. The importance of such simple

[16] The initial name of today's Central University Library of Bucharest (BCUB)

rules as the need of a space after the initial of a name or a comma, or of the difference that it makes whether a name is written in upper case or in upper and lower case characters was underestimated by many of the librarians. It took some time before librarians became aware of the major effects of not following such minor rules.

As to the subject indexing, the coming of automated library routines did not bring much change in it. The UDC was still the only indexing language available to define the subject of documents. Initially the library used CDS/ISIS for cataloguing and the catalogue thus created was not more than a computerized version of the traditional card catalogue. This was the more so as additional editing work was needed in order to provide the catalogue with cards as customary with any of the Romanian libraries.

Fundamentally, the change in the cataloguing and indexing activities at the BCUB started when the VUBIS integrated library system[17] was implemented and a new online catalogue was designed. The UDC was no longer the only indexing tool used as long as other subject fields were provided for in the system. The subject proper of the documents was just one of the subject fields available. To this some more were added such as: a field for geographical subjects, one for personal name as subject, another one for institution name as subject. All these further possibilities of the library system considerably enhanced the information retrieval potential of the database.

The transition from the CDS/ISIS catalogue to the VUBIS catalogue was not quite straightforward. Apart from the technical efforts required by the conversion from one database to another, the additional subject fields available in VUBIS made the concern of the cataloguers and indexers alike. The existence in the catalogue of two kinds of bibliographic records (old ones from the CDS/ISIS database with only UDC notations for the subject as in *Figure 26* and new ones created in VUBIS with descriptors as in *Figure 27*) would confuse the searcher. The number of records that needed to be added descriptors (about 60,000) demanded painstaking and time-consuming efforts from an indexer that should have been dispossessed of any other activity for a while. Indeed the experiment of about half a year of additional indexing and reclassification proved the activity not to be productive at that moment in the history of our catalogue.

```
+---------------------------------------------------------------- BCUB  ----+
¦TIT: Introduccion general a las ciencias sociales, política, sociología,  ¦
¦     antropología, economía, geografía humana y económica, psicología,    ¦
¦     [ed. Leopoldo Fornés Bornavia, Antonio Izquierdo Escribano, Armando  ¦
¦     Pérez Pino,...].- Madrid, Playor, 1989.- 386p., il., 24cm. -         ¦
¦     Bibliogr. & glosar dupa cap.. - ISBN 84-359-0631-0                   ¦
¦UDC: 316/32                                                               ¦
¦UDC: 159.9                                                                ¦
¦MAIN:  ^a159.9^ePsychology^xPsychology                                    ¦
¦MAIN:  ^a316^eSociology^xDemography. Sociology. Statistics                ¦
¦MAIN:  ^a32^ePolitics^xPolitics                                           ¦
¦PDES:  ^z159.9^ePsychology^fPsychologie^rPsihologie                       ¦
¦PDES:  ^z316^eSociology^fSociologie^rSociologie                           ¦
¦PDES:  ^z32^ePolitics^fPolitique^rPolitică                                ¦
¦LDES:  ^z159.9^ePsychology^fPsychologie^rPsihologie                       ¦
¦LDES:  ^z316^eSociology^fSociologie^rSociologie                           ¦
¦LDES:  ^z32^ePolitics^fPolitique^rPolitică                                ¦
¦LNDES: ^z316^eSocial classes^fClasses sociaux^rClase sociae               ¦
¦                                                          MFN: 36418      ¦
+--------------------------------------------------------------------------+
```

Figure 26. Example of bibliographic record with subject description without descriptors

[17] VUBIS is a product of GEAC ('s-Hertogenbosch, The Netherlands).

The need to create a database able to reflect as fast as possible the large collection of documents waiting to be available to the users of the library determined the decision makers to abandon the idea of additional indexing. The main concern of the institution – imposed by the existence of large amounts of documents not reflected in the database – was the development of the online catalogue although this priority would impede on its accuracy and consequently affect the information retrieval.

```
+------------------------------------------------------------- BCUB  ---+
¦TIT:  Traité pratique d'analyse du caractère / par Gaston Berger.- 4e éd.¦
¦      - Paris : Presses Universitaires de France, 1958. - 251 p. ; 20 cm ¦
¦DES:  Psihologie individuala                                             ¦
¦DES:  Personalitate (Psihologie)                                         ¦
¦UDC:  159.923                                                            ¦
¦MAIN:  ^a159.923^eType psychology. Individual psychology. Psychology of  ¦
¦       individualities. Individuality. Personality. Character psychology.¦
¦       Characterology. Idiosyncrasies. Personal equation. Personality    ¦
¦       types. ^xPsychology                                               ¦
¦                                                     MFN 71959           ¦
+-----------------------------------------------------------------------+
```

Figure 27. Example of bibliographic record with subject description including Romanian descriptors

6.3 Purpose of our case study

The research of Riesthuis (1999) may be a real solution in a situation like the one just described above. While going back to each record and checking the validity of the UDC numbers would be a tremendous and time consuming work the possibility of assigning automatically descriptors to the bibliographic records lacking them in the database might save a great deal of time and considerable intellectual effort.

The remarkable benefit in terms of time and effort would have its drawbacks as well. To begin with, the possibility of assigning descriptors to bibliographic records as an alternative subject representation mode along with the UDC codes brings forward the problem of vocabulary control. But will the vocabulary control still be problematic in case of automatic assignment of descriptors taken from a UDC-based thesaurus? This together with other problems involved by the management and use of the descriptors derived from the text of the UDC notations will be investigated one by one in the ongoing.

Considering the conclusions of the above described research as a starting point we shall make an attempt to solve some of the problems just mentioned avoiding as much the manual approach as possible. Additionally, based on our previous studies on the feasibility of a UDC-based multilingual thesaurus (Frâncu, 1996, 1999b, 2000), we shall add multilingual descriptors to the records in an experimental bibliographic database thus offering multilingual access to the subject of the documents. The main reason for using thesaurus terms rather than classification codes in information retrieval is that words are friendlier to the user than numbers. As long as there is a reasonable high degree of compatibility between the two, it is advisable to switch in favour of the thesaurus terms for information search and retrieval.

As user-friendliness of information retrieval systems became a highly regarded imperative sceptical voices were heard that the class numbers are not desirable and have no future. However, the numerical representation is advisable to be preserved in such systems in order "to preserve the original meaning and scope of a particular concept" (McIlwaine, 2000).

The problems encountered, their solutions in as much as the input and the output are concerned, the advantages and the drawbacks of such a system just as well will be presented in the upcoming paragraphs of this thesis.

To sum up the purpose of the present research is to prove that in a bibliographic database with subjects represented initially by UDC notations and partly by manually assigned descriptors, the automatically added descriptors from thesauri based on the UDC structure enhance the search results and this is more so as these descriptors are in more than one language. For this purpose we shall use a comparative approach testing the extent to which the search result is influenced by alternative search methods. Different degrees of specificity in thesaurus structure will have as consequences different recall and precision rates in information retrieval, as we shall demonstrate in the following chapter.

6.4 The structure of the experimental database

The database used in our experiment consists of three parts:

- the bibliographic database as such
- the Master Reference File (MRF) and
- the shortened variant of the main tables of the MRF.

The database was created in Micro CDS/ISIS format Version 3.08 (c)UNESCO 1997 and includes a total of 231411 records of which 165305 are bibliographic records, 65931 belong to the UDC MRF as of the year 2000 (Extensions & Corrections vol. 22) and the rest of 175 are high level UDC notations that represent each of the disciplines/classes by their first digits (see the table below). The two multilingual thesauri (PTHES and LTHES) are embedded in the MRF part of the experimental database where descriptors and non-descriptors belonging to each of them are put in separate fields: 160 and 165 for PTHES and 170 and 175 for LTHES. The mentioned fields in the MRF play the active role in this research.

Sections of the experimental database	**No. of records**
Records in the bibliographic database	165305
Records in the Master Reference File (MRF)	65931
Records in the short MRF	175
Total number of records	**231411**
Bibliographic records indexed with manually assigned descriptors	112888
Bibliographic records indexed with automatically assigned multilingual descriptors from PTHES	148432
Bibliographic records indexed with automatically assigned multilingual descriptors from LTHES	162574

Looking at the table that shows the structure of the experimental database we can easily notice the difference between the number of records in the bibliographic database (165305) and the number of bibliographic records indexed with automatically assigned multilingual descriptors (148432 records have descriptors from PTHES and 162574 have descriptors from LTHES). The number of bibliographic records indexed manually is also important to be known. It is presumed that all bibliographic records are classified with UDC numbers. Actually there are bibliographic records that have no UDC number, many of those being multi-level bibliographic records for multi-volume documents. The automatic indexing had as result a lower number of records indexed with multilingual descriptors. On the one hand the difference in numbers 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 basis for the multilingual thesauri. On the other hand, if we compare the number of records automatically indexed with LTHES terms with the number of records indexed with PTHES terms the difference is in favour of the former because the latter has descriptors for only some of the classes of the UDC Pocket Edition (see **§5.5**). But even so, the number of bibliographic records indexed automatically is rather high and therefore serves our purposes.

The data entry worksheet (BCUB) for the experimental database comprises the following main fields[18]:

```
200 Title etc. [TIT:]
610 Subject headings in Romanian [DES:]
611 Geographical Subject headings [DES:]
675 UDC subject notation [UDC:]

701 Aux. for language [LANG:]
702 Aux. for bibliographic form [FORM:]
703 Aux. for place [PLAC:]
704 Aux. for ethnic group & nation [ETHN:]
705 Aux. for time [TIME:]
706 Alphabetical addition [TEXT:]
707 Aux. for point of view [VIEW:]
708 Aux. for material / persons [CHAR:]
709 Main numbers [MAIN:]
711 PTHES descriptors in bibliographic record [PDES:]
712 PTHES non-descriptors in bibliographic record [PNDES:]
713 LTHES descriptors in bibliographic record [LDES:]
714 LTHES non-descriptors in bibliographic record [LNDES:]

001 UDC number
002 Table
100 Description
105 Verbal examples
160 PTHES descriptors
165 PTHES non-descriptors
170 LTHES descriptors
175 LTHES non-descriptors
```

The fields in the first part of this list – 200 through 714 – are set up for the bibliographic part of the BCUB database; those in the second part – 1 through 175 – belong to the MRF part of the experimental database. We give below an example of an MRF record:

Format: MRF		MFN: 177669
UDC number:	025.4	Table: M
Description:	Classification and indexing	
Verbal examples:	Indexing and retrieval languages. Classifications, thesauruses etc. and their construction	
References:	001.82 168.2	
Pthes Descriptors:	Classification and indexing, Classification et indexation. Clasificare şi indexare	
Pthes Non-Descriptors:	Classifications, thesauri and their construction, Langages d`indexation. Limbaje de indexare şi regăsire	
Pthes Non-Descriptors:	Indexing and retrieval languages. Sisteme de clasificare, tezaure şi construcţia lor	
Lthes Descriptors:	Classification, Classification. Clasificare	
Lthes Non-Descriptors:	Classification and indexing, Classification et indexation. Clasificare şi indexare	

[18] The square brackets contain the field indicators as they appear in the bibliographic database (see also Appendix 1)

Taking a closer look at the bibliographic records given as examples throughout the thesis from this point one will certainly remark the 3 subfields of the 709 fields (*Figures 28, 29*). As they are important for the progress of the study we feel necessary to give here their meanings:

^a is meant for the separate parts of the UDC notation in field 675;
^e contains the text belonging to the part of the UDC notation in ^a;
^x provides the context for the main number and has as source the short MRF.

At the same time, for better understanding the mechanism of automatic assignment of descriptors from LTHES and PTHES it is helpful to give the meanings of the subfields likely to occur in 71- fields (*Figures 30, 31*):

^z and ^y co-exist in a record when the UDC number in the bibliographic record is not found in the MRF and the program truncates it until the combination of digits found has a corresponding descriptor (*Figure 30*);
^z exists alone in a record when the UDC number in the bibliographic record is found as such in the MRF and has a corresponding descriptor (*Figure 31*).

6.5 Work method: steps taken and stages of the case study

The ultimate goal of our case study is to give an answer to the question: does converting the UDC numbers into words for searching enhance the information retrieval? If so, to what extent?

For that purpose we undertake the following procedures which, in our opinion, will enable us to formulate the answer we hope for:

1. "Cleaning-up" the database;
2. Mapping the UDC numbers with descriptors;
3. Making multilingual subject headings available.

The first action taken over the selected database is the running of the split up program against the classification notations assigned to each of the bibliographic description. This is done in two steps: first, the split up of the combinations of *main numbers* and *main and auxiliary numbers connected by relation (: or ::) and coordination (+)* and second, the split up of the rest of complex notations using *ranges, special auxiliaries and parallel subdivisions*. The intricate process of splitting up the complex UDC notations is comprehensively described by Riesthuis in his study published in volume 21of Extensions and Corrections to the UDC (Riesthuis, 1999, 25-31).

The next step, namely the addition of text to the decomposed parts of the complex notations (and to the main numbers as much as they exist) is carried out by running 9 different programs against the separate parts resulted thereby, one for each kind of UDC notation:

- main numbers or combinations of main numbers;
- strings of main numbers and common auxiliaries;
- strings of special auxiliaries and parallel subdivisions;
- ranges using a stroke.

The result of the operations described above is illustrated in *Figure 28*. The example shows the text added to and derived from a simple UDC notation consisting of only one *main number*. Note the automatically added descriptors from LTHES.

```
+ 9 / 11 ---------------------------------------------- Format: BCUB  --+
¦TIT:  Bazele acusticii moderne / Eugen Badarau si Mircea Grumazescu. -  ¦
¦      Bucuresti : Editura Acad. R.P.R., 1961. - 508p. : fig. ; 24cm. -  ¦
¦      Include bibliogr. si index.                                        ¦
¦DES:  Acustica                                                           ¦
¦UDC:  534                                                                ¦
¦MAIN: ^a534^eVibrations. Acoustics^xPhysics                              ¦
¦LDES: ^z534^eAcoustics^fAcoustique^rAcustică                             ¦
¦LNDES:^z534^eVibrations^fVibrations^rOscilații mecanice^rVibrații        ¦
¦      mecanice                                                           ¦
¦                                                             MFN: 96463 ¦
+------------------------------------------------------------------------+
```

Figure 28. Bibliographic record with text added to a simple UDC notation

When more sophisticated UDC codes are used to represent the subject of a document more fields come into play and consequently a larger amount of text is provided. The bibliographic record in *Figure 29* illustrates this situation where complex UDC notations, several auxiliaries of language and an auxiliary of form being included represent the subject of the document.

```
+- 6 / 15 --------------------------------------------- Format: BCUB  --+
¦TIT: Thésaurus de l'Éducation Unesco : liste par facettes de termes     ¦
¦     destinés à la recherche des documents et données relatifs à        ¦
¦     l'éducation, avec leurs équivalents anglais et espagnols / préparé ¦
¦     par le Bureau International d'Éducation. - 15e éd. rév. et augm.   ¦
¦DES: Tezaur de termeni                                                  ¦
¦DES: Educatie                                                           ¦
¦UDC: 025.4.06:37=133.1=111=134.2                                        ¦
¦UDC: 37(038)=133.1=111=134.2                                            ¦
¦LANG: ^a=133.1^eFrench                                                  ¦
¦LANG: ^a=^e111English                                                   ¦
¦LANG: ^a=^e134.2Spanish                                                 ¦
¦FORM: ^a(038)^eDictionaries. Language dictionaries. Special subject and ¦
¦      technical dictionaries                                            ¦
¦MAIN: ^a025.4.06^eClassification and indexing. Indexing and retrieval   ¦
¦      languages. Classifications, thesauruses etc. and their            ¦
¦      construction Indexing and retrieval languages for special subjects¦
¦      Special classifications. Special thesauruses^xLibrarianship       ¦
¦MAIN: ^a37^eEducation. Teaching. Training. Leisure^xEducation. Teaching.¦
¦      Training. Leisure                                                 ¦
¦                                                           MFN: 55596   ¦
+------------------------------------------------------------------------+
```

Figure 29. Bibliographic record with text added to UDC notations including auxiliaries of language and form

The *ranges of UDC numbers* graphically represented by a stroke in the tables bring together the meanings of all the classification codes included in between the digits mentioned before and after the stroke (the stroke has the same effect as the Boolean operator OR). The result of the way Riesthuis (1999, 27) treated the ranges of UDC numbers consists of the display of all the numbers included in the expression along with their corresponding text as illustrated in *Figure 30*.

The UDC notations built up by *parallel subdivisions* and using *special auxiliaries* are added text as it is shown in *Figure 31* where the number representing the individual literature is the result of the parallel subdivision of 821 by the numbers representing the language code, in this case '=133.1' for French. That number is subsequently added the special auxiliary representing the literary genre i.e. '-4' for Essays.

```
+ 1 / 13 ----------------------------------------------- Format: BCUB  --+
¦TIT:  Stammesgeschichte der Menschheit / Hans Weinert. - Stuttgart :     ¦
¦      Kosmos, 1941. - 80p. : fig. ; 20cm                                 ¦
¦DES: Antropogenie                                                        ¦
¦UDC: 572.1/.4                                                            ¦
¦MAIN: ^a572.1^eUnity of the human species. Monophyletic or polyphyletic  ¦
¦      origin. Monogenism. Polygenism^xHuman anthropology                 ¦
¦MAIN: ^a572.2^eHeterogeneity of the human species: races, physical types¦
¦      varieties^xHuman anthropology                                      ¦
¦MAIN: ^a572.3^eAnthropology [~] xxx^xHuman anthropology                  ¦
¦MAIN: ^a572.4^ePlace and time of origin of the human species^xHuman      ¦
¦      anthropology                                                       ¦
¦LDES: ^y572^z572.1^eAnthropology^fAnthropologie^rAntropologie            ¦
¦LDES: ^y572^z572.2^eAnthropology^fAnthropologie^rAntropologie            ¦
¦LDES: ^y572^z572.4^eAnthropology^fAnthropologie^rAntropologie            ¦
¦                                                         MFN: 67997      ¦
+-------------------------------------------------------------------------+
```

Figure 30. Bibliographic record with text added to a UDC notation using a stroke

```
+ 211 / 234763 ----------------------------------------- Format: BCUB  --+
¦TIT:  Proust, Freud et l'autre / Jean-Louis Baudry. - Paris, Minuit,     ¦
¦      1984. - 154p., 22cm.. -  L'ecrit du temps. - ISBN 2-7073-0698-3    ¦
¦BDES: Psihanaliza                                                        ¦
¦BDES: Literatura franceza                                                ¦
¦BDES: Eseuri                                                             ¦
¦UDC:  821.133.1-4                                                        ¦
¦UDC:  159.964.2                                                          ¦
¦MAIN: ^a821.133.1-4^eFrench^oFranceză - Essays^gEssais                   ¦
¦      ^xLiterature^oLiteratură                                           ¦
¦MAIN: ^a159.964.2^ePsychoanalysis^oPsihanaliză^xPsychology^oPsihologie   ¦
¦PDES: ^y159.964^z159.964.2^ePsychanalyse^fPsychanalyse^rPsihanaliză      ¦
¦PDES: ^y821^z821.133.1-4^eLiteratures of individual                      ¦
¦      languages^fLittératures relatives à des langues                    ¦
¦      particulières^rLiteratura limbilor individuale                     ¦
¦PDES: ^z82-4^eEssays^fEssais^rEseuri                                     ¦
¦PDES: ^z=133.1^eFrench^fFrançais^rFranceză                               ¦
¦LDES: ^y159.9^z159.964.2^ePsychology^fPsychologie^rPsihologie            ¦
¦LDES: ^z821.133.1-4^eFrench literature^fLittérature                      ¦
¦      française^rLiteratură franceză                                     ¦
¦LDES: ^z82-4^eEssays^fÉssais^rEseuri                                     ¦
¦LDES: ^z=133.1^eFrench^fFrançais^rFranceză                               ¦
¦LDES: ^y82^z821^eLiterature^fLittérature^rLiteratură         MFN 211     ¦
+-------------------------------------------------------------------------+
```

Figure 31. Bibliographic record with text added to a UDC notation resulting from parallel subdivision and a special auxiliary

Mention should be made here on the fact that if both a parallel subdivision and a range of numbers coexist in the structure of a complex UDC notation e.g. 821.1/.2 for Indo-European literatures (correctly built from the UDC grammar point of view) the two algorithms cannot handle them correctly and therefore an error is given e.g.

```
709:  ^a821.1^eLiteratures of individual languages [~] -
       [Literature] [P] xxx^xLiterature^oLiteratură
709:  ^a821.2^eLiteratures of individual languages [~] -
       [Literature] [P] xxx^xLiterature^oLiteratură
```

At this point we can conclude that these are the types of bibliographic records existing in the experimental database. Irregularities may come up as some of the records have descriptors assigned alternatively with the UDC numbers, others do not. This is just a question of the time the records were created. As a rule, the earlier records have only UDC numbers, the later have also descriptors (see *Figures 26* and *27* respectively). The most important thing is that they all have classification notations so our needs and purposes are thoroughly accomplished. Although not all of them are correct the simple fact that they exist enables us to make something out of their form and meaning. In case of doubt the formal description in any of the bibliographic records turns to be a valuable source of validation for the classification code.

6.5.1 "Cleaning-up" the database

For bibliographic records that contain UDC numbers with errors in them the program is designed in such a way that the errors are pointed out (*Figure 32*). They are graphically marked by either or both of these signs: ~ and **x** repeated as many times as the number of wrong digits indicates.

There are several types of recurrent errors in this database of which some can be quite easily corrected. This is even more so as they are systematic or typical or 'consistent' errors, so to speak, and therefore methods to identify them and amend them can be found. Some of the most frequently met categories of errors are:

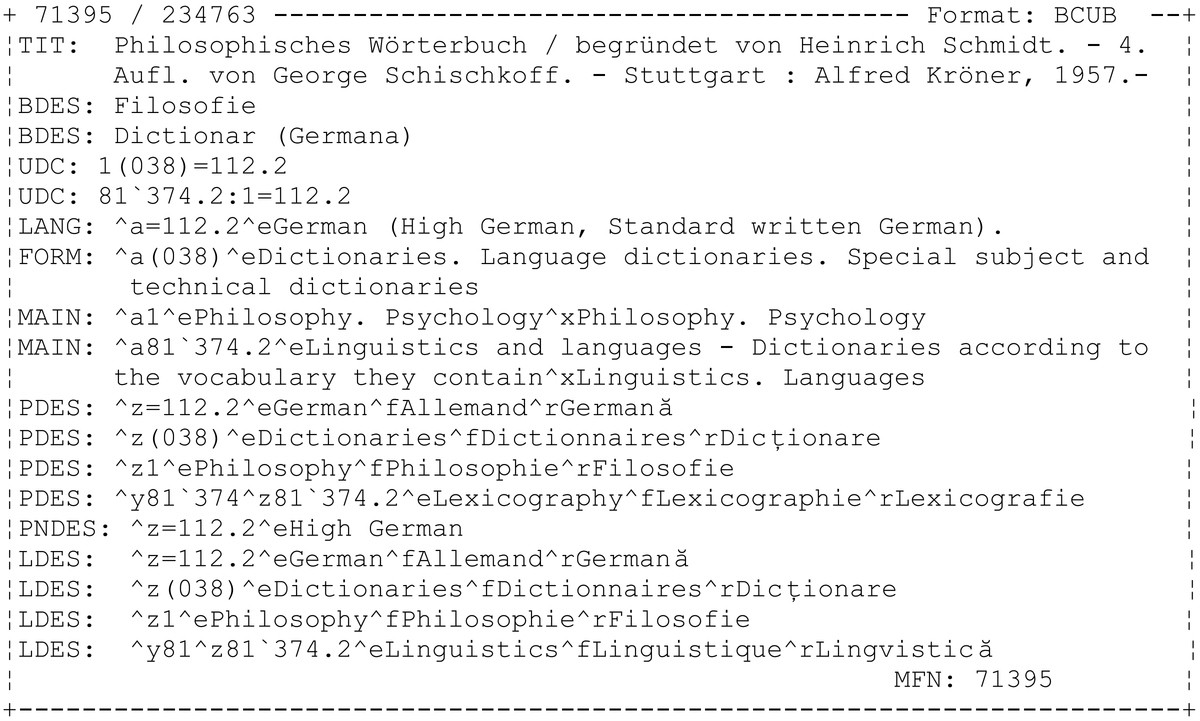
- typing errors
- errors deriving from misuse of the UDC
- errors generated by lack of update according to the E&C level in the MRF

Each identifiable type of errors will be studied and the correcting solution will be mentioned in the upcoming. Of course there are errors that cannot be automatically detected therefore they are not made visible by those marks earlier mentioned. If a change in the order of the digits composing a UDC number occurs and the resulting number has a meaning in the UDC, although in a different class, the consequences may be quite troublesome on the subject of the indexed document. Such a case will be described later in this chapter (see *Figure 42*).

```
+  3 / 176 --------------------------------------------- Format: BCUB  --+
¦TIT: A concise dictionary of law / edit. Elizabeth A. Martin. - Oxford: ¦
¦     Oxford University Press, 1988. -  394 p., 21cm. - Text pe doua     ¦
¦     coloane.- ISBN 0-19-825399-0                                        ¦
¦UDC: 801.32:34=111                                                       ¦
¦UDC: 34(03)=111                                                          ¦
¦LANG: ^a=111^eEnglish                                                    ¦
¦FORM: ^a(03)^e Reference works                                           ¦
¦MAIN: ^a801.32^eProsody. Auxiliary sciences and sources of philology -   ¦
¦       xxx ^xProsody.Auxiliary sciences and sources of philology         ¦
¦       ^oProzodie.                                                       ¦
¦      `Stiin`te auxiliare `si surse ale filologiei                       ¦
¦MAIN: ^a34^eLaw. Jurisprudence^oDrept jurisprudenţă^xLaw Jurisprudence.  ¦
¦       ^oDrept. Jurisprudenţă                                            ¦
¦PDES: ^z=111^eEnglish^fAnglais^rEngleză.                                 ¦
¦PDES: ^z(038)^eDictionaries^fDictionnaires^rDicţionare                   ¦
¦PDES: ^z34^eLaw^fDroit^rDrept                                            ¦
¦PNDES:^z34^eJurisprudence^fJurisprudence^rJurisprudenţă                  ¦
¦      ^fLégislation^rLegislaţie                                          ¦
¦LDES: ^z=111^eEnglish^fAnglais^rEngleză                                  ¦
¦LDES: ^z(038)^eDictionaries^fDictionnaires^rDicţionare                   ¦
¦LDES: ^z34^eLaw^fDroit^rDrept                                            ¦
¦LNDES:^z34^eJurisprudence^fJurisprudence^rJurisprudenţă^rLegislaţie      ¦
¦                                                             MFN: 23     ¦
+-------------------------------------------------------------------------+
```

Figure 32. Bibliographic record with text added to current and outdated UDC notations (note the xxx mark)

If we consider the example shown in *Figure 32* we can start at this point our discussion about one of the most frequent errors in Class 8 bibliographic records, i.e. invalid or outdated UDC number. However, this is only one side of the problem that we have here: ‘801.32’ is an invalid UDC number for Lexicography as before the major change stipulated in the 14th volume of Extensions and Corrections. But do we really need it at all? Is it really necessary to group all the dictionaries, particularly technical dictionaries or specialised dictionaries, in a compartment belonging to Class 8 of the classified catalogue? The answer is definitely ‘no’ to such a question. Not in an online catalogue that has other options that can successfully be used for that purpose. An additional descriptor for the bibliographic form of the document to accompany the descriptor that stands for the subject itself would do. In our case, ‘Law’ and ‘Dictionary’ might be sufficient and perhaps a language mention might have accomplished the correct representation of the subject of our document. The reason for using ‘801.32:0/9’ for all kinds of dictionaries was the necessity to have them all reflected in one place in the systematic card catalogue. So they were all put in their right place according to their contents but also in a separate section for dictionaries in the systematic catalogue. The practice is still kept nowadays yet practised with a different UDC number i.e. ‘81’374.2:0/9’ (*Figure 33*).



```
+ 71395 / 234763 ---------------------------------------- Format: BCUB  --+
¦TIT:  Philosophisches Wörterbuch / begründet von Heinrich Schmidt. - 4.  ¦
¦      Aufl. von George Schischkoff. - Stuttgart : Alfred Kröner, 1957.-  ¦
¦BDES: Filosofie                                                          ¦
¦BDES: Dictionar (Germana)                                                ¦
¦UDC: 1(038)=112.2                                                        ¦
¦UDC: 81`374.2:1=112.2                                                    ¦
¦LANG: ^a=112.2^eGerman (High German, Standard written German).           ¦
¦FORM: ^a(038)^eDictionaries. Language dictionaries. Special subject and  ¦
¦       technical dictionaries                                            ¦
¦MAIN: ^a1^ePhilosophy. Psychology^xPhilosophy. Psychology                ¦
¦MAIN: ^a81`374.2^eLinguistics and languages - Dictionaries according to  ¦
¦      the vocabulary they contain^xLinguistics. Languages                ¦
¦PDES: ^z=112.2^eGerman^fAllemand^rGermană                                ¦
¦PDES: ^z(038)^eDictionaries^fDictionnaires^rDicționare                   ¦
¦PDES: ^z1^ePhilosophy^fPhilosophie^rFilosofie                            ¦
¦PDES: ^y81`374^z81`374.2^eLexicography^fLexicographie^rLexicografie      ¦
¦PNDES: ^z=112.2^eHigh German                                             ¦
¦LDES:  ^z=112.2^eGerman^fAllemand^rGermană                               ¦
¦LDES:  ^z(038)^eDictionaries^fDictionnaires^rDicționare                  ¦
¦LDES:  ^z1^ePhilosophy^fPhilosophie^rFilosofie                           ¦
¦LDES:  ^y81^z81`374.2^eLinguistics^fLinguistique^rLingvistică            ¦
¦                                                         MFN: 71395      ¦
+-------------------------------------------------------------------------+
```

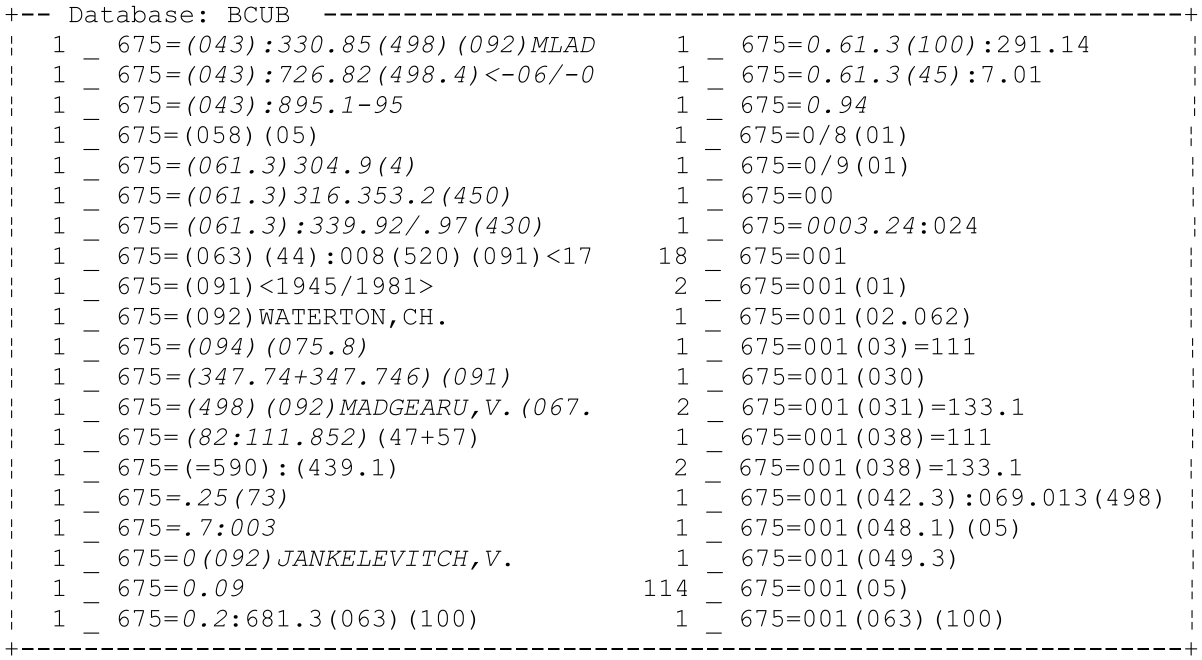
Figure 33. Bibliographic record with UDC text, UDC-based descriptors and non-descriptors and descriptors (subject headings) assigned during the indexing process in the BCUB database

The benefit of having this outdated UDC code in the experimental database is that we can find out all the specialised dictionaries dating from that time when traditional thinking of the classified catalogue was not abandoned yet. Not only can we identify them but we can also correct them. Out of 165305 bibliographic records in the database, a number of 56889 are Class 8 records, which represent 34.41% of the total. Of these, 176 records have a ‘801.32’ number as resulted from our search. Having found these records we can afford to handle them according to our purposes.

The text added to the studied UDC number *801.32* and mentioned in subfield ^e of field 709 i.e. *Prosody. Auxiliary sciences and sources of philology* might at least confuse the potential user especially when the text is found in documents dealing with a variety of subjects like: parasitology, sciences, law, children’s literature, art etc. as the first of the retrieved records do. The program automatically adds this text given the algorithm finds it by looking at the digits in the UDC notation one by one: the last digit that makes sense to the program when matching the notation found in the record with the text of the UDC MRF is the third. Therefore this text is corresponding to class number 801 in the MRF. The program prompts us that two digits are the wrong ones here by the ***[~~]*** inserted in the text.

But as previously argued, records with erroneous classification notations can all be selected and corrected. The correcting methods shall be described in the forthcoming.

In order to retrieve the erroneous UDC numbers several methods of investigating the experimental database can be used. The easiest though most cumbersome way is to browse all UDC numbers by using the ‘675=’ search option. This is mostly the case when the searcher does not know the number he is looking for (option T: display terms dictionary in CDS/ISIS). Each faulty UDC number is selected and the correction procedure is followed within each bibliographic record at a time. The large amount of time it takes and the painstaking work it involves make this method unproductive, therefore not recommendable. *Figure 34* illustrates the first screen displayed as result of using the above mentioned search key.



```
+-- Database: BCUB  ------------------------------------------------------+
¦  1 _ 675=(043):330.85(498)(092)MLAD     1 _ 675=0.61.3(100):291.14      ¦
¦  1 _ 675=(043):726.82(498.4)<-06/-0     1 _ 675=0.61.3(45):7.01         ¦
¦  1 _ 675=(043):895.1-95                 1 _ 675=0.94                    ¦
¦  1 _ 675=(058)(05)                      1 _ 675=0/8(01)                 ¦
¦  1 _ 675=(061.3)304.9(4)                1 _ 675=0/9(01)                 ¦
¦  1 _ 675=(061.3)316.353.2(450)          1 _ 675=00                      ¦
¦  1 _ 675=(061.3):339.92/.97(430)        1 _ 675=0003.24:024             ¦
¦  1 _ 675=(063)(44):008(520)(091)<17    18 _ 675=001                     ¦
¦  1 _ 675=(091)<1945/1981>               2 _ 675=001(01)                 ¦
¦  1 _ 675=(092)WATERTON,CH.              1 _ 675=001(02.062)             ¦
¦  1 _ 675=(094)(075.8)                   1 _ 675=001(03)=111             ¦
¦  1 _ 675=(347.74+347.746)(091)          1 _ 675=001(030)                ¦
¦  1 _ 675=(498)(092)MADGEARU,V.(067.     2 _ 675=001(031)=133.1          ¦
¦  1 _ 675=(82:111.852)(47+57)            1 _ 675=001(038)=111            ¦
¦  1 _ 675=(=590):(439.1)                 2 _ 675=001(038)=133.1          ¦
¦  1 _ 675=.25(73)                        1 _ 675=001(042.3):069.013(498) ¦
¦  1 _ 675=.7:003                         1 _ 675=001(048.1)(05)          ¦
¦  1 _ 675=0(092)JANKELEVITCH,V.          1 _ 675=001(049.3)              ¦
¦  1 _ 675=0.09                         114 _ 675=001(05)                 ¦
¦  1 _ 675=0.2:681.3(063)(100)            1 _ 675=001(063)(100)           ¦
+-------------------------------------------------------------------------+
```

Figure 34. Display of search result used in retrieving erroneous UDC numbers

Another way of finding the erroneous UDC numbers in the database is specific for the situation in which the searcher knows the expression of that number (option S: search formulation in CDS/ISIS). That being the case the searcher formulates the search expression like this: “675=…” filling after the = sign the precise UDC number searched for e.g.

```
Search expression?
"675=801.32"
```

The search result will be given in the following way:

```
Set   6: "675=801.32"
P=       1   675=801.32
T=       1 - #6: 675=801.32
```

where P represents the number of postings and T the number of records retrieved. This will be followed by the display of the only one bibliographic record retrieved (*Figure 35*). If in most cases this UDC notation is more or less legitimately assigned to the specialised dictionaries, in this particular one it really represents the subject of the document[19].

```
+  1 / 1 ------------------------------------------------ Format: BCUB  --+
¦TIT:  Dictionaries / Kenneth Whittaker. - London, Clive Bingley, 1966. - ¦
¦      88p., 22cm. -  The readers guide series. -  Index p.80-88          ¦
¦UDC:  801.32                                                             ¦
¦MAIN: ^a801.32^eProsody. Auxiliary sciences and sources of philology[~~] ¦
¦      xxx^xProsody. Auxiliary sciences and sources of philology          ¦
¦      ^oProzodie. `Stiin`te auxiliare `si surse ale filologiei           ¦
¦                                                            MFN: 35591   ¦
+-------------------------------------------------------------------------+
```

Figure 35. Display of search result given a particular UDC number as search expression

When the search expression is a right truncated UDC number the result of the search is in most cases a greater number of bibliographic records. In our example the search expression was found in 177 postings, which are distributed in the 176 records that we have already discussed about.

```
Search expression?
"675=801.32$"
P=      176 - #177: 675=801.32$
T=      176 - #1: #177
```

There is another way of making a search when a given number of digits in the UDC notation is known. In this expression 675 is the field number in which the expression is found and the digit after the dot represents the number of characters the UDC notation starts with i.e.

```
Search expression?
? v675*0.6 : '801.32'

    +---MFN---+         +---Hits--+        +----%----+          +---Recs--+
    ¦  231411 ¦         ¦   177   ¦        ¦   0.08  ¦          ¦  231411 ¦
    +---------+         +---------+        +---------+          +---------+
```

The advantage of this search method is that it allows for the formulation of some conclusions. For example here we can reckon with the percentage of this UDC notation in the total number of bibliographic records in the database. We can also remark that this notation is mostly found in the first part of the database, given its condition of outdated UDC notation.

The correction of invalid UDC notations can be done in two ways in our database:

a. single or individual changes
b. global changes

For the *single type of changes* or corrections of UDC numbers, from the main menu of CDS/ISIS one should type E - ISISENT for Data entry services and hence go to Data Base Maintenance. Then type W, select the worksheet called BCUBT and type E that is asking for the MFN of the bibliographic record with an incorrect UDC number. From this moment on it makes a difference whether the correction applies to a main number or to an auxiliary one.

In order to correct a main number one should go to field 675 in the bibliographic record and change that number. Then go to the main menu, select A - ISISPAS for Advanced programming services and run the program named UDCCON. When the MFN of the bibliographic record is asked it has to be typed in and then by pressing ‘return’ the change in text is automatically made. As to an auxiliary number, the same procedure should be followed.

[19] In the old UDC the meaning of ‘801.32’ was Word presentation and arrangement (in dictionaries, encyclopaedias etc.)

An illustration of the mentioned procedures is given below by means of an example of a record from the database before and after the change in the UDC notation. The example shows a bibliographic record with an invalid UDC number changed into a valid one according to the restructured table of common auxiliaries of language (*Figures 36 - 37*). Our point here is to show the confusing way the added text looks in field 709 [MAIN:] before the change and the difference this change made on the updated bibliographic record.

```
+--1334 / 231411 --------------------------------------- Format: BCUB  --+
¦TIT:  Hainele cele noi ale imparatului / H.C. Andersen. - Bucuresti,    ¦
¦      Evenimentul, 1990. -  16p., il. color, 23cm . -  Mythos. -        ¦
¦UDC:  839.8-34=135.1                                                    ¦
¦LANG: ^a=135.1^eRomanian.                                               ¦
¦MAIN: ^a839.8-34^eLanguage. Linguistics. Literature [~~~~~] - xxx       ¦
¦       xxx^xLanguage. Linguistics. Literature                           ¦
¦PDES: ^z=135.1^eRomanian^fRoumain^rRomană.                              ¦
¦PDES: ^y821^z821.113.4-34^eLiteratures of individual languages          ¦
¦      ^fLittératures relatives à des langues particulières              ¦
¦      ^rLiteratura limbilor individuale                                 ¦
¦PDES: ^z82-34^eTales^fContes^rPoveşti                                   ¦
¦PDES: ^z=113.4^eDanish^fDanois^rDaneză                                  ¦
¦PNDES:^z=135.1^eRumanian.                                               ¦
¦PNDES:^y82-3^z82-34^eProse narrative                                    ¦
¦PNDES:^y=113^z=113.4^eNordic languages^fLangues nordiques^rLimbi nordice¦
¦LDES: ^z=135.1^eRomanian^fRoumain^rRomâna.                              ¦
¦LDES: ^z821.113.4-34^eFrench literature^fLittérature                    ¦
¦      française^rLiteratură franceză                                    ¦
¦LDES: ^z82-34^eTales^fContes^rPoveşti                                   ¦
¦LDES: ^z=113.4^eDanish^fDanois^rDaneză                                  ¦
¦LDES: ^y82^z821^eLiterature^fLittérature^rLiteratură.                   ¦
¦LNDES:^y82-3^z82-34^rProză scurtă.                                      ¦
¦                                                           MFN: 1334    ¦
+------------------------------------------------------------------------+
```

Figure 36. Bibliographic record before modification of the UDC main number

```
+--1334 / 231411 --------------------------------------- Format: BCUB  --+
¦TIT:  Hainele cele noi ale imparatului / H.C. Andersen. - Bucuresti,    ¦
¦      Evenimentul, 1990. -  16p., il. color, 23cm . -  Mythos. -        ¦
¦UDC:  821.113.4-34=135.1                                                ¦
¦LANG: ^a=135.1^eRomanian.                                               ¦
¦MAIN: ^a821.113.4-34^eDanish^oDaneză - Tales^gErdichtete                ¦
¦      Erzählungen. Legenden. Kunstmärchen^xLiterature^oLiteratură       ¦
¦PDES: ^z=135.1^eRomanian^fRoumain^rRomană                               ¦
¦PDES: ^y821^z821.113.4-34^eLiteratures of individual languages          ¦
¦      ^fLittératures relatives à des langues particulières              ¦
¦      ^rLiteratura limbilor individuale                                 ¦
¦PDES: ^z82-34^eTales^fContes^rPoveşti                                   ¦
¦PDES: ^z=113.4^eDanish^fDanois^rDaneză                                  ¦
¦PNDES:^z=135.1^eRumanian                                                ¦
¦PNDES:^y82-3^z82-34^eProse narrative                                    ¦
¦PNDES:^y=113^z=113.4^eNordic languages^fLangues nordiques^rLimbi nordice¦
¦LDES: ^z=135.1^eRomanian^fRoumain^rRomâna                               ¦
¦LDES: ^z821.113.4-34^eFrench literature^fLittérature                    ¦
¦      française^rLiteratură franceză                                    ¦
¦LDES: ^z82-34^eTales^fContes^rPoveşti                                   ¦
¦LDES: ^z=113.4^eDanish^fDanois^rDaneză                                  ¦
¦LDES: ^y82^z821^eLiterature^fLittérature^rLiteratură                    ¦
¦LNDES:^y82-3^z82-34^rProză scurtă                                       ¦
¦                                                           MFN: 1334    ¦
+------------------------------------------------------------------------+
```

Figure 37. Bibliographic record after modification of the outdated UDC number

The *global changes* are carried through a program called GLOB that has to be run after a query has been formulated. Once the program is started the options for the following operations are given:

FIELD OPERATIONS

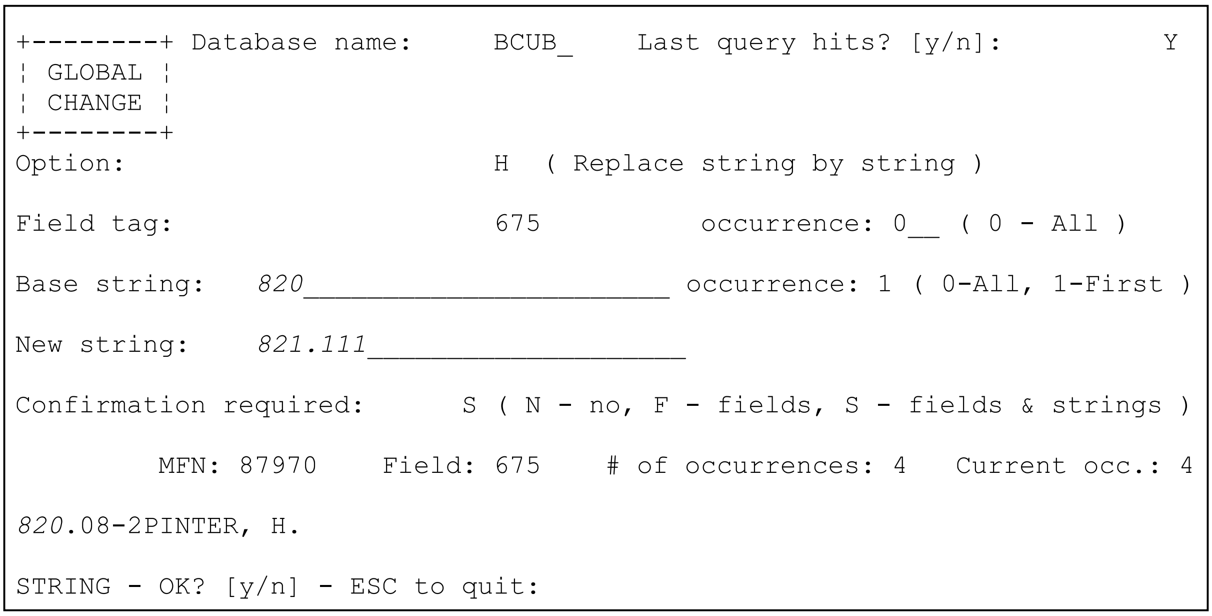
A - insert a new occurrence of the field after the current occurrence
B - insert a new occurrence of the field before the current occurrence
C - replace the current occurrence of the field
D - delete the current occurrence of the field
E - edit the current occurrence of the field

STRING OPERATIONS

F, J - insert string after the source string
G, K - insert string before the source string
H, L - replace the source string
I, M - delete the source string

F - I : source string
defined by content
J - M : source string
defined by position

The most convenient of these options that works for our purposes is option H. We can efficiently use it in order to replace the old strings of digits representing languages and literatures with the updated ones according to the MRF. This way the old UDC numbers in the query hits that gave errors in field 709 for text added to main numbers are replaced automatically within reasonably short time in each bibliographic record at a time as it is shown in *Figure 38*. The GLOB program is run separately for each field where the outdated UDC numbers exist. Any other information having been included in any of the fields of the bibliographic records can be modified according to the user's needs.



```
+--------+ Database name:     BCUB_    Last query hits? [y/n]:        Y
¦ GLOBAL ¦
¦ CHANGE ¦
+--------+
Option:                       H  ( Replace string by string )

Field tag:                    675          occurrence: 0__  ( 0 - All )

Base string:   820_______________________ occurrence: 1 ( 0-All, 1-First )

New string:    821.111____________________

Confirmation required:       S ( N - no, F - fields, S - fields & strings )

         MFN: 87970    Field: 675    # of occurrences: 4   Current occ.: 4

820.08-2PINTER, H.

STRING - OK? [y/n] - ESC to quit:
```

Figure 38. First screen of the GLOB program for global changes in CDS/ISIS

The start of the global change operations is made by running the GLOB program for field 675. Then comes field 709 for the same string of digits. When the new string of digits is placed in field 709, program us709 is run in order to add new text to the updated UDC notation. This is done automatically in each record belonging to the last query hits. The same procedure can be followed for each of the fields that need correction.

6.5.2 Mapping the UDC numbers with UDC descriptors

If we consider *Figure 33* we have before us the image of a most complex bibliographic record in that it has several alternative forms of subject representation. Chronologically speaking, they are: UDC notations, manually assigned descriptors, text added to the complex UDC notations after they were separated into their parts and finally, UDC descriptors.

The automated allocation of descriptors to the bibliographic records in the database adds a new dimension to the information retrieval facilities and enlarges the accessibility in terms of user-friendliness. Experience with Class 8 – Linguistics. Literatures has demonstrated that in most of the cases descriptors are not correctly assigned if only taken over from the multilingual thesaurus and added to the bibliographic records. Class 8 having a faceted structure, many of the class numbers are constructed by synthesis and therefore the UDC notation used in subject representation is not always present in the UDC table as such. If we only think of all the individual languages and literatures we can say that all linguistic aspects and all literary genres require preceding alteration in order to be assigned to a document according to the UDC grammar. In a UDC-based thesaurus built for Class 8 only the common auxiliaries of language in Table 1c[20] will take a considerable part of the total size of it all. About 1,360 language codes would have to be repeated in different combinations having as result a particular language, a particular literature and the language of a particular document whose subject is denoted by a main number (see also **§5.4**). Given this multitude of languages and literatures that are likely to occur in the subjects of documents the sizes of a thesaurus should be huge in order to cover them all. For this reason it is necessary that the language category in any of its occurrences be treated as a separate descriptor and used post-coordinately in searching. The application of the algorithms for the decomposition of the UDC complex notations into their component parts might well do the job here, the complete meaning of the UDC number being given by a prescribed combination of descriptors (see *Figure 33*).

The next example of bibliographic record (*Figure 39*) shows how the subject of the same document can be treated by different indexing languages. Primarily the subject was classified under both Classical Latin literature and under Philosophy. Indeed a writer like Cicero can hardly be considered as belonging to only one of these two domains. But this is beyond our concern now. We are now interested in how these two facets of the subject complete each other in all the indexing methods.

A careful look at the captions the UDC numbers have – according to the MRF – enables us to conclude that with one exception, all four types of subject representation say almost the same about the subject of our document. The language of this particular literature makes the exception. Classical Latin being ‘the missing piece’ of the ‘puzzle’ we can come without hesitation to the conclusion that languages do need a special consideration. Particularly those that are not included in the MRF as examples of subdivisions of Class 811 Languages as subjects and Class 821 Literatures of individual languages along with Table 1c for common auxiliaries of language. The existing thesaurus for Class 8 (like all the others mentioned previously – see **§5.3**) that we use in our case study was not intended for automated use. It was built in order to be used manually in the first place by way of assigning descriptors from the printed thesaurus to the bibliographic records during the indexing process and secondly to

[20] Table 1c is the main place in the UDC tables for enumeration of languages, and serves as the source for the subdivision of class ‘811’ Languages (as subjects of study), class ‘821’ Literatures of individual languages, and (=...) Table 1f – Common auxiliaries of ethnic grouping

enable guided search in the online catalogue. This explains why the Romanian descriptor existing in the database is complete, Literatura latina clasica, as it is in the printed thesaurus, and the descriptor based on the same UDC number but automatically assigned is lacking this piece of information i.e. Classical Latin literature.

Another difference in subject indexing is that the degree of coordination is higher in the existing descriptors than in the automatically assigned ones in Class 1 – Philosophy e.g. Destin (Filosofie). Likewise, there is a higher compatibility between the UDC numbers in the MRF and the UDC-based multilingual descriptors than between the class numbers and manually assigned descriptors existing in the VUBIS database e.g.: *Metempsihoza* (129) vs. *Metafizica* (11).

```
+    26 / 36 -------------------------------------------- Format: BCUB --+
¦TIT:  Traité du destin / Cicéron ; texte établi et trad. par Albert Yon. ¦
¦      - 5-e tirage. - Paris : Les Belles Lettres, 1991. - 1 vol. în pag. ¦
¦      multipla ; 20 cm. - Contine index. - Texte paralele în lb. la si fr¦
¦      . - 2-251-01081-5                                                  ¦
¦BDES:  Literatura latina clasica                                         ¦
¦BDES:  Scrieri filosofice                                                ¦
¦BDES:  Metafizica                                                        ¦
¦BDES:  Destin (Filosofie)                                                ¦
¦UDC:   124.6                                                             ¦
¦UDC:   129                                                               ¦
¦UDC:   821.124`02-96                                                     ¦
¦UDC:   821.124`02-96=133.1                                               ¦
¦LANG: ^a=133.1^eFrench.                                                  ¦
¦MAIN: ^a124.6^eDetermination. Destiny^xPhilosophy. Psychology            ¦
¦MAIN: ^a129^eOrigin and destiny of the individual soul. Transmigration of¦
¦       souls. Metempsychosis. Incarnation. Reincarnation.                ¦
¦       Immortality^xPhilosophy. Psychology                               ¦
¦MAIN: ^a821.124`02^eLatin - Origins and periods of languages. Phases of  ¦
¦       development^xLiterature                                           ¦
¦MAIN: ^a821.124-96^eLatin - Works of science and philosophy as           ¦
¦       literature^xLiterature                                            ¦
¦PDES: ^z=133.1^eFrench^fFrançais^rFranceză                               ¦
¦PDES: ^z124.6^eDestiny^fDestiné^rDestin                                  ¦
¦PDES: ^z129^eMetempsychosis^fMétempsycose^rMetempsihoză                  ¦
¦PDES: ^y821^z821.124`02^eLiteratures of individual languages             ¦
¦       ^fLittératures relatives à des langues particulières^rLiteratura  ¦
¦       limbilor individuale                                              ¦
¦PDES: ^y82^z82`02^eLiterature^fLittérature^rLiteratură                   ¦
¦PDES: ^z=124^eLatin^fLatin^rLatină                                       ¦
¦PDES: ^z821^eLiteratures of individual languages^fLittératures relatives ¦
¦       à des langues particulières^rLiteratura limbilor individuale      ¦
¦PDES: ^y821^z821.124-96^eLiteratures of individual anguages^fLittératures¦
¦       relatives à des langues particulières^rLiteratura limbilor        ¦
¦       individuale                                                       ¦
¦PDES:  ^y82^z82-96^eLiterature^fLittérature^rLiteratură                  ¦
¦LDES:  ^z=133.1^eFrench^fFrançais^rFranceză                              ¦
¦LDES:  ^y1^z124.6^ePhilosophy^fPhilosophie^rFilosofie                    ¦
¦LDES:  ^y1^z129^ePhilosophy^fPhilosophie^rFilosofie                      ¦
¦LDES:  ^z821.124`02^eFrench literature^fLittérature française^rLiteratură¦
¦       franceză                                                          ¦
¦LDES:  ^y82^z82`02^eLiterature^fLittérature^rLiteratură                  ¦
¦LDES:  ^z=124^eLatin^fLatin^rLatină                                      ¦
¦                                                          MFN: 77281     ¦
+-------------------------------------------------------------------------+
```

Figure 39. Bibliographic record with automatically assigned descriptors

Our next example (*Figure 40*) shows full compatibility between all information languages used. The difference from the earlier one is that the English literature is among the few individual literatures given as such in the MRF. This is just a simple example of subject representation. The more complicated the subject the bigger the complexity of problems.

```
+     5 / 16 ----------------------------------------- Format: BCUB   ---+
¦TIT:  Moulds of understanding: a pattern  of natural philosophy / Joseph ¦
¦      Needham, ed. and introd. by Gary Werskey. - London, George Allen & ¦
¦      Unwin, 1976. -  320p., 22cm. -  Note dupa capitole. -  Bibliogr. si¦
¦      index p. 305-320. - ISBN 0-04-925011-6                              ¦
¦DES:  Metafizica                                                          ¦
¦DES:  Literatura engleza                                                  ¦
¦DES:  Scrieri filosofice                                                  ¦
¦UDC:  11                                                                  ¦
¦UDC:  821.111-96                                                          ¦
¦MAIN: ^a11^eMetaphysics^xMetaphysics                                      ¦
¦MAIN: ^a821.111-96^eEnglish - Works of science and philosophy as          ¦
¦       literature^xLiterature                                             ¦
¦PDES: ^z11^eMetaphysics^fMétaphysique^rMetafizică                         ¦
¦PDES: ^y821^z821.111-96^eLiteratures of individual anguages^fLittératures¦
¦       relatives à des langues particulières^rLiteratura limbilor         ¦
¦       individuale                                                        ¦
¦PDES: ^y82^z82-96^eLiterature^fLittérature^rLiteratură                    ¦
¦PDES: ^z=111^eEnglish^fAnglais^rEngleză                                   ¦
¦LDES: ^z11^eMetaphysics^fMétaphysique^rMetafizică                         ¦
¦LDES: ^z821.111-96^eEnglish literature^fLittérature anglaise^rLiteratură ¦
¦       engleză                                                            ¦
¦LDES: ^y82^z82-96^eLiterature^fLittérature^rLiteratură                    ¦
¦LDES: ^z=111^eEnglish^fAnglais^rEngleză                                   ¦
¦LDES: ^y82^z821^eLiterature^fLittérature^rLiteratură                      ¦
¦                                                              MFN: 28077 ¦
+-------------------------------------------------------------------------+
```

Figure 40. Example of total compatibility between the UDC notations and descriptors in Class 8

Hence some conclusions can be formulated regarding issues of *recall* and *precision*, *compatibility* and *complementarity* of the four types of indexing languages under consideration:

1. The monolingual descriptors manually assigned at the moment of indexing can hardly meet the requirements of *consistency* and *control*.

Discussion:

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 (Frâncu, 2002).

```
159.923    Type psychology. Individual psychology. Psychology of
           individualities. Individuality. Personality. Character
           psychology. Characterology. Idiosyncrasies. Personal
           equation. Personality types
159.923.3  Composition of the personality. Character traits. Psychogram
159.925    Study of expression. Physical manifestation of mentality.
           Bodily expression of character
```

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

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

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

Obviously the diversity of manually assigned descriptors can confuse the searcher not only because of total inconsistency in the indexing terms assigned for similar subjects 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 'healthier' way to ensure consistency in indexing is to map univocal descriptors onto UDC notations. By such a procedure the control on terms is (automatically) imposed and this is specifically what our approach is about.

Some more examples of inconsistencies in classification and indexing and the consequences they have on information retrieval will emphasize the above-mentioned statement. We give as example the way the International Labour Organization (ILO) was represented in UDC codes in several documents in the VUBIS bibliographic database of the Central University Library of Bucharest. To begin with, consider the MRF record below. Note that there is an application note and a combination example in the MRF that give just the International Labour Office as example:

```
MFN: 182522
             UDC number: 331.07
                  Table: M
      Special aux. type: B
            Description: ^eAdministrative organizations, authorities and official
                         bodies in the field of labour
       Application note: ^eAuxiliaries <3.07...> (as at <35.07>)
Combination examples: ^y(100)ILO^dInternational Labour Office
```

The correct combination resulting from the MRF instructions will be: *331.07(100)ILO* (see the subject representation in the bibliographic record under Title 4). So let us see the way the ILO was represented in the classified catalogue. The search query used was the name of the institution *International Labour Organization (ILO).* The asterisk (*) points to the different ways the UDC combination was interpreted.

Title 1:

Record of proceedings : International Labour Conference, eighty-ninth session, Geneva, 2001. - Geneva : International Labour Office, 2001. - vol. ; 29 cm

UDC notations:
061.1ILO:331.91(063)(100)(047)*
331.91:061.1ILO(063)(100)(047)
061.3(494):351.83(047)
351.83(063)(100)(047)

Descriptors:
1 Organizarea internationala a muncii
2 Organizatii internationale
3 Dreptul muncii
4 Conferinta internationala
5 Raport

Institutions: International Labour Organization (ILO)

Title 2:

Your voice at work : global report under the Follow-up to the ILO Declaration on Fundamental Principles and Rights at Work : Report I(B) / Michel Hansenne. - Geneva : International Labour Office, 2000. - X, 88 p.

UDC notations:
061.1ILO:331.1(063)(047)*
331.1:061.1ILO(063)(047)
316.334.22(063)(047)
061.3(494):316.334.22

Descriptors:
1 Teoria muncii
2 Dreptul la munca
3 Sociologia muncii
4 Conferinta internationala
5 Raport

Institutions: International Labour Organization (ILO)

Title 3:

International labour standards : a workers'education manual / International Labour Organization. - 4th rev. ed. – Geneva : International Labour Office, 1998. - VIII, 148 p.; 24 cm

UDC notations:
061.1ILO:331(100)(075)*
006.4:331(100)(075)
331.1(100):061.1(075)
331(100)(083.74)(075)

Descriptors:
1 Normarea muncii
2 Organizatii de munca
3 Organizatii internationale
4 Manual

Institutions: International Labour Organization (ILO)

Title 4:

Codes of conduct and multinational enterprises *: [Resursa electronica]*. - Geneva : International Labour Office, 2002. - 1 disc optic (CD-ROM) ; 12 cm. - Configuratia minima necesara: Microsoft Windows 9x, Me, 2000 or NT

UDC notations:
331.07(100)ILO(086.76)*
334.726:331.07(100)ILO(086.76)
304.4:331.07(100)ILO(086.76)

Descriptors:
1 Organizatii de munca
2 Organizatii internationale
3 Economia muncii
4 Politica sociala
5 Intreprinderi multinationale
6 (CD-ROM)

Institutions: International Labour Organization (ILO)

2. The *UDC numbers used along with corresponding text and descriptors as information access facilities* have the following advantages:

- the classified catalogue collocates related subjects;
- the enumerative notation avoids language problems providing universal access;
- the classified catalogue allows convenient generation of lists of documents on a given subject area such as specialised bibliographies;
- the alphabetical list of subjects available each time a search is performed via words from the UDC text is friendlier to the searcher;

- the application of Boolean logic in both class numbers and words can modify the search result according to the user's needs;
- related subjects may be collocated and serendipity taken advantage of when truncation is used e.g. the stem *bank$* used as search element gives 'bank', 'banking', 'banks' as queries;
- all the documents relating to a specific geographic area will be retrieved without the user having to know the UDC number for that particular area (field 703); the drawback in this method is that fiction, for instance, will be found among these documents if that fiction is represented by a UDC number including an auxiliary of place (e.g. American literature, literature of multilingual countries);
- likewise all translated documents having a particular language as target language will be collocated by using field 701 (language of a document) and the name of the language as search key;
- manually assigned descriptors are not so reliable as the automatically assigned ones given the inconsistencies that may occur when different descriptors are linked with the same UDC number; the UDC-based descriptors will always be the same for a particular UDC number.

3. Despite all inconsistencies, there is a reasonably high *precision rate* in the manually assigned descriptors given (to a certain extent) the precoordination of terms;

4. However, the *recall rate* is quite low particularly because of the information loss generated by poor consistency in the use of descriptors – sometimes the same UDC notation may have as counterpart completely different descriptors (see the discussion);

5. This drawback is balanced by the use of the automatically assigned multilingual descriptors that will always be the same for a UDC number, hence the high *recall rate*;

6. There is a higher *compatibility* degree between the UDC notations in the bibliographic description and the UDC based descriptors, hence the advantages of consistency and correctness in indexing and control on terms;

7. What is still problematic is the *level of specificity* of the multilingual thesaurus that has to be as close as possible to the specificity of the classified catalogue, or else we deal with information loss (compare the number of bibliographic records with automatically assigned descriptors – 156618 as against the total number of bibliographic records – 165370 shown in the table at page 93);

8. If all the modes of subject representation are considered we have a complete image of the subject of the document given the *complementarity* of the indexing languages used; the stress should be laid here on the possibility of the UDC to express via alphabetical extension sides of the subject that cannot be expressed in a controlled language like a thesaurus (unique entities); this is also true for the auxiliaries of time except for those provided by the MRF.

6.5.3 Making multilingual subject headings available

Multilingual descriptors are assigned to bibliographic records in as much as their UDC source number is found in the MRF and there are descriptors available in more than one language to be imbedded in the database. This statement has far going consequences on our experimental database in general and on information retrieval in particular.

Reference was made in the previous paragraph to the automatic assignment of descriptors and a few hints were made on the multilingual aspects involved by the existing three language descriptors in Class 8 (cf. Frâncu, 1999) of the database. What we are now going to talk about is the way the multilingual interdisciplinary thesaurus described earlier is acting in the bibliographic database environment, in other words how can the multilingual descriptors be inserted in the bibliographic records and then used in information retrieval.

It is convenient at this point to consider the operations undertaken before the multilingual descriptors are available in the bibliographic records of the database. The first step taken is the selection of only three fields out of the thesaurus database built in MTM3, the program specially designed for multilingual thesaurus construction in CDS/ISIS format (see **§5.3**). These fields contain the UDC numbers, the descriptors based on them and the non-descriptors. This is a temporary database including all the three languages: English, French and Romanian. Next, this database is merged with the UDC MRF to the purpose that the UDC numbers as such are mapped onto the UDC-based multilingual descriptors whenever such descriptors exist. This operation is done through a very sophisticated program, which is not of our purpose to discuss here. We give in *Figure 41* below one record from this MRF database with a field (170) specially designed for LTHES descriptors (see *Appendix 1*).

```
+     1 / 1 -------------------------------------------- Format: MRF -----+
¦ MFN: 197940                                                             ¦
¦       UDC number : 582                                                  ¦
¦            Table : M                                                    ¦
¦       Description: ^eSystematic botany                                  ¦
¦       Scope note : ^eFor systematic palaeobotany see 561 which is       ¦
¦                    parallel with 582                                    ¦
¦ Lthes Descriptors: ^eSystematic botany^fBotanique systématique          ¦
¦                    ^rBotanică sistematică                               ¦
+-------------------------------------------------------------------------+
```

Figure 41. Example of an MRF record after the multilingual descriptors were mapped onto the UDC number

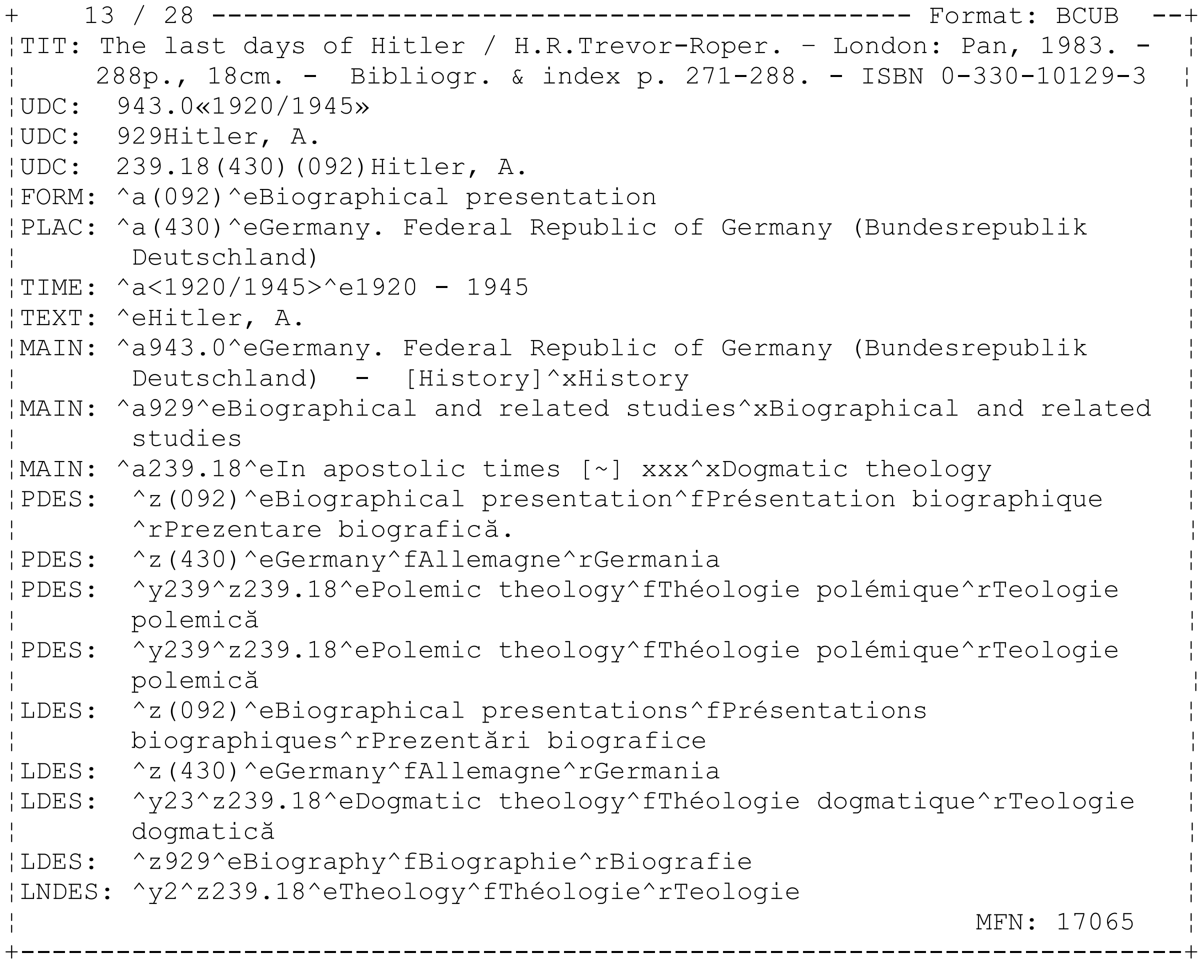
Once the multilingual descriptors are inserted in the MRF, the next thing to do is to merge it with the bibliographic database. This is the third and final step the result of which is the automatic allocation of multilingual descriptors throughout the database in its totality.

Our main concern at this point is to see what the consequences of these three operations are on our experimental database.

One searcher may want to have descriptors in only one language displayed on the screen. The only thing to do is to use LMB followed by the initial of the search language: E, F or R and his wish is accomplished. Thus for English descriptors he will use LMBE. Likewise for French the search starts from the code LMBF and for Romanian from LMBR e.g.

```
   2 _ LMBE=BIOGAS AS FUEL      2 _ LMBF=BIOGAZ            2 _ LMBR=BIOGAZ
4925 _ LMBE=BIOGRAPHY        4925 _ LMBF=BIOGRAPHIE        2 _ LMBR=BIOGENEZA
 286 _ LMBE=BIOLOGY           286 _ LMBF=BIOLOGIE         16 _ LMBR=BIOGEOGRAFIE
  25 _ LMBE=BIRDS               6 _ LMBF=BLOCS          4925 _ LMBR=BIOGRAFIE
   1 _ LMBE=BOOKBINDING               D'APPARTEMENTS      30 _ LMBR=BIOGRAFII
           AND STATIONERY       1 _ LMBF=BOIS COMME      286 _ LMBR=BIOLOGIE
  10 _ LMBE=BORROWED WORDS            COMBUSTIBLE          4 _ LMBR=BIOLOGIE
   1 _ LMBE=BORROWING         108 _ LMBF=BOTANIQUE                APLICATA
 108 _ LMBE=BOTANY              6 _ LMBF=BOTANIQUE        62 _ LMBR=BIOLOGIE
   9 _ LMBE=BRIQUETTING               APPLIQUEE                   MOLECULARA
  22 _ LMBE=BUDGETS            26 _ LMBF=BOTANIQUE        34 _ LMBR=BIOLOGIE
  22 _ LMBE=BUILDING                  GENERALE                    TEORETICA
            MATERIALS          59 _ LMBF=BOTANIQUE        33 _ LMBR=BIOTOP
  57 _ LMBE=BUILDING TRADE            GEOGRAPHIQUE        98 _ LMBR=BISERICA
```

Automatically added captions (but also descriptors) to the UDC notations, in case of a typing mistake in a particular notation, may have embarrassing consequences on information retrieval in that it can produce confusion like in the following example. The error here consists of a switch between two digits in the UDC notation i.e. 239.18 *Dogmatic theology,* instead of 329.18 *Fascist attitude* (*Figure 42*). The beneficial side of such a situation is that the natural language text can make the identification of a classification (here typing) mistake much more easily noticeable than the classification notation.



```
+    13 / 28 ------------------------------------------ Format: BCUB  --+
¦TIT: The last days of Hitler / H.R.Trevor-Roper. - London: Pan, 1983. -  ¦
¦     288p., 18cm. -  Bibliogr. & index p. 271-288. - ISBN 0-330-10129-3  ¦
¦UDC:  943.0«1920/1945»                                                   ¦
¦UDC:  929Hitler, A.                                                      ¦
¦UDC:  239.18(430)(092)Hitler, A.                                         ¦
¦FORM: ^a(092)^eBiographical presentation                                 ¦
¦PLAC: ^a(430)^eGermany. Federal Republic of Germany (Bundesrepublik      ¦
¦       Deutschland)                                                      ¦
¦TIME: ^a<1920/1945>^e1920 - 1945                                         ¦
¦TEXT: ^eHitler, A.                                                       ¦
¦MAIN: ^a943.0^eGermany. Federal Republic of Germany (Bundesrepublik      ¦
¦       Deutschland)  -  [History]^xHistory                               ¦
¦MAIN: ^a929^eBiographical and related studies^xBiographical and related  ¦
¦       studies                                                           ¦
¦MAIN: ^a239.18^eIn apostolic times [~] xxx^xDogmatic theology            ¦
¦PDES:  ^z(092)^eBiographical presentation^fPrésentation biographique     ¦
¦       ^rPrezentare biografică.                                          ¦
¦PDES:  ^z(430)^eGermany^fAllemagne^rGermania                             ¦
¦PDES:  ^y239^z239.18^ePolemic theology^fThéologie polémique^rTeologie    ¦
¦       polemică                                                          ¦
¦PDES:  ^y239^z239.18^ePolemic theology^fThéologie polémique^rTeologie    ¦
¦       polemică                                                          ¦
¦LDES:  ^z(092)^eBiographical presentations^fPrésentations                ¦
¦       biographiques^rPrezentări biografice                              ¦
¦LDES:  ^z(430)^eGermany^fAllemagne^rGermania                             ¦
¦LDES:  ^y23^z239.18^eDogmatic theology^fThéologie dogmatique^rTeologie   ¦
¦       dogmatică                                                         ¦
¦LDES:  ^z929^eBiography^fBiographie^rBiografie                           ¦
¦LNDES: ^y2^z239.18^eTheology^fThéologie^rTeologie                        ¦
¦                                                          MFN: 17065     ¦
+-------------------------------------------------------------------------+
```

Figure 42. Example of the significant change in the meaning of a UDC number caused by a typing mistake

6.6 Is the information retrieval enhanced? If so, to what extent?

The ultimate goal of our case study is to demonstrate that the information retrieval is improved by means of tests done in the experimental database. For this purpose we start from real-life examples of queries and execute searches in the database to see the result. In the end we need to have a multiple choice of search methods so that accessibility of information is capable to satisfy a large scale of users with different degrees of expertise in searching.

Think about any of the bibliographic records from the experimental database and in which all possible subject fields are correctly represented. There are fields containing the subject matter of the document as they existed in the database before our experiment (UDC notations and descriptors in Romanian) on one hand and in addition to those there are other fields containing parts of the complex UDC notations and their corresponding text as of the MRF plus multilingual descriptors on the other hand.

The potential user is offered multiple options to choose from and use in searching the database. Some of them are meant for expert searchers others are not. The queries can be formulated by using the T option (Display terms dictionary) in Information retrieval services according to the Field Select Table (FST). The search codes fall into two groups: search codes addressing the descriptive part of the bibliographic records in the database and search codes addressing the subjects of the documents in the database.

Since the first group of codes concerns the descriptive part of the records they refer to either the title or the whole body of bibliographic description (see *Appendix 1* for more details):

TI = Title of the document
KW = Words from the bibliographic description (title included).

The second grouping of search codes can be divided further in three categories. In the first place, this category of searches regards the UDC numbers and they give the knowledgeable searchers several possibilities to formulate their queries such as:

NC = UDC number in the MRF
NK = UDC number in the shortened version of MRF
675 = UDC number in the bibliographic record in the database
701 = Auxiliaries of language
702 = Auxiliaries of bibliographic form
703 = Auxiliaries of place
704 = Auxiliaries of ethnic grouping and nationality
705 = Auxiliaries of time
706 = Alphabetic addition
707 = Auxiliaries of point of view
708 = Auxiliaries for material / persons
709 = UDC main numbers

As previously stated the 700 fields were especially created to enable searches via parts of the decomposed UDC complex notations.

The next category of search codes is closely related to the above-mentioned one. These search codes are addressing the MRF part of the experimental database:

UE = UDC description (caption) in the MRF in English
UR = UDC description (caption) in the MRF in Romanian
DPE/DPF/DPR = UDC-based descriptors in English/French/Romanian from PTHES in MRF
NPE/NPF/NPR = non-descriptors in English/French/Romanian from PTHES in MRF
DLE/DLF/DLR = UDC-based descriptors in English/French/Romanian from LTHES in MRF
NLE/NLF/NLR = non-descriptors in English/French/Romanian from LTHES in MRF

Additional possibilities of search are represented by the third category of codes meant to *enhance the user-friendliness of the UDC as an information language*. These are meant for information retrieval via the three types of descriptors existing in the bibliographic database, Romanian descriptors assigned manually by indexers at the moment of subject indexing and multilingual descriptors in English, French and Romanian automatically imbedded in the database from the two thesauri. Along with those the text of the decomposed UDC notations can also be used in searching.

DU = Words from the UDC text automatically added in a 70- field in the database
DE = Romanian descriptors manually assigned to bibliographic records in the database

PMBE/PMBF/PMBR = PTHES automatically assigned descriptors in English, French and Romanian
LMBE/LMBF/LMBR = LTHES automatically assigned descriptors in English, French and Romanian
PNBE/PNBF/PNBR = PTHES automatically assigned non-descriptors in English, French and Romanian
LNBE/LNBF/LNBR = LTHES automatically assigned non-descriptors in English, French and Romanian

The above-mentioned codes for multilingual descriptors used in searching will retrieve those records whose subject was classified with a UDC number that corresponds to the user's query. For example:

```
Set    Data Base    Hits    Query element       Current Data Base name = BCUB
1      BCUB          10     "PMBR=AFGANISTAN"
2      BCUB          10     "703=(581)"
```

As shown in the simple example above the number of records retrieved by the Romanian PTHES descriptor 'Afganistan' is exactly the same as the number of records in which the corresponding UDC auxiliary of place (581) is mentioned.

The forthcoming examples of searches in the bibliographic database will illustrate the retrieval power of the search devices available.

6.6.1 Searches using words from the bibliographic description

As one will easily notice the first two search keys – TI = title and KW = words from the bibliographic description – give little problems in use (if at all) but the precision rate of the information retrieved will not be too high. The search is made in the descriptive part of the bibliographic records. The queries used are the well-known English word '*bank*' and its Romanian and French equivalents, '*bancă'* and '*banque*' respectively. Their plural forms were also used. Basically, the queries used may have the following meanings:

- Financial institution and related activities
- Institution name in the title
- Personal (author) name
- Publisher / Series title
- Computer data bases (data banks)

1. For the *English* words '*bank/banks'* the number of occurrences in bibliographic records go up 70 and their meanings are divided such as:

Financial institution	– 21 hits
Institution name in the title	– 14 hits
Personal (author) name	– 11 hits
Publisher / Series title	– 16 hits
Computer databases (data banks)	– 8 hits

Mention should be made on the presence of 4 bibliographic records in German among the English ones. In addition to that there is one of the records in which the studied word appears in the title: "*Wanganella Bank bathymetry"* with a completely different meaning from the studied one i.e. riverside.

2. For the *Romanian* words '*bancă/bănci'* the number of occurrences is 55 and their meanings belong to fewer categories such as:

Financial institution	– 15 hits (of which 10 are Romanian and 5 Italian)
Institution name in the title	– 10 hits
Publisher / Series title	– 17 hits (of which 14 are Romanian and 3 Italian)
Computer databases (data banks)	– 13 hits

Here we have also a special situation in which the meaning of the title word '*bancă'* is a place to sit: "*Calculatorul, coleg de banca"*. An example of inter-lingual cognates is found here again: the Italian common noun *banca* as title word and the same word as part of an institution name: *Banca Comerciala Italiana*, formally and semantically identical with their Romanian correspondent. It is the diacritical marks that make the difference between the two languages. If diacritics are not used then the records of both Romanian and Italian will be mixed up in the search result.

Other grammatical variants of the same Romanian word e.g. genitive singular and plural forms BANCII, BANCILOR, the adjectival form of the nouns with their singular/plural and masculine/feminine variants BANCAR, BANCARA, BANCARI, BANCARE plus the difference between the studied words with or without definite article BANCILE bring us to the total amount of 85 hits in Romanian. Therefore truncation to the very stem of the word used as query will influence the search result particularly in a language whose grammatical clitics are so important.

3. The *French* records in the database represent only half of the number of occurrences for the studied queries in English. The 35 records containing the words '*banque/banques'* divide their meanings as follows:

Financial institution	– 13 hits
Institution name in the title	– 14 hits
Publisher / Series title	– 4 hits
Computer databases (data banks)	– 4 hits

A Romanian record was intermingled among the French ones. The exception was caused by the French word '*banques'* mentioned as title word in an annotation including the original title of a document, i.e. *"Les banques"*, in one of the Romanian bibliographic records.

A brief look at the figures resulted from our survey brings us to the conclusion that it is not necessarily true that the presence of a particular word, a meaningful word or a key word in the bibliographic description has much to do with the subject of the document described. The greatest number of occurrences in the title category is related to the subject matter 'banks and banking'. The number of occurrences of the studied words in English bibliographic records is remarkably increased by the category of personal name, missing in both the other languages. This category, of course, has no relation whatsoever with the meaning of studied concepts as such. The same thing can be said about the compound word in which 'bank' is a component part in all the studied languages, i.e. 'data bank'. As far as meaning is concerned, the closest meaningful category related to our query statement is the second in our list: institution name in the title.

Let us investigate now another search key and study the effect this kind of queries has on information retrieval.

6.6.2 Searches using Romanian descriptors manually assigned by indexers

Starting from the truncated query ‘de=banc$’ for Romanian descriptors (as far as they are found in the database as an outcome of the indexing process) the search result was 122 hits. The query formulations are the following:

"DE=BANCHERI" + "DE=BANCHERI (BIOGRAFII)" + "DE=BANCI" + "DE=BANCI AGRICOLE" + "DE=BANCI DE CREDIT" + "DE=BANCI INTERNATIONALE" + "DE=BANCI NATIONALE" + "DE=BANCI POPULARE"

Given the above-mentioned query elements, the records displayed are in various languages, actually in any of the languages existing in the database. However, this result does not reflect all the relevant information about *banks* and *banking* available in the database. The simple reason for that is that descriptors were assigned to bibliographic records only at a certain moment in the history of the online catalogue. Therefore, a far more complete search result will be given by using the UDC notations representing the studied concepts or subjects such as: *336 Finance. Public finance Banking. Money, 336.7 Money. Monetary system. Banking. Stock exchanges and 336.71 Banks. Banking.* If the query is formulated like this:

```
? v675*0.5 : '336.7'
```

then the search result will be 674 hits, including all records having one or more of the above UDC notations occurring as such or in combination with auxiliaries of place, of form, or being lower subdivisions of these numbers. [??? different from the printed text, p. 29 ???]

6.6.3 Searches using words from the UDC text (captions)

A third approach that may bring about relevant information to our survey is using words from the UDC text to search with. For the query formulation "DU=BANKS" we have 116 hits and for "DU=BANKING" other 731 hits. If we compare the number of hits retrieved, the last ones are amazingly big. There is yet need for some discussion here given the word ‘banks’ is found in some other semantic areas but that we are interested in. One of them was also found in the first approach undertaken (words from title), i.e. ‘*Data banks*’, found in Mathematics, under the UDC number 519.256. Another one is ‘*Banks. Banking business*’, found in Law, under the UDC number 347.734. A third one is ‘*International banks*’, from International finance, under the UDC number 339.732. Let alone another combination, i.e. *‘Savings banks*’ which is still a related concept and is found under 336.722.

The condition for getting a higher rate of precision in such circumstances like the one just described is that context is specified. Disambiguation is a necessity with a view to lowering such a high recall rate. To this purpose that additional subfield ^x was created. The use of Boolean operators AND, OR and NOT will enable complex step-by-step search strategies to modify the search result.

Going briefly back to our survey if we apply the Boolean logic operators to our search strategy the search expression DU=BANKS + DU=BANKING will retrieve 832 records and DU=BANKS * DU=BANKING will only retrieve 15. In both sets of retrieved records (although the number of hits is lower than for each separate query formulation) the precision is still hard to be satisfactory. Despite the restriction operated by the Boolean logic, there are still records mixed up from other domains in the search results, particularly when the OR (+) operator was used. That is the reason why subfield ^x is indexed in order to permit the intersection between the subject as derived from the UDC text and the context given by the class.

Another solution that leads to the same result i.e. restriction of the number of hits, eliminate ambiguity, hence decrease the recall rate and improve precision is the use of a string search. The search expression will be formulated as follows:

```
?v709^e : 'banking' and v709^x : 'economics'
```

and the search result – 728 hits – will only include those bibliographic records that deal with both these topics: one as subject and the other one as context.

In addition to the earlier mentioned search methods, the use of multilingual descriptors as queries will make our demonstration more pertinent. And indeed, the use of thesaurus descriptors from LTHES (‘*banking*’) bring about a number of records that is very close to that resulting from the search using a significant word from the UDC caption, i.e. banking modified by the context provider i.e. economics (713 vs 728). However surprising it may seem, the use of one of the more specific thesaurus terms from PTHES (‘*banks*’) bring only 194 records. The reason why it is that PTHES, given its higher specificity, has a larger variety of terms for related categories of *banks* and *banking* whereas LTHES has descriptors for only a higher level of division i.e. banking – UDC 336.7. The comparative results of these searches are given in the table below):

Type of search or search mode		**Number of hits**		
Searches using words from the bibliographic description	*Language of the search result*	En	Ro	Fr
	Financial institution and related activities	21	15	13
	Institution name in the title	14	10	14
	Personal (author) name	11	-	-
	Publisher / Series title	16	17	4
	Computer data bases (data banks)	8	13	4
	Total number of retrieved records	70	55[21]	35
Romanian manually assigned descriptors (related to banks)		122		
Words from the UDC captions: banks + banking		116 + 731		
Multilingual descriptors from LTHES		713		
Multilingual descriptors from PTHES		194		

Other examples of string searches will show the difference between context dependent and context free query formulations:

```
Set  Data Base   Hits  Query element
---  ---------   ----  --------------
 1     BCUB       148  ?v709^e : 'acoustics' and v709^x : 'physics'
 2     BCUB        10  ?v709^e : 'acoustics' and v709^x : 'music'
 3     BCUB         2  ?v709^e : 'acoustics' and v709^x : 'linguistics'
 4     BCUB       159  ?v709^e : 'acoustics'
```

The difference in the number of records retrieved as results of the four searches above illustrates the possibility of combination between a subject and its context in order to provide

[21] If grammatical clitics are included the total number of Romanian retrieved records goes up to 85.

higher precision of the search result (compare the number of the last query hits with each of the three preceding ones).

The multilingual applications of this string search method can also be tested using other languages included in the MRF part of the database. As long as other language variants of the UDC MRF exist they can be used on a par with English. Likewise, the shortened version of the MRF providing context to the UDC notations used can also have other language variants but English. So, we can have as many languages as needed or available in the fields designed for text added to the UDC notations in the bibliographic records, i.e. 70- fields, hence information retrieval will be available in either of these languages. Moreover, other fields can be involved in combinations of the type we described above permitting the restriction of the subject to a certain geographic area or a certain moment or period of time.

6.7 Conclusions: summary of methodology and final results

By virtue of the described methods of updating the UDC numbers and adding text and descriptors to bibliographic records new life is evolving from the old in our experimental database. However complex it may seem the methodology used in making information retrieval possible via text added and descriptors derived from UDC notations proved its feasibility. Our purpose was reached: we demonstrated how good both of them work together.

Moreover multilingualism adds an extra advantage to information access. Once the thesaurus terms are multilingual and the link exists between the UDC notations and the descriptors the various language variants of the descriptors will be automatically available to a wider range of users.

To summarize our methodology let us briefly consider *the main stages of our case study:*

1. In order to start the procedures of making information retrieval possible using words in the UDC captions and descriptors it is recommendable that the database is “cleaned up” of errors, especially in the subject field. Basically there are three *categories of errors* that may be found more often such as: typing errors, errors deriving from misuse of the UDC, errors generated by lack of update according to the E&C level in the MRF. We have described and illustrated with examples the methods of identification of the wrong UDC numbers. There are two situations that may come up: (1) the erroneous UDC notation is known – then one has to select and use option S for ‘search expression’ and type in the wrong number and (2) the erroneous UDC notation is not known – then option T has to be selected and used to display the terms dictionary.

2. Once *the invalid UDC numbers* are identified there are approaches of *correcting* them: by making single or individual changes, which is painstaking and time consuming, and by global changes. For the individual changes we have described the procedure to be followed for correcting a main number and an auxiliary number. For the first kind of UDC numbers the change has to be done in field 675, the updated number is repeated in field 709 and then the program us709 is run against the record and adds the correct text to that number. For the auxiliary numbers the change is done in a 68 field, repeated in the corresponding 69 field and then by running the program us 68 followed by the particular digit representing the type of auxiliary we are concerned with the appropriate text will appear in a 70- field.

3. *Global changes* are extremely helpful and easy to manage. The first necessary thing is to formulate a query. Then go to the main menu and select A, run program GLOB,

select option H for field 675, change the source string with the new string of digits included in the UDC number and the program will go through each record of the last query and make the changes. Then run the program UDCCON. Nothing has to be done manually which saves a great deal of time.

4. When the database is reasonably clear of invalid or erroneous UDC notations the *search procedures* may start up. Information can be retrieved in two different ways depending on whether the subject is specified (has a known name) or not (the user just wants to browse the index and select different related topics belonging to a certain semantic area). There are two groups of search codes approaching the information retrieval in two ways (see Appendix 1). The first has as destination *the descriptive part of the bibliographic records* and the second *the subject of the documents in the bibliographic database*. In the first case the search request is filled in after the introducing codes used for the display of the terms dictionary, e.g. TI = Title of the document; KW = Words from the bibliographic description (title included). For the second type of searches the potential user will place his queries after such codes as: DU = Words from the UDC text automatically added in a 700 field in the database, DE = Descriptors previously assigned to bibliographic descriptions in the database. The multilingual access is possible in two different ways depending on the thesaurus terms chosen to search with. The search codes for automatically assigned descriptors from the multilingual thesaurus vary according to either the language of the query formulation (either English or French or Romanian) or the expected degree of specificity. The thesaurus terms in LTHES and PTHES are available as explicitly shown in Appendix 1. The use of option T from the Information Retrieval Services opens the gateway to this types of searches. Another search method starts from option S – search formulation – of the same menu and is based on search formulas able to be used either separately or in combinations, following the rules of the Boolean logic. Thus the symbols of the Boolean operators OR (+), AND (*) and NOT (^) can modify the search result according to the search strategy used. The query elements as parts and stages in a search strategy can bring together information from either different fields or from fields and subfields and thus enable complex searches. The validity of a search result generated by a search key can be checked by another search key.

5. The whole descriptor index is displayed by choosing the T option from the Information Retrieval Services without any other code and all language descriptors are interfiled.

6. As far as the *multilingual access* is concerned we have seen that multiple language variants of the MRF make possible multilingual access to the information contained in the database. As long as the UDC MRF is provided in a particular language in the database that particular language can be used to search with. Any other language variant of the MRF added to the existing one enables information retrieval via words from that natural language. Moreover, the multilingual thesaurus embedded in the database enables multilingual access by means of the multilingual descriptors available as an alternative search method. The searcher may have his own choice for a particular language to search with. To this purpose there is a separation of languages in the multilingual descriptor files.

CHAPTER 7
THE IMPACT OF SPECIFICITY ON THE RETRIEVAL POWER OF A UDC-BASED MULTILINGUAL THESAURUS

In his "Rules for a Dictionary Catalogue", formulated as early as 1876, Cutter insisted on the importance of the *specific entry* in subject cataloguing: "Enter a work under its subject heading, not under the heading of the class which includes it" (Rule 106). It is specificity that this chapter will focus on.

In the same context, Charles Ammi Cutter said: "The *ideal catalogue* would give under every subject its complete bibliography, not only mentioning all the monographs on that subject, but all the works which in any way illustrate it, including parts of books, magazine articles, and the best encyclopedias that treat of it… ". Cochrane (1985, 287) cites Cutter in a paper in which she attempts to find the answer to the question: "How can we create a catalog that brings works together, does not separate related subjects or conceal information, and allows the user to search with ease and little difficulty no matter whether the query is specific or general". In so doing she mentions the 4 types of catalogues existing in Cutter's times:

1. The dictionary catalogue
2. The alphabetico-classed catalogue
3. The classed catalogue
4. The combined catalogue

Cochrane concludes that the combined catalogue best matches the requirements of the ideal catalogue and if the question inspired by Cutter's statements makes the online catalogue designers concern, then nowadays catalogues can come closer to the ideal catalogue than those in Cutter's times.

Therefore, the ideal catalogue should permit browsing but also navigation, thus bringing documents dealing with related subjects together (1), should reduce to the minimum possible information loss (2) and also provide the users with help messages that assist them during the search process (3).

We shall examine in the coming paragraphs to what extent these requirements are met by our experimental database and in particular make evident the effects different information languages and different degrees of specificity within the same type of information languages have on information retrieval.

7.1 Specificity and exhaustivity

The most outstanding information scientists admit as highly important issues for the study of information retrieval systems key concepts like *recall*, *precision* and *relevance*. These are topics always dealt with in textbooks and manuals on information languages but also preferred subjects for system evaluation procedures. Information system designers and compilers of information languages spend large amounts of time and considerable efforts to make information retrieval as effective as possible and the output product of the IR system as convenient to the user as expected within the shortest time possible. For that purpose the three concepts mentioned earlier are considered and in particular the ways to improve their ratios so that the information need is fulfilled in optimal conditions.

Among the authors who studied in details aspects of *recall*, *precision* and *relevance* as *indicators of performance* of information retrieval systems we shall mention only a few such as: A. C. Foskett (1982), Robert Fugmann (1993), F. W. Lancaster (1998), Gerard Salton and M. J. McGill (1983).

There is a close relationship between the performance indicators just mentioned and the specificity of the indexing language as already said (see **§5.6**). The size of the document collection is also important in this respect and Salton and McGill (1983, 187) appreciate that "the desirable level of precision and thus the importance of language specificity varies with collection size, high precision being most crucial for very large collections."

Before moving forward to demonstrating the impact specificity has on the retrieval power of our UDC-based indexing language, let us see what are the meanings of the main concepts we deal with in this chapter.

Recall, according to Salton and McGill (1983, 55), measures the proportion of relevant information actually retrieved in response to a search. That is, they argue, the number of relevant items actually obtained divided by the total number of relevant items contained in the collection.

Precision, in the opinion of the same authors, measures the proportion of retrieved items actually relevant, which means, the number of relevant items actually obtained divided by the total number of retrieved items.

In other words, recall measures the ability of the system to retrieve useful documents whereas precision measures its ability to reject the useless ones (Salton and McGill, 1983, 160).

Relevance, a concept already used in the definitions of recall and precision, is extensively used in information retrieval environment. It is the expression of the degree to which the retrieved documents correspond with the user's information need in response to the query statement he used to inquire the system. According to Fugmann (1993) "a message is considered relevant if it comprises all the concepts and concept relations (as far as they can be defined in advance in the inquiry) in the desired degree of specificity." Mostly relevance is strongly connected with the quality of indexing or the information system performance.

Other two concepts are often mentioned in close relation with these three parameters. These are *specificity* and *exhaustivity*. They are cited as "factors which affect the overall performance of an information retrieval system and its potential in terms of recall and precision" (Foskett, 1982, 25).

Specificity is, according to the previously cited author, the extent to which the system permits the indexer to be precise when specifying the subject of a document while processing it. The higher the specificity the more likely for the system to achieve high relevance. On the contrary, a system that permits limited specificity is likely to achieve reasonably high recall but correspondingly low relevance. Lack of specificity cannot be amended at the output stage the outcome of it being the necessity imposed on the user to browse unnecessarily large output in order to find or discover the information needed. Serendipity – "the faculty of making happy and unexpected discoveries by accident" – can play an important role in cases of low specificity of indexing languages when the undecided user is not in the position to express precisely his information need.

Exhaustivity of an indexing language refers to its capacity to include terms covering all subject areas mentioned in the document collection (Salton and McGill, 1983, 160). A high level of indexing exhaustivity tends to imply a high recall rate by making it possible to retrieve most of the potentially relevant documents. This will affect precision because some less relevant documents are also likely to be retrieved when many different subject areas are covered by the indexing terms. *The higher the specificity of the index terms, the higher the precision rate since most of the retrieved documents are expected to be relevant. Conversely, the broader or more general the indexing terms, the lower the precision because the broad terms will not distinguish the less relevant documents from the truly relevant ones.* This is the hypothesis on which we build our demonstration in the forthcoming chapter.

7.2 Searches conducted in the experimental database after implementing the second multilingual thesaurus

As already said, the configuration of the experimental database (BCUB) has become more complex with the addition of the second multilingual thesaurus (PTHES) terms. The newly implemented thesaurus will certainly have influence on information retrieval and particularly increase the relevance of the retrieved documents in relation to the query formulation.

We give below some examples of searches and, as we feel necessary, also comment on one or another of the aspects of information retrieval encountered.

Searching for a certain topic, say *"development of libraries"*, will have the following search results depending on the language used in information retrieval:

Search No. 1[22]

```
Set   Data Base   Hits   Query element
---  --------- ------  -------------
1      BCUB         9    "DE=DEZVOLTAREA BIBLIOTECILOR"
2      BCUB        30    "675=021" + "675=021(038)=111" +
                         "675=021(063)(100)" + "675=021(100)" +
                         "675=021(100)(063)(100)" + "675=021(410)" +
                         "675=021(438)(063)(082)" + 675=021(44)(091)"
3      BCUB        49    "709=021"
4      BCUB        49    "LMBR=DEZVOLTAREA BIBLIOTECILOR"
5      BCUB        49    "PMBE=FUNCTION, VALUE, UTILITY,"
```

Some explanations and also some comments are due here:

- DE='Dezvoltarea bibliotecilor' gives the lowest number of hits (only 9) because the manually assigned descriptors were introduced at a later stage in the development of the bibliographic database; in addition to that, even when descriptors are given in, to the same UDC number i.e. 021 some other descriptors were assigned as equivalents such as: *Biblioteconomie*, (Librarianship in Romanian), *Biblioteci* (Libraries in Romanian) and the like.

- The main discussion is demanded by the difference in between the number of hits resulted from 675='021' (30 hits) and 709=021 respectively (49 hits). The 70- field in our case reads: 709: *^a021^eFunction, value, utility, creation, development of libraries^xLibrarianship^oBiblioteconomie*. Where does the difference come from? Field 675, according to the Field Select Table stands for the UDC subject notation. Field 709 is meant for main UDC numbers (and their captions). The comparison between the bibliographic records found in either of the two searches gives the answer to our question: the second set of retrieved bibliographic records does not include those records having *021 either as second member of a colon relation or as part of a range with stroke*.

[22] See Appendix 1 for the prefixes

- The following two searches are identical as far as the search results are concerned: the 709 field and the LTHES descriptor in Romanian, LMBR (replaceable at any time with LMBE or LMBF as needed). We deal in this situation with full compatibility between the UDC main notation and the automatically assigned multilingual descriptors belonging to the LTHES.
- The fifth query, PMBE='Function, value, utility, creation, development of libraries', gives the same number of retrieved records because it is based on the same UDC number as the fourth. Only the form of the descriptor is different. It is important to know that the query has to be formulated in exactly the same way the descriptor is worded, or else the response is 'no hit'.

Our first search proved that manually ascribed subject headings at the time of indexing gave a lower result when used as search element than the automatically assigned ones did. At the same time, using the UDC notations in all their possible combinations is somehow cumbersome.

On the contrary the thesaurus terms demonstrated their effectiveness in finding the relevant information for this query. Yet, a critical point can be made here about the precise form of a descriptor. If this form is not precisely known to the user then the search result will be 'no hit'. A printed version of the thesaurus indicating the wording of the descriptors as much as their relations with other thesaurus terms would be a really helpful tool in such situations and others alike.

Search No. 2

```
Set   Data Base   Hits   Query element
---  --------- ------  -------------
1     BCUB         199   "708=-053.2"
2     BCUB         199   "PMBE=CHILDREN"
3     BCUB         199   "LMBE=CHILDREN"
4    BCUB        1348   "DU=CHILDREN"
```

A search using field 708='-053.2' as query[23] gives 199 hits and the same search results is given by the descriptors of either of the two multilingual thesauri. If the search statement is C*hildren* as keyword from the UDC caption, then DU=Children will give 1348 records as result. But it is not only the term 'children' as equivalent of the common auxiliary of person that we have in these UDC captions. The search is made for all occurrences of this word in any of the 70- fields, including those for main numbers such as:

708: ^a-053.2^e**Children** and infants (in general)
708: ^a-053.4^ePreschool-age **children**
709: ^a027.625^eFor **children**, juniors, juveniles, young people, adolescents. Library work with children^xLibrarianship
709: ^a087.5^ePublications for young people. **Children**`s, juvenile literature. Books for infants. Picture books. Story books^xPolygraphs. Collective works
709: ^a159.922.76^eAbnormal **children**
709: ^a613.22^eNutrition of **children**
709: ^a821.133.1-93^eFrench - Literature for **children**. Juvenile literature^xLiterature

Summing up, the search for a topic like 'children' as agents or regarding personal aspects connected with the main subject will give more relevant results if the search element is a

[23] Field 708 is designed for common auxiliaries of materials and persons as parts of complex UDC notations

descriptor directly related to the UDC number that expresses such aspects. Using a search element like ‘children’ in all its possible occurrences (see search set no. 4 above) the search result will bring too many hits and the precision will be lower in this case than in the previous. The recall rate being so overwhelmingly high browsing the search result may prove helpful for an undecided user. The higher the specificity of the search element is, the higher the accuracy (precision) of the search result (see also our statements in the introductory paragraphs of this chapter). Both types of multilingual descriptors proved their efficiency in this search.

Search No. 3

Searching with an auxiliary of place e.g. 703=’(469)’ and with the pocket UDC thesaurus descriptor PMBE=’Portugal’ will give the same 45 hits as result. The manual descriptor ‘Portugalia’ was assigned to only 27 bibliographic records. LMBE=’Portugal’ will give a ‘no hits’ answer to the search.

```
Set   Data Base   Hits      Query element
---   ---------  ------     -------------
1      BCUB        45     "703=(469)"
2      BCUB        45     "PMBE=PORTUGAL"
3      BCUB        27     "DE=PORTUGALIA"
```

So far we have tested the database and demonstrated that automatically assigned multilingual descriptors give highly reliable results in searching and in most of the cases the number of retrieved records for these searches matches those that use the UDC numbers (or their component parts) as search elements. In the 3rd search we’ve come to a ‘no hit’ result. Let us examine the possible reasons why a ‘no hit’ answer can be given to the search.

Portugal in the above given example is not found among the descriptors in LTHES. That is why the result is ‘no hits’. And indeed, if we search in the MRF for either the number ‘(460)’ or the word ‘Portugal’ the only descriptors found belong to PTHES and they are: ^ePortugal^fPortugal^rPortugalia. Therefore A ‘no hit answer’ may occur in case of *inexistence of a particular descriptor* or, as we saw in the preceding search, the exact form of the descriptor is not known to the user.

Imagine this scenario: the searcher starts by looking in the pocket UDC thesaurus for the term ‘Pottery’. It is not found among the descriptors therefore it might be a non-descriptor. Indeed, NPE=pottery gives 1 hit.

```
          UDC number: -033.6
               Table: k
         Description: ^eCeramics. Pottery. Clayware in general
          References: ^a-032.61
    Pthes Descriptors: ^eCeramics^fCéramique^rCeramică
Pthes Non-Descriptors: ^eClayware^fPoterie^rProduse ceramice
                       ^ePottery^fProduits céramiques^rProduse din argilă
```

The non-descriptor is found along with the corresponding preferred term, i.e. *Ceramics* in English, *Céramique* in French and *Ceramică* in Romanian. In the same record the UDC number is included and can thus be used in searching. The next step further is to use the auxiliary for materials -033.6 as search element, therefore 708=-033.6. The display of the

MRF record gives the user all the elements necessary to start a search by exploring all the possibilities offered by the system such as: UDC number, corresponding descriptors and non-descriptors in one or both UDC-based thesauri.

```
Set   Data Base   Hits  Query element
---   ---------   ----  -------------
1     BCUB        1     "NPE=POTTERY"
2     BCUB        5     "708=-033.6"
3     BCUB        5     "PMBE=CERAMICS"
4     BCUB        52    "DU=CERAMICS"
5     BCUB        32    "709=666.3" + "709=666.3-135" +
                        "709=666.3-978" + "709=666.3.014" +
                        "709=666.3.017" + "709=666.3.018-977" +
                        "709=666.3.04" + 709=666.3/.7" +
                        "709=666.31.7" + "709=666.32" +
                        "709=666.327" + "709=666.327.081.2" +
                        709=666.33" + "709=666.363"
6     BCUB        16    "LMDE=CERAMICS"
7     BCUB        17    "LMBE=CERAMICS"
8     BCUB        32    #5 + #6
```

The number of hits found by using the PTHES thesaurus term ‘Ceramics’ in any of the 3 languages equals the number of those found by the UDC number in field 708 (as search sets 2 and 3 show). If the search is enlarged by using as query ‘ceramics’ as word from the UDC text, the result will be considerably better (52 hits in set number 4). By browsing the resulted 52 bibliographic records the searcher will notice the presence of LTHES descriptors having the same form i.e. ‘Ceramics’ but corresponding to a main number this time. Which means that a descriptor with the same form is included in LTHES for 666.3 as a UDC code. Therefore the next search is based on this UDC number and its subdivisions (set number 5). The two search modes with descriptors in LTHES (sets 6 and 7) are following. If the results of these two sets are put together the result will be the same 32 hits as set 5 based on UDC codes. Therefore, the UDC number and subdivisions gave the same search result when used as search statement as the corresponding descriptors either with or without truncation of the UDC notations in the bibliographic records. As to PTHES, since Class 6 – Applied sciences is only partly represented in the thesaurus, most of the notations in this class are left-truncated to the first digit and given the class description as indexing term (*Figure 43*):

```
+-  10 / 32 -------------------------------------------   Format: BCUB   -+
¦ TITEL: Concise encyclopedia of advanced ceramic materials / Brook, R.J. ¦
¦ - Oxford ; New York ; Beijing[etc.] : Pergamon, 1991. - XVI, 588p.. -   ¦
¦ UDC:   666.3/.7(031)=111                                                ¦
¦ LANG:  ^a=111^eEnglish                                                  ¦
¦ FORM:  ^a(031)^eEncyclopaedias                                          ¦
¦ MAIN:  ^a666.3/.7^eCeramics^xGlass industries. Ceramics. Cement and     ¦
¦         concrete                                                        ¦
¦ PDES:  ^z=111^eEnglish^fAnglais^rEngleză                                ¦
¦ PDES:  ^z(031)^eEncyclopaedias^fEncyclopédies^rEnciclopedii             ¦
¦ PDES:  ^y6^z666.3/.7^eApplied sciences^fSciences appliquées^rStiinţe    ¦
¦         aplicate                                                        ¦
¦ LDES:  ^z=111^eEnglish^fAnglais^rEngleză                                ¦
¦ LDES:  ^y(03)^z(031)^eReference material^fOuvrages de référence^rLucrări¦
¦         de referinţă                                                    ¦
¦ LDES:  ^y666.3^z666.3/.7^eCeramics^fCéramique^rCeramică                 ¦
¦ LNDES: ^y666.3^z666.3/.7^eCeramic materials^fArgiles céramiques^rArgile ¦
¦         pentru ceramică                                                 ¦
¦                                                            MFN: 36865   ¦
+-------------------------------------------------------------------------+
```

Figure 43. Bibliographic record with automatically assigned descriptors resulted from truncation of UDC codes

The multiple search possibilities offered by our approach enhance the information retrieval in the experimental database as demonstrated by the searches above. Even if a ‘no hits’ result is given to one or another of the search elements, alternatives are always available and meant to satisfy the user queries. What is still to be evaluated is how relevant the search results are when compared to the user information need.

Search no. 4

Several types of searches have been conducted for ‘Experimental psychology’ as query. The results of these alternative searches are given below.

```
Set   Data Base   Hits   Query element
---   --------- ------  -------------
1     BCUB        173    ? v675 : '159.9.07'
2     BCUB         86   "675=159.9.07"
3     BCUB        111   "DE=PSIHOLOGIE EXPERIMENTALA"
4     BCUB        123   "PMBE=EXPERIMENTAL PSYCHOLOGY"
5     BCUB         48   "PMDE=EXPERIMENTAL PSYCHOLOGY"
6     BCUB        170    #4 + #5
7     BCUB       1891   "LMDE=PSYCHOLOGY"
8     BCUB        559   "LMBE=PSYCHOLOGY"
```

The first of these searches gives all the occurrences of 159.9.07 as UDC notation corresponding to the query. According to the formulation of the query the result will contain all the bibliographic records having this UDC number and its subdivisions plus the auxiliaries of form accompanying it (the first search formulation covers also the truncated variants of the number). Additionally, records having this UDC number as the first or the second member of a colon relation, are also included. The second search set gives the result for the records indexed with exactly this number. The latter are included in the former.

Next, the search was made for the manually assigned Romanian descriptor ‘Psihologie experimentala’. The number of retrieved records (111) is rather misleading because of two reasons: (1) not all the bibliographic records have manually assigned descriptors and (2) some of the records indexed with this descriptor have a completely different UDC number but 159.9.07 as they should. This is an example of inconsistency in indexing and a proof that the required one-to-one correspondence between UDC notations and manual descriptors is not respected.

The forthcoming searches used *‘Experimental psychology’* as query element. The PTHES includes this descriptor therefore there is a search result for this query element. But the LTHES does not have it therefore a ‘no hit’ response is received. For ‘experimental psychology’ yet, there is a reference pointing to ‘Psychology’ as the discipline covering all the related subjects. Therefore among the high numbers of records retrieved some will have the subject searched for. The expected result will come up after browsing the list of retrieved records and selecting the needed ones. We deal here with a high recall rate but little precision. And it is so because of lack of specificity in LTHES thesaurus terms.

Let us go back to the other type of searches, i.e. those using the PTHES descriptors. There are two instances again: one with the descriptor itself based on exactly that particular UDC number (123 hits), and another one with the descriptor assigned to records where the UDC number has either subdivisions or auxiliaries of form attached to it (48 hits). The total result is rather close to the first search result, i.e. 170 hits vs. 173. The difference can be explained by the way the descriptors are handled in case of very long UDC notations for example 159.9.072.533.6. There are still typing errors in the experimental database and they can influence the search result. But the precision rate in the 6th search set, when PTHES terms were used as queries, is much higher than in the 7th and 8th search sets, when the query element – from LTHES – was much broader.

We insist here on the high precision of the search result in searches that used the multilingual descriptors based on the UDC Pocket Edition. This one, unlike the initial thesaurus includes among its entries the concept used as query therefore the search result more precisely reflects the real number of documents dealing with this subject than any of the last two searches using the very broad term 'Psychology'. The huge number of hits retrieved by those searches – 1968 and 579 – contain only a small part of relevant documents. Yet, this is the only solution offered by LTHES as a thesaurus with restricted coverage and little opportunity to offer ways for the user to discover information of a refined category as this. Browsing almost 2000 bibliographic records makes his task much more difficult than expected or desired and a good way to avoid such inconvenience is to provide the information retrieval system with help messages. Such a message could suggest the user to try another search key – one with a higher degree of specificity – that might give a result closer to reality. This kind of messages together with a field designed for feedback from the information searchers are meant to enhance the quality of the retrieval and thus the reliability of the system as a whole. The capital requirement proclaimed by the slogan "save the time of the user" has to be respected and system designers have to keep that in mind at any time when conceiving their products.

We shall see in the following examples the importance the *authority control* and *indexing consistency* have for information retrieval, how much the precision ratio is influenced by this control and we shall use for this purpose the same query statement in order to have the possibility to compare our findings.

In the MRF part of the experimental database we read:

```
MFN: 178312
      UDC number: 159.9.07
               Table: M
 Special aux. type: B
        Description: ^eExperimental psychology. Psychological research
 Pthes Descriptors: ^eExperimental psychology^fPsychologie expérimentale
                    ^rPsihologie experimentală
 Pthes Non-Descriptors: ^ePsychological research^fRecherche psychologique
                         ^rCercetare psihologică^ePsychological tests^fTestes
                        psychologiques^rTeste psihologice

MFN: 178313
      UDC number: 159.9.072
              Table: M
 Special aux. type: B
        Description: ^ePsychological experiment, investigation, tests,
           measurements
```

The close meaning of these two UDC notations in the table can induce confusion and hence misinterpretation of the meaning of each. That explains how the document in the bibliographic record below[24] was classified with a subdivision of 159.9.07 although the

[24] This as much as the following searches were made in the online catalogue of the Central University Library accessible at www.bcub.ro as of October 2002

selected subject heading was not ‘Teste psihologice’ (also existing in the subject heading list) but ‘Psihologie experimentala’ a subject heading more commonly used for this kind of subjects. (The search keys are in bold letters).

Complete description   1 of 1

Psihologie experimentala
1 Cifrul vietii psihice / Adina Chelcea si Septimiu Chelcea. - Bucuresti : Editura Stiintifica si
Enciclopedica, 1978. - 144 p. : fig. ; 17 cm

Code(s)   :   159.9.072(0.062)
                159.922(0.062)
Subj.headings: 1 **Psihologie experimentala**
               2 Psihologia dezvoltarii
               3 Literatura de popularizare

Let us consider the titles included in the alphabetic lists resulted from the different searches and analyse them in relation with the search methods used.

The first search had as result 164 retrieved records of which we give the first 9 as examples. Here the search method used was the subject heading ‘*Psihologie experimentala*’ assigned manually to the bibliographic records at the time of indexing:

alphabetic list of titles   164 titles

Psihologie experimentala
1 Abandonarea fumatului prin metode psihologice / Daniel Costa. - Buc
2 Activitatea cognitiva a operatorului uman / Gh. Iosif. - Bucuresti
3 Analiza factorilor psihici / Nicolae Margineanu ; [pref. de Fl. Ste
4 Aphasie als kritisches Lebensereignis und Strategien ihrer Bewõltig
5 Aplicatii de curs psihoterapie : [manual pentru uzul studentilor] /
6 The appraisal of intelligence / A. W. Heim. - Oxford : NFER-Nelson,
7 Aptitudinea technica si inteligenta practica / Liviu Rusu. - Cluj-N
8 L'association des idées / Edouard ClaparÞde. - Paris : Octave Doin,
9 Les automatismes cognitifs / sous la dir. de Pierre Perruchet. - Li

Next, the search method used is the UDC number corresponding to the query i.e. *159.9.07* as shown in the MRF record above. The difference between the number of hits resulting from these two searches cannot be explained other than by the various relations or concordances between UDC numbers and subject headings. The control over the indexing languages used has to exist not only inside each of the two but also in between the two. An authority file stipulating which UDC number or combination of numbers corresponds to which descriptor or combination of descriptors would be of great help in such circumstances. Our approach of using a combined indexing language that automatically adds descriptors to the UDC notations used in classifying the subjects of documents proves once more its advantage.

```
alphabetic list of titles                    140 titles

  159.9.07
  1  Activitatea cognitiva a operatorului uman / Gh. Iosif. - Bucuresti : Ed
  2  Analiza factorilor psihici / Nicolae Margineanu ; [pref. de Fl. Stefane
  3  Aptitudinea technica si inteligenta practica / Liviu Rusu. - Cluj-Napoc
  4  L'association des idées / Edouard ClaparÞde. - Paris : Octave Doin, 190
  5  Les automatismes cognitifs / sous la dir. de Pierre Perruchet. - LiÞge
  6  Chestionarele de personalitate în evaluarea psihologica / Mihaela Minul
  7  The cognitive control of motivation : the consequences of choise and di
  8  Community psychology:theory and practice / Jim Orford. - Chichester ;
```

If we modify the search result previously obtained by *restricting* it with the subject heading *'Psihologie experimentala'* we have the resulting number of hits (97) that are indexed with both the descriptor and the UDC number as appropriate (see the list in the box below). Therefore only 97 of the 140 documents have been appropriately indexed with both the UDC number for Experimental psychology and its corresponding subject heading. This is not likely to happen in the approach we undertake here, the correspondence between the UDC notations and index terms being established in a prior stage to that of the indexing and implemented automatically in the database.

A comparison of the titles in the alphabetical list will confirm the difference between the two sets of retrieved records. However small those differences might be the improved precision makes our approach worthwhile.

```
list of titles                         97 titles

  RESULT AFTER 2 SEARCHES
  1  Activitatea cognitiva a operatorului uman / Gh. Iosif. - Bucuresti
  2  Analiza factorilor psihici / Nicolae Margineanu ; [pref. de Fl. Ste
  3  Aplicatii de curs psihoterapie : [manual pentru uzul studentilor] /
  4  Aptitudinea technica si inteligenta practica / Liviu Rusu. - Cluj-N
  5  L'association des idées / Edouard ClaparÞde. - Paris : Octave Doin,
  6  Les automatismes cognitifs / sous la dir. de Pierre Perruchet. - Li
  7  Bucura-te de ceea ce ai : cum sa-ti controlezi sentimentele si sa-ti
  8  Chestionarele de personalitate în evaluarea psihologica / Mihaela M
  9  The cognitive control of motivation : the consequences of choice an
```

Search No. 5

It would be interesting to examine a number of searches for *'Field crops'* as topic statement and see what the difference is whether we use the small thesaurus in searching compared with the Pocket Edition thesaurus. In the first example (MFN 7190) the subject is represented by a relation between two main UDC numbers of which one is a range using the stroke (633.1/.8). The decomposition algorithm will list all the subdivisions and their captions according to the MRF. Likewise, the PTHES descriptors corresponding to this notation, since they are included in the thesaurus, will show the three language versions of all these subdivisions. This means that a search for any of the descriptors corresponding to this range of UDC numbers will retrieve all the bibliographic records indexed with them (some of which can be seen in the subsequent examples of bibliographic records: MFN 132058, 135756 and 157591). Moreover, the existing non-descriptors will increase the possibilities of access to subjects in this area.

On the contrary, as clearly seen in this example (*Figure 44*), the LTHES descriptors, being much more restrictive and therefore less specific, will make no difference between one or another of the subdivisions of 633, the only available one being the corresponding descriptor of 633, i.e. ‘field crops’. Therefore all the information on subjects related with ‘field crops’, irrespective of their particularities will be put together in only one cluster under this descriptor.

```
+     3 / 30 ------------------------------------------- Format: BCUB ----+
¦TITEL: Curs de protectia plantelor tropicale, subtropicale si           ¦
¦mediteraneene / Pßll, Olga / Musat, Despina. - Cluj-Napoca : Facultatea ¦
¦Agronomie, Sectia de agricultura, 1989. - 249p. multigr.                ¦
¦UDC:   632:633.1/.8(213)(075.8)                                         ¦
¦UDC:   633.1/.8(213):632(075.8)                                         ¦
¦UDC:   663.9:632(075.8)                                                 ¦
¦FORM:  ^a(075.8)^eTexts for university, higher education                ¦
¦PLAC:  ^a(213)^eSubtropical and tropical regions generally              ¦
¦MAIN:  ^a632^ePlant damage, injuries. Plant diseases. Pests, organisms  ¦
¦        injurious to plants. Plant protection^xPlant damage, injuries.  ¦
¦        Plant diseases. Pests, organisms injurious to plants. Plant     ¦
¦        protection                                                      ¦
¦MAIN:  ^a633.1^eCereals. Grain crops^xField crops and their production  ¦
¦MAIN:  ^a633.2^eForage grasses. Meadow and pasture grasses^xField crops ¦
¦        and their production                                            ¦
¦MAIN:  ^a633.3^eForage plants except grasses^xField crops and their     ¦
¦        production                                                      ¦
¦MAIN:  ^a633.4^eEdible roots and tubers. Root crops^xField crops and    ¦
¦        their production                                                ¦
¦MAIN:  ^a633.5^eTextile and fibre plants^xField crops and their         ¦
¦        production                                                      ¦
¦MAIN:  ^a633.6^eSugar and starch plants^xField crops and their production¦
¦MAIN:  ^a633.7^ePlants yielding stimulants. Beverage plants^xField crops ¦
¦        and their production                                            ¦
¦MAIN:  ^a633.8^eAromatic plants. Condiment plants. Oleaginous plants. Dye¦
¦        plants. Tanning plants. Medicinal plants^xField crops and their ¦
¦        production                                                      ¦
¦MAIN:  ^a663.9^eChocolate. Cocoa. Coffee. Tea. Tobacco^xIndustrial      ¦
¦        microbiology. Industrial mycology                               ¦
¦PDES:  ^y(075)^z(075.8)^eEducational texts^fTextes pour l’enseignement  ¦
¦       ^rTexte pentru învăţământ                                        ¦
¦PDES:  ^z(213)^eSubtropical and tropical regions^fRégions subtropicales ¦
¦        et tropicales^rRegiuni subtropicale şi tropicale                ¦
¦PDES:  ^y6^z632^eApplied sciences^fSciences appliquées^rStiinţe aplicate ¦
¦PDES:  ^z633.1^eCereals^fCéréales^rCereale                              ¦
¦PDES:  ^z633.2^eForage grasses^fHerbes fourragères^rIerburi furajere    ¦
¦PDES:  ^z633.3^eForage plants^fPlantes fourragères^rPlante furajere     ¦
¦PDES:  ^z633.4^eEdible roots and tubers^fRacines comestibles et         ¦
¦        tubercules^rRădăcini comestibile şi tuberculi                   ¦
¦PDES:  ^z633.5^eTextile plants^fPlantes textiles^rPlante textile        ¦
¦PDES:  ^z633.6^eSugar plants^fPlantes sucrières^rPlante de zahăr        ¦
¦PDES:  ^z633.7^ePlants yielding stimulants^fPlantes stimulantes^rPlante ¦
¦        stimulante                                                      ¦
¦PDES:  ^z633.8^eAromatic plants^fPlantes aromatiques^rPlante aromatice  ¦
¦PDES:  ^y6^z663.9^eApplied sciences^fSciences appliquées^rStiinţe plicate¦
¦PNDES: ^y(075)^z(075.8)^eSchoolbooks^fManuels écoliers^rManuale scolare ¦
¦PNDES: ^z633.1^eGrain crops                                             ¦
```

```
¦PNDES: ^z633.2^eMeadow and pasture grasses^rPlante de nutreț           ¦
¦PNDES: ^z633.7^eBeverage plants^fPlantes à boisson                     ¦
¦PNDES: ^z633.7^eNarcotic plants                                        ¦
¦PNDES: ^z633.7^eTobacco^fTabac^rTutun                                  ¦
¦PNDES: ^z633.8^eCondiment plants^rPlante pentru condimente             ¦
¦PNDES: ^z633.8^eTanning plants^fPlantes à tanin                        ¦
¦PNDES: ^z633.8^eMedicinal plants^fPlantes médicinales^rPlante medicinale ¦
¦PNDES: ^z633.8^eOleaginous plants^fPlantes oléagineuses^rPlante        ¦
¦       oleaginoase                                                     ¦
¦LDES:  ^y(075)^z(075.8)^eManuals^fManuels^rManuale                     ¦
¦LDES:  ^y(21)^z(213)^eContinents^fContinents^rContinente               ¦
¦LDES:  ^z632^ePlant injuries^fAffections des plantes^rBoli și dăunători ¦
¦        ai plantelor                                                   ¦
¦LDES:  ^y633^z633.1^eField crops^fPlantes de culture^rPlante de cultură ¦
¦LDES:  ^y633^z633.2^eField crops^fPlantes de culture^rPlante de cultură ¦
¦LDES:  ^y633^z633.3^eField crops^fPlantes de culture^rPlante de cultură ¦
¦LDES:  ^y633^z633.4^eField crops^fPlantes de culture^rPlante de cultură ¦
¦LDES:  ^y633^z633.5^eField crops^fPlantes de culture^rPlante de cultură ¦
¦LDES:  ^y633^z633.6^eField crops^fPlantes de culture^rPlante de cultură ¦
¦LDES:  ^y633^z633.7^eField crops^fPlantes de culture^rPlante de cultură ¦
¦LDES:  ^y633^z633.8^eField crops^fPlantes de culture^rPlante de cultură ¦
¦LDES:  ^z663.9^eStimulant products^fProduits stimulants^rStimulente    ¦
¦LNDES: ^y633^z633.1^eCereals^fCéréales^rCereale                        ¦
¦LNDES: ^y663^z663.9^eFermentation industry^fIndustrie des fermentations ¦
¦       ^rIndustria fermenților                                         ¦
¦BCUB#: 18825                                                MFN: 7190  ¦
+-----------------------------------------------------------------------+
```

Figure 44. Different degrees of specificity in automatically assigned descriptors for a range of UDC notations

The real difference in specificity of the two thesauri and its impact on retrieval will however be proved by the next example (*Figure 45*). The variety of subordinate index terms more clearly show the advantage in precision resulting from higher specificity of the information language.

```
+    23 / 30 ------------------------------------------- Format: BCUB  --+
¦TITEL: Aspecte privind cultura cerealelor în Franta / Baicu, T. / Baicu, ¦
¦Tudorel M.. - Bucuresti : Editura Agro-Silvica, 1968. - 167 p.          ¦
¦BDES:  Cultura plantelor                                                ¦
¦BDES:  Cereale                                                          ¦
¦UDC:   633.1(44)                                                        ¦
¦SIMP:  633.1(44)                                                        ¦
¦PLAC:  ^a(44)^eFrance. French Republic. République Française            ¦
¦MAIN:  ^a633.1^eCereals. Grain crops^xField crops and their production  ¦
¦PDES:  ^z(44)^eFrance^fFrance^rFranța                                   ¦
¦PDES:  ^z633.1^eCereals^fCéréales^rCereale                              ¦
¦PNDES: ^z(44)^fRépublique Française                                     ¦
¦PNDES: ^z633.1^eGrain crops                                             ¦
¦LDES:  ^z(44)^eFrance^fFrance^rFranța                                   ¦
¦LDES:  ^y633^z633.1^eField crops^fPlantes de culture^rPlante de cultură ¦
¦LNDES: ^y633^z633.1^eCereals^fCéréales^rCereale                         ¦
¦BCUB#: 179508                                              MFN: 132058  ¦
+------------------------------------------------------------------------+
```

Figure 45. Different degrees of specificity in automatically assigned descriptors for a single UDC notation

The following sets of searches will show comparatively the importance specificity has for the performance of the information language. While bringing a great number of retrieved records for only one query, the use of the LTHES descriptor 'Field crops' will affect negatively the search result by lack of precision. On the contrary, the descriptors in PTHES, a thesaurus having a more specialized vocabulary and an entry for each of the subdivisions of the UDC number underlying the considered descriptor, will prove the second thesaurus better performance.

Our first query statement at this point is ‘Textile plants’, a descriptor that has an entry only in the Pocket Edition thesaurus (PTHES).

```
Set  Data Base    Hits  Query element
---  ---------    ----  -------------
1    BCUB          11   "709=633.5"
2    BCUB          11   "PMBE=TEXTILE PLANTS"
```

```
+    10 / 11 ------------------------------------------- Format: BCUB  ---+
¦TITEL: Combaterea bolilor plantelor textile / Becerescu, D. - Bucuresti: ¦
¦Ceres, 1979. - 251 p.                                                    ¦
¦BDES:  Fitopatologie                                                     ¦
¦BDES:  Combaterea bolilor si daunatorilor                                ¦
¦BDES:  Plante textile                                                    ¦
¦UDC:   632:633.5                                                         ¦
¦UDC:   633.5:632                                                         ¦
¦MAIN:  ^a632^ePlant damage, injuries. Plant diseases. Pests, organisms   ¦
¦        injurious to plants. Plant protection^xPlant damage, injuries.   ¦
¦        Plant diseases. Pests, organisms injurious to plants. Plant      ¦
¦        protection                                                       ¦
¦MAIN:  ^a633.5^eTextile and fibre plants^xField crops and their roduction¦
¦PDES:  ^y6^z632^eApplied sciences^fSciences appliquées^rStiințe aplicate ¦
¦PDES:  ^z633.5^eTextile plants^fPlantes textiles^rPlante textile         ¦
¦PNDES: ^y633^z633.5^rPlante de câmp                                      ¦
¦LDES:  ^z632^ePlant injuries^fAffections des plantes^rBoli şi dăunători  ¦
¦        ai plantelor                                                     ¦
¦LDES:  ^y633^z633.5^eField crops^fPlantes de culture^rPlante de cultură  ¦
¦LNDES: ^z632^eControl of plant diseases and pests^fDommages des          ¦
¦        plantes^rCombaterea bolilor şi dăunătorilor                      ¦
¦LNDES: ^y633^z633.5^eCereals^fCéréales^rCereale                          ¦
¦BCUB#: 182532                                              MFN: 135756   ¦
+-------------------------------------------------------------------------+
```

Figure 46. Example of how a lower specificity degree affects automatic indexing and information retrieval

As expected, the number of documents retrieved by the appropriate UDC number in a 700 field equals the number of documents retrieved by the descriptor in PTHES. One of the 11 records is given as example in order to illustrate the way the automatic indexing is done (*Figure 46*). Mention should be made on the absence of the specific entry for the UDC number 633.5 in LTHES where the subdivisions of 633 are not included in the abridged schedule used as a basis for the thesaurus construction. Consequently, all documents having been classified with one of the subdivisions of 633 will be automatically indexed with the broader term ‘Field crops’. This way the automatic indexing process is affected with respect to the specificity that would confer more precision in information retrieval.

Another example is meant to show how manual indexing is influenced by factors that prevent the one-to-one correspondence between a UDC notation and a descriptor thus putting consistency in indexing under question (*Figure 47*). The existence of more than one category or concept in the caption of the UDC number (633.8) found in the bibliographic record determines the use of alternative descriptors each time the title of the document is changed irrespective of the fact the UDC number remains the same. An even higher degree of specificity in our thesauri would permit the use of each alternative concept or category for a separate descriptor. But at this level of division the corresponding descriptor is automatically and consistently given according to our methodology.

```
   Set  Data Base    Hits  Query element
   ---  ---------    ----  -------------
   3    BCUB          24  "709=633.8"
   4    BCUB          24  "PMBE=AROMATIC PLANTS"
```

```
+    19 / 24 -------------------------------------------- Format: BCUB  --+
¦TITEL: Plante medicinale si aromatice / Coiciu, Evdochia, Rßcz, Gabriel.-¦
¦[Bucuresti] : Editura Academiei Republicii Populare Române, 1962. - 683p.¦
¦BDES:  Plante aromatice                                                  ¦
¦BDES:  Plante medicinale                                                 ¦
¦UDC:   633.8                                                             ¦
¦MAIN:  ^a633.8^eAromatic plants. Condiment plants. Oleaginous plants. Dye¦
¦        plants. Tanning plants. Medicinal plants^xField crops and their  ¦
¦        production                                                       ¦
¦PDES:  ^z633.8^eAromatic plants^fPlantes aromatiques^rPlante aromatice   ¦
¦PNDES: ^z633.8^eCondiment plants^rPlante pentru condimente               ¦
¦LDES:  ^y633^z633.8^eField crops^fPlantes de culture^rPlante de cultură  ¦
¦LNDES: ^y633^z633.8^eCereals^fCéréales^rCereale                          ¦
¦BCUB#: 208649                                              MFN: 157591   ¦
+-------------------------------------------------------------------------+
```

Figure 47. Example of the way manual indexing is tributary to alternative indexing terms for one and the same UDC notation

Our search statement is 'Aromatic plants' and the result is comparable with the previous one. A brief look at the way the manual indexing was made shows that the decision of the indexer is visibly influenced by the title of the document. And again, the LTHES terms being too general, gave no opportunity for higher specificity with bad consequences on precision.

Another search statement is using another subdivision of the considered UDC number for Field crops (*Figure 48*). Here, like in the previous examples of bibliographic records, the automatically assigned descriptors from the pocket UDC thesaurus allow a higher degree of specificity of the indexing terms with direct consequences on precision and relevance in information retrieval.

The third and last search in this series ('Forage plants') is meant to confirm our statement i.e. *the higher the specificity of the indexing terms, the higher the precision since most of the retrieved documents are expected to be relevant.* The alternative solution for the user would be to browse all the 81 documents indexed with 'Field crops' and select the desired ones out of them (see the 3 sets of searches following *Figure 48*). But this operation would contradict every principle of quality, performance and user-friendliness of an information system.

```
   Set  Data Base    Hits  Query element
   ---  ---------    ----  -------------
   5    BCUB          24  "709=633.3"
   6    BCUB          24  "PMBE=FORAGE PLANTS"
```

```
+     2 / 24 -------------------------------------------- Format: BCUB  --+
¦TITEL: Tehnologii actuale de însilozare a nutreturilor / Vintila, Mircea.¦
¦Bucuresti : Ceres, 1989. - 244p.. - 973-40-0071-3                        ¦
¦UDC:   633.3                                                             ¦
¦MAIN:  ^a633.3^eForage plants except grasses^xField crops and their      ¦
¦        production                                                       ¦
¦PDES:  ^z633.3^eForage plants^fPlantes fourragères^rPlante furajere      ¦
¦PNDES: ^y633^z633.3^rPlante de câmp                                      ¦
¦LDES:  ^y633^z633.3^eField crops^fPlantes de culture^rPlante de cultură  ¦
¦LNDES: ^y633^z633.3^eCereals^fCéréales^rCereale                          ¦
¦BCUB#: 42731                                                 MFN: 9814   ¦
+-------------------------------------------------------------------------+
```

Figure 48. Example of different degrees of specificity in automatic indexing of a document

The ideal situation is to have search results as in the 3 sets below:

```
Set     Data Base      Hits     Query element
---     ---------      ----     -------------
1       BCUB           81      "709=633"
2       BCUB           81      "PMBE=FIELD CROPS"
3       BCUB           81      "LMBE=FIELD CROPS"
```

This will be indeed “one utterance for one subject” in any of the search modes the user might choose. And that’s the way it happens in the experimental database in most of the searches performed to demonstrate the retrieval power of the two thesauri (see also Search No. 1 and Search No. 2 in the foregoing). The difference between the two is that LTHES only permits searching for information on broad subjects – such as the above mentioned one for ‘Field crops’ – whereas PTHES enables more refined searches resulting in more precision and hence higher relevance of the search results. To get the expected response for a query statement like ‘*Cereals’* or ‘*Textile plants’* or ‘*Aromatic plants’* or ‘*Forage plants’* the searcher using the small thesaurus will have to browse all the 81 records retrieved when using the query ‘*Field crops’*. An alternative solution would be the use of each of these queries in particular and retrieve all the documents automatically indexed with these descriptors from the Pocket Edition thesaurus in only one step.

Depending on the expertise of the searcher since the experimental database is provided with more than these two methods of approaching the information in its document collection the searcher is offered several other search methods. They have been described earlier and we shall not resume the demonstration about how they work.

As well known by everyone familiar with the structure and functioning of the UDC as a classification scheme, not all the UDC numbers are usable the way they are found in the tables. Many of them have to be built up by synthesis and this is done either through common auxiliaries or through special auxiliaries according to the instructions given in the schedule. Among the combinations built by means of the common auxiliaries of language those meant to express aspects of linguistics and literature are most frequently used. We give below several examples of searches to illustrate the possibilities offered by our indexing language for retrieval of this kind of information.

As previously stated (see **§5.4**) a combination of descriptors that can effectively be used in multiple variants at the moment of searching seems to be the appropriate solution for dealing with UDC numbers that are not found as such in the MRF. In our case the name of the language and the aspect concerning that particular language are combined. For this purpose the same algorithm developed by Riesthuis (1998) is used in factoring or decomposition of the complex UDC notations so that each component part can be subsequently used in an indefinite number of other combinations. The special auxiliaries for literary genres are treated by the factoring algorithms just as well. Our tests focus on Romanian language, a language that was not included among the few examples of combinations given in the MRF.

Several alternative searches were conducted using different search methods for the same search topic i.e. ‘Romanian language’. As much as descriptors are concerned, the first two search modes use descriptors from the two UDC-based thesauri; the sixth search uses manually assigned descriptors and the seventh also a descriptor from a UDC-based thesaurus, PTHES, one which is assigned automatically starting from erroneous UDC numbers that couldn’t be found in the MRF left-truncated to the first digit of the class, i.e. 8).

The first two search sets have almost the same result. The two types of descriptors used as queries in those have been controlled by a third search, using field 711 for ‘Romanian’ as the counterpart of a main number. This set of retrieved records includes translated documents.

The lower number of records retrieved (28024) demanded the use of a Boolean operator to see what the difference consists of.

Browsing the result of the fourth search makes it clear that the difference comes from documents whose subject include the common auxiliary of language to express the language of translated documents. Although the difference is not mathematically correct (there are still typing errors here too) it gives an idea about its meaning.

```
Set  Data Base  Hits   Query element
---  ---------  ----   -------------
1    BCUB       30749  "PMBE=ROMANIAN"
2    BCUB       30770  "LMBE=ROMANIAN"
3    BCUB       28024  ? v711^e : 'romanian' and v711^e : 'literatures'
4    BCUB       2837   #1 ^ #3
5    BCUB       16680  "DE=LITERATURA ROMÂNA" + "DE=LITERATURA ROMÂNA
                       (AMERICA)" + "DE=LITERATURA ROMÂNA (BELGIA)" +
                       "DE=LITERATURA ROMÂNA (BUCOVINA)" + "DE=LITERATURA
                       ROMÂNA (BULGARIA)" + "DE=LITERATURA ROMÂNA
                       (CANADA)" + "DE=LITERATURA ROMÂNA (ELVETIA)" +
                       "DE=LITERATURA ROMÂNA (FRANTA)" + "DE=LITERATURA
                       ROMÂNA (GERMANIA)" + "DE=LITERATURA ROMÂNA (ISRAEL)
                       + "DE=LITERATURA ROMÂNA (IUGOSLAVIA)"+
                       "DE=LITERATURA ROMÂNA (MACEDONIA)" + "DE=LITERATURA
                       ROMÂNA (REPUBLICA MOLDOVA)" + "DE=LITERATURA ROMÂNA
                       (UNGARIA)" + "DE=LITERATURA ROMÂNA VECHE"
```

Another search was made using manually assigned descriptors as query. ‘Romanian literature’ with all its variants (such as Romanian literature published in countries other than the country of origin) makes the subject of 16680 documents manually indexed with this subject heading. *Figure 49* shows an example of the bibliographic records retrieved in the 5th search set.

```
+    38 / 20018 ----------------------------------------- Format: BCUB  --+
¦TITEL: Cartea amagirilor : [eseuri] / Cioran, Emil . - Bucuresti :       ¦
¦       Humanitas, 1991. - 224 p.. - 973-28-0198-0                        ¦
¦BDES:  Literatura româna                                                 ¦
¦BDES:  Eseuri                                                            ¦
¦UDC:   821.135.1-4                                                       ¦
¦SIMP:  821.135.1-4                                                       ¦
¦MAIN:  ^a821.135.1-4^eRomanian - Essays^xLiterature                      ¦
¦PDES:  ^y821^z821.135.1-4^eLiteratures of individual languages           ¦
¦       ^fLittératures relatives à des langues particulières^rLiteratura  ¦
¦        limbilor individuale                                             ¦
¦PDES:  ^z82-4^eEssays^fEssais^rEseuri                                    ¦
¦PDES:  ^z=135.1^eRomanian^fRoumain^rRomâna                               ¦
¦PDES:  ^z821^eLiteratures of individual languages^fLittératures relatives¦
¦        à des langues particulières^rLiteratura limbilor individuale     ¦
¦PNDES: ^z=135.1^eRumanian                                                ¦
¦LDES:  ^z82-4^eEssays^fÉssais^rEseuri                                    ¦
¦LDES:  ^z=135.1^eRomanian^fRoumain^rRomâna                               ¦
¦LDES:  ^y82^z821^eLiterature^fLittérature^rLiteratura        MFN: 678    ¦
+-------------------------------------------------------------------------+
```

Figure 49. Bibliographic record showing the treatment of concepts that are not found in the MRF

In case of Class 9 for instance, another example of built up classification numbers, things are less complicated. The only difficulty is the addition of the descriptor corresponding to the common auxiliary of place to the main descriptor representing the discipline ‘History’. Here again there is no descriptor for the main class number belonging to PTHES but there are descriptors for both numbers from LTHES (Figure 50).

```
+- 28 / 56 -------------------------------------------- Format: BCUB  ---+
¦TITEL: Studies in Romanian history / Dennis Deletant, Alexandru Dutu.   ¦
¦- Bucharest : Editura Enciclopedica, 1991. - 352 p. - 973-45-0005-8     ¦
¦ BDES:  Istorie                                                         ¦
¦ BDES:  Romnia                                                          ¦
¦ UDC:   94(498)                                                         ¦
¦ SIMP:  94(498)                                                         ¦
¦ PLAC:  ^a(498)^eRomania. Republic of Romania                           ¦
¦ MAIN:  ^a94^eGeneral history^xHistory                                  ¦
¦ PDES:  ^z(498)^eRomania^fRoumanie^rRomânia                             ¦
¦ LDES:  ^z(498)^eRomania^fRoumanie^rRomânia                             ¦
¦ LDES:  ^z94^eRegional history^fHistoire régionale^rIstorie regională   ¦
¦ LNDES: ^z(498)^eRoumania                                               ¦
¦                                                          MFN: 150185   ¦
+------------------------------------------------------------------------+
```

Figure 50. Example of automatically assigned descriptors in Class 9

7.3 Conclusions

Starting from the good old "Rules for a Dictionary Catalogue", formulated as early as 1876 by Cutter and particularly from his assessment on the importance of the *specific entry*, this chapter attempts at demonstrating the extent to which information retrieval is influenced by the specificity of the information language used in searching.

Out of the four types of catalogues existing in Cutter's time Cochrane argues that the ideal catalogue defined by Cutter, the combined catalogue, is in many ways comparable to the online catalogue nowadays provided that it accomplishes Cutter's formulated requirements since:

1. they permit browsing but also navigation, thus bringing documents dealing with related subjects together;
2. they endeavour to reduce to the minimum possible information loss;
3. they provide the users with help messages that assist them during the search process

By means of the multiple search examples provided in this chapter we demonstrated the effects different information languages and different degrees of specificity within the same type of information languages have on information retrieval.

In order to fix the theoretical background of our demonstration we focused on some very important concepts like *recall*, *precision* and *relevance*. In so doing we cited the definitions of these concepts as most outstanding information scientists gave them.

Our objective was to prove that the higher specificity of the information languages used in an information system implies higher precision therefore higher relevance of the search results. To sum up, we came to the following conclusions regarding the impact of specificity of the information language on the quality of information retrieval:

- There has to be a direct relationship between several factors within the framework of our approach. The *size of the document collection and the level of specificity of the information language* determine the performance of the information retrieval and Salton and McGill appreciate as crucial to have the right balance between them, precision being of utmost importance for very large collections. Therefore, the correlation between the collection size and the specificity of the information language determine the performance of the information system as a whole.

- The combined information language we compiled and explored in this research is based on a selection of UDC numbers that is different from the International Medium Edition employed in the classified catalogue that our experimental database is taken from. Two multilingual thesauri with different levels of specificity have been developed and used in information retrieval procedures to show their performance. Our tests proved that there is a direct relation between the *level of specificity of the classified catalogue and that of the UDC selection the thesaurus is based on or else the precision rate is affected.* In the first instance LTHES, the broad thesaurus, proved to be too general for the in-depth classified catalogue its *broad terms being not able to distinguish the less relevant documents from the truly relevant ones.* The high recall thus obtained implies browsing great numbers of retrieved records in order that the user eventually finds what he was looking for at the expense of his own time. The searches testing the second thesaurus, PTHES, resulted in improved precision hence higher relevance of the documents retrieved (see the tests using the narrower terms of 'Field crops' as search statements).

- The automatic indexing procedure used in our research proved its superiority in terms of precision and relevance of information retrieval since it is based on several strong attributes of the information language used in our study: *indexing consistency and control of indexing terms resulting from the one-to-one correspondence between the UDC number used as a basis and the descriptor derived from its caption*. While manual indexing shows sometimes the indecision or confusion of the indexer the automatic indexing will always overcome such shortcomings by assigning the established descriptor – always the same – to a certain UDC number. *Ambiguity* normally occurring in manual index terms with little control over them *is eliminated*. As long as the UDC number assigned to a bibliographic record is found both in the MRF and in the selection of UDC numbers the thesaurus is based on the automatic indexing is feasible. Accuracy of information retrieval is highly dependent on the correctness of the classification notations assigned in subject representation. The rules of building up complex UDC notations have to be respected for an improved quality of the automatic indexing (see the examples using 'Experimental psychology' as query).

- As a rule the search results in three search modes (the PTHES mode, the LTHES mode and the 700 field mode – used as controlling device since it reflects the meaning of the UDC notation in the classified catalogue) are quasi-equivalent with little if any difference deriving in most of the cases from either obsolete UDC notations, or typing errors or extremely long strings of digits that are difficult to be managed.

This is in straightforward statements the advantage of having a high specificity in the information language used in information retrieval. It has been demonstrated that the higher the specificity the higher the precision of the results and hence the improved relevance of the documents retrieved. An additional advantages of our UDC-based thesauri accounting for the consistency in indexing and control on indexing terms derive from the classification system notation underlying each of the indexing terms. Furthermore, the multilingual character of the indexing terms adds extra value to the information language enlarging the scope of the retrieval.

CHAPTER 8
FINAL REMARKS AND GENERAL CONCLUSIONS

8.1 Purpose of the research

The challenge of easier access to the information contained in bibliographic databases where subjects of the documents are represented by numeric classification codes that seem unfamiliar and rather frustrating to the average end-user – in our case the UDC notations – makes the objective of our research. Furthermore, our purpose was to demonstrate that multilingual access to information contained in bibliographic databases is possible via multilingual descriptors derived from UDC numbers. The present research has demonstrated that this can be automatically done, while subjects of the documents are manually indexed with UDC notations alone in a classified library catalogue.

In order to address the multilingual aspects of subject indexing and information access (or information retrieval) in bibliographic databases we have developed our approach from two perspectives:

- the existence of more than one information language (either enumerative or word-based) used in indexing and information retrieval;
- the existence of word systems based on vocabulary elements of more than one natural language, in our case Romanian, English and French.

Our ultimate goal was to prove that information retrieval can be remarkably enhanced by the use of multilingual thesaurus terms based on an intermediate language, in our case, the Universal Decimal Classification (UDC). For this purpose we explored and used the basic qualities of the UDC such as: language independence, hierarchical structure, terminological richness, consistency and control of notation.

The continuing role classification has in information retrieval was underlined and its logical structure favourably used in building updated tools to enable enhanced access to information. Specificity was seen as a primary precision device along with other methods meant to increase the reliability of the information language that makes the object of this research.

8.2 Methodological outlines

In the introductory part of our research a number of *linguistic aspects of information languages and information retrieval* are discussed The examples used were meant to illustrate the difficulties the searchers are confronted with in their attempts to find potentially useful information. It was illustrated that the expected good results are not always met and that shortcomings are mainly represented by: multiple character sets, particularly by diacritics, transliteration systems, name variants, acronyms, and above all by the ambiguities generated by homonyms and polysemic words. These controversial linguistic problems are examined in as much as they occur in the three languages employed in the study.

Problems of compatibility and integration of information languages are also discussed. Since quite an important part of our research is based on it, compatibility is seen in its major types namely: *structural and semantic* compatibility. As far as its coverage is concerned *full and partial compatibility* are considered along with a few hints at the complementarity of indexing languages.

Issues of *homonymy* and hence *ambiguity of meaning* are argued about, the strongest argument bringing forward the role of context as disambiguating device. Needless to say that the main point at issue here is vocabulary control.

In order to provide a comprehensive image of the *state-of-the-art in the multilingual access problems* nowadays some of the latest developments in the field are described. Particular emphasis is given to the project of CoBRA+ working group on Multilingual Subject Access (MACS) that had encouraging results though progress is quite slowly made. The ETHICS system in Zurich is also described and the more so as it was based on the structure of the UDC as well. For such an approach to the much discussed topic of multilingual access the achievements of the TREC conferences are inevitable so outlines of the latest experiments of the Cross-Language Information Retrieval (CLIR) tracks are also considered.

Thoroughgoing methodological *issues of harmonizing a classificatory structure with a thesaurus structure* are discussed and significant *aspects of multilingualism* are presented in a rather comprehensive manner. The UDC-based multilingual thesauri employed in this research were built according to the norms established by the existing international standards – ISO 2788 (1986) and ISO 5964 (1985). The first thesaurus (LTHES) has some particular features generated by its broad coverage on the one hand and its lack of specificity on the other. To have a better perception of the problems arisen another thesaurus with a higher degree of specificity was developed in order to prove by comparison the influence specificity has on the performance of the information retrieval system. It is in this chapter that the UDC is reiterated as candidate to the position of an intermediate or a switching language and pros and cons of different authors are considered and debated..

In order to provide evidence to all previously stated issues of multilingual access to information a case study is presented in Chapter 6. To accomplish that, an experimental database was created consisting of three segments: in the first place a bibliographic segment, secondly, the machine readable version of the UDC, the Master Reference File (MRF), and lastly, a shortened version of the MRF containing high level UDC notations that represent each of the disciplines/classes by their first digits. More than one language version of the MRF was used. It is important to underline that the terms – descriptors and non-descriptors – of both multilingual thesauri, discussed in detail formerly in this thesis, were embedded in the MRF in order to make the automatic indexing procedures possible.

As already stated the ultimate purpose of our case study was to give an answer to the question: does converting the UDC numbers into corresponding vocabulary terms enhance information retrieval? If it does, to what extent? But then, some other questions arose such as: should those words belong to a controlled or to an uncontrolled vocabulary? What is the effect on information retrieval in either of these situations? In order to find the answers to these questions the following preliminary steps were taken: 1) "Cleaning-up" the database; 2) Mapping the UDC numbers with descriptors; 3) Making multilingual subject headings available.

The task of the first step in our approach was considerably lessened by a helpful feature of CDS/ISIS – the software used in building the experimental database – *the global changer.* This function permits the change of the UDC notation in a matter of minutes on condition that a search query is formulated. By running the conversion programme over the resulted UDC numbers the UDC corresponding text is changed accordingly.

Our experiment was carried out with the help of several conversion programs that mapped the UDC numbers onto their corresponding descriptors in specially designed fields of the bibliographic records in the experimental database. This was done gradually according to the complexity of the programs. Some of the descriptors were simply added in the bibliographic database as they were found in the thesaurus as a result of the existence in the MRF of the UDC number these were based on. The difficulties arose when the UDC numbers in the database were the result of building up procedures. For complex UDC notations using either the common auxiliary tables (all existing in PTHES as descriptors) or the special auxiliaries, some complicated programs were created many of them using the same algorithm procedures as in the case of decomposition of the complex UDC notations

In the end we demonstrated that information retrieval can be enhanced by using automatically assigned descriptors derived from the UDC captions in as many languages as available. Our demonstration was based on real-life searches performed in the experimental database and using a variety of search methods. Specifically the search methods addressed the descriptive part of the bibliographic records on the one hand and the subjects of the documents on the other, with query statements ranging from titles and title words to classification codes, words from captions and descriptors. The descriptors used as queries were those initially existing in the database – Romanian manually assigned descriptors – plus automatically assigned multilingual descriptors having as source one or both of the UDC-based multilingual thesauri.

8.3 Conclusions

The most important findings of our research have already been exposed in the previous chapters of this thesis (see the conclusions formulated in **§6.7** and **7.3**). However, we shall not hesitate to underline the fundamental one that most of this research is focusing on i.e. *information retrieval is possible via text added and thesaurus descriptors derived from the UDC notations in more than one language with notably improved results*. Once the MRF and the thesaurus terms are multilingual and the link exists between the UDC notations and the UDC text and descriptors respectively, the different language variants of those can be made automatically available to a wider range of users conferring added value to the information retrieval facilities.

At a preliminary indexing stage, the addition of text to the component parts of complex UDC notations in 700 fields may provide the indexer with a good device for detecting errors in classification. Likewise it may suggest indexing terms according to the meaning of the classification notation if the indexing is manually and not automatically done. The association of words can act interactively influencing the user's search behaviour. Domain specific searches can be associatively derived from the context of the UDC captions enabling further steps in the search strategy.

The *level of specificity* of the first multilingual thesaurus used in our study (LTHES) yet not always concordant with the depth of the classification notations used by the indexers is a shortcoming relatively easy to overcome. The *temporary solution* to this problem is the existence of *a greater number of non-descriptors* making reference to the preferred terms. The *long-term solution* and a thoroughgoing one is *the adjustment of the specificity level of the multilingual thesaurus to the requirements of the database*. That was our rationale in proceeding at the enlargement of the thesaurus sizes to the coverage of the UDC Pocket Edition (1999). Consequently the more specific thesaurus (PTHES) was created. The non-descriptors are numerous in each of the participating languages and queries can be formulated to include them by means of search codes separately defined for English, for French and for Romanian. After they are selected from the term dictionary the corresponding records in the MRF are displayed and the preferred terms can be visualized and considered for subsequent searches in the bibliographic database.

For a better perception of the impact of specificity on the retrieval power of a UDC-based multilingual thesaurus our approach was undertaken comparatively (like other stages of our research) with a view to bring clearer evidence of the assumptions and hypothesis formulated.

A brief look back at the previous paragraph gives us the opportunity to say more about the *structural configuration of different information languages*. There is a question of *language compatibility* at issue here. According to one of our findings, the basic requirement for two indexing languages to be compatible is that their respective configurations are at least partly compatible if not fully compatible (in both meaning and structure). If these requirements were not fulfilled, many of the features and functionalities of the described information retrieval tools would not have been appropriately used. Hence the requirement of compatibility of information languages that can co-exist and be used alternatively in accessing information in our case study.

The experiment presented in the preceding chapter proved that *there is a reasonably high degree of compatibility between the UDC numbers used in indexing and the descriptors used in information retrieval*. We also demonstrated that there is a still higher degree of compatibility between the UDC notations separated in their meaningful component parts and the descriptors derived from them. Hence we can say that there is a considerable degree of compatibility between the four information languages used in our experimental database. The degree of this compatibility is variable and it strongly depends on the intrinsic relation between the UDC numbers and the descriptors provided during the indexing process, on the one hand and between the UDC numbers and the different language variants of the thesaurus on the other.

We can also speak about the *complementarity of the four indexing languages* in that alphabetical additions existing in the UDC and taken over by the 706 field are not represented in the automatically added multilingual descriptors. The alphabetical additions and the chronological aspects of the subjects of documents need a special treatment and they have to be included in separate authority files. Alternatively they can be manually entered in the appropriate fields of the database.

Once this research has proved its *feasibility and advantages over the manual indexing with descriptors*, sooner or later this hard manual indexing work should be given up and efforts should be concentrated on building a high quality interdisciplinary thesaurus based strictly on the UDC Master Reference File. Once this thesaurus is accomplished the classification notations are the only things needed for subject representation. The intimate relation between the UDC number and the corresponding descriptor can be appropriately hidden and the searcher can use directly the textual vocabulary elements in searching regardless of the mapping between the two information languages.

The most important outcome of this simplified indexing method, although highly demanding in terms of correctitude and accuracy, is that *postcoordinate search via words from the UDC captions and via multilingual descriptors is made possible*. Postcoordinate searching by means of logical operators: OR, AND, NOT can be used to refine the search and particularly the use of the Boolean operator NOT can filter the irrelevant information.

A major advantage of using the above-described development in classified catalogues is the avoidance of the painstaking work of re-indexing the classified documents. By applying our work method step by step, the updating of the outdated UDC numbers can be performed and as long as the UDC numbers are found in the MRF, descriptors are automatically added and available for searching. We insist on the crucial importance of classification numbers being correctly assigned and of consistency being respected throughout the bibliographic databases. Assigning the same category of documents to the same class is not always easy and competence and skills are required for each of the discipline/domain in question. That makes us consider J.-E. Mai totally right to say, "classification is an art that goes beyond any formal rules of logic and science" (Mai, 2000, 27). As long as the classification notations are appropriately assigned and the UDC grammar rules are strictly followed, all the preliminary procedures to information retrieval go the right way. It is only on this condition that our research can have the expected results.

From all the assumptions that we made and demonstrated here we can draw as conclusions for the whole study a *few principles underlying the building of thesauri based on a semi-enumerative hierarchical classification system like the Universal Decimal Classification*:

1. Descriptors derived from the UDC numbers should have each only one number (or component part of a complex one) as counterpart and not more. In other words, the UDC number should correspond to one descriptor or prescribed combination of descriptors. Examples are easily found in Class 8 Linguistics and literatures but also elsewhere in the UDC tables where the notations are built up by parallel subdivision.

2. The *general* or *common auxiliaries of bibliographic form* having often homonymous equivalents that overlap with topical descriptors should be either accompanied by parenthetical qualifiers or kept in separate fields. This way there will be no confusion whether the subject of a document is, say, 'Periodicals' or this is just the bibliographic form of that document.

3. The *alphabetical additions* as much as the *common auxiliaries of time* can never be controlled neither predicted while included in a thesaurus. Therefore they must be treated a *unique entities* after the application of the decomposition algorithms. Alternatively, they could be kept in separate (authority) files, particularly recommended for proper names, and henceforward controlled.

4. *Node labels* or *dummy terms* (NISO, 1994) can be profitably used to show the logical hierarchies necessary in a thesaurus. They are meant to group narrower terms in categories. They have similar functions to BT's but they are not descriptors therefore they should not be used as indexing terms.

5. For those *UDC numbers that collect the meanings of lower levels of subdivision* a separate treatment is recommended according to each particular case (see **§5.5**).

Our approach has a certain advantage undisputable among others (see **§6.7**) i.e. the multilingual thesaurus being embedded in the database is always available to the searcher with its entire structural configuration.

However, a real shortcoming of our approach is that the interterm relationships (BT's, NT's and RT's) specific to the syndetic structure of a thesaurus are not functional in the database. Consequently, the user can only use the browsing function and not the navigating function of such a database[25]. At first sight the hardships and complex problems of building a multilingual thesaurus structure based on the UDC might prove ineffective. A list of equivalent descriptors to the UDC notations implemented in the MRF would do the job with equally good results. The navigating function can theoretically be achieved through the UDC notations that permit the user to go up and down the schedule's hierarchical structure. This is not practically applicable in our case, yet.

A good reason to motivate our endeavour is that a printed thesaurus can successfully accomplish this functionality not available in CDS/ISIS. This way the UDC-based multilingual thesaurus can be used as a tool to assist the information retrieval procedures by showing all the relations such a thesaurus is equipped with. Not only the hierarchical relations, that are visible in the UDC-based display of the thesaurus, but also the associative ones will help the user in making out logical associations and better understand the knowledge domain he is interested in (see Appendices No. 2 and 3 at the end of the thesis for samples of both thesauri).

For all that, MTM3, the thesaurus-building program used for LTHES and PTHES, has a quality that cannot be overlooked: it controls the terms i.e. the software does not permit the existence of a descriptor twice in the thesaurus (see **§5.2**). The alternative of providing equivalent descriptors to the UDC numbers in the MRF directly will save time and intellectual effort but will not permit any control on terms. As a result doublets might occur and cause ambiguities and confusion in searching.

Before concluding we enumerate some guiding principles along with methodological issues that we followed throughout our research for a better understanding of our objectives:

- There are certain requirements that the UDC-based thesaurus builder has to be aware of when approaching the creation of such an indexing and retrieval tool. In the first place the thesaurus building guidelines and rules should be accurately followed. Should any restrictions be imposed on the structure or usage of the thesaurus, they have to be dealt with such a way that on the whole the thesaurus remains within the limits of acceptance of the existing international standards. One has to keep in mind the two sides of such an approach:

    - the thesaurus building side with all its particularities
    - the thesaurus usage side giving account on its effectiveness

[25] However, depending on the way the OPAC is configured and on the software used, navigation is possible through the hierarchies of the UDC notations and their built-in relations.

- Another requirement would be one of a totally different nature, i.e. the thesaurus builder should know what kind of a bibliographic database the thesaurus is meant for (in other words, the document collection), the growing rate of the database, how much detail the users need and what category of users the collection is addressing.

- If the UDC-based thesaurus is implemented in a database whose bibliographic records are indexed with UDC numbers alone and those numbers are included in the thesaurus structure, there are two ways the automatic assignment of descriptors should be made possible:

  - based on links between bibliographic databases and authority files containing both UDC numbers and thesaurus terms;
  - within the same database with fields specially designed to hold different indexing languages or indexing methods.

- The higher the degree of specificity the more concepts have homonymous expressions therefore disambiguation is demanded; the main disambiguating device used in our approach is the addition of a qualifier in brackets, specifying the discipline that concept belongs to. Context is a powerful disambiguating device and apart from this method working in the thesaurus search mode the 700 field search modes are provided with a subfield ^x that proved to be extremely helpful whenever information on context was needed. As a matter of fact, homographs and scope notes, while increasing the level of pre-coordination increase also the precision, preventing retrieval of non-relevant documents (Aitchison & Gilchrist, 1987, 8-9).

- Speaking about the 700 fields, in many of the searches conducted especially to demonstrate the impact of specificity on the retrieval power of our indexing language we used it successfully. In so doing we permanently had a device to compare the effectiveness of other search methods used by controlling their result with results of the 700 field searches. The reliability of the later is doubtless because the information these fields and subfields contain is nothing but the text attributed to the UDC numbers or their parts in the classification notations.

- For concepts that are not expressed in the tables but are possible to be built up with the existing numbers (as for example those expressed by parallel subdivisions or special auxiliaries) the solution used in PTHES is different from LTHES; in the former we adopted the combination of descriptors (e.g. Romanian linguistics and Romanian literature) whereas in the later, the list of descriptors the thesaurus is based on includes a number of predictable combinations in an enumerative sequence (See the languages and literatures part of the LTHES thesaurus in Appendix No. 2).

- There are also some weak points we still have to deal with in the future. One of them is that the auxiliaries for peoples and ethnic grouping go wrong in conversion because the descriptors are not taken from the MRF (e.g. (=135.1) denoting Romanian people or Romanians). So even if they are included in the thesaurus they cannot find their match in the MRF. The solution for this kind of problems is yet to be found. Another weak point is that relations between the thesaurus terms are not shown. Since nowadays *information discovery* is becomming by far more important than *information retrieval* the relational network or thesaurus terms grow in importance. For that reason the presence of as many as possible synonyms, near-synonyms and also related terms is considerably practical for the purposes of a reliable information language. Navigation rather than browsing is a functionality that permits an expanded scope of and hence a larger view on the relevant information available in an information system (see **§5.3**).

- With a view to enhance the quality of the thesaurus content and develop it in compliance with the users needs, it is highly recommendable that an additional field is provided in the system for suggestions of new entry terms or changes in the already existing ones. By users here we mean both indexers and searchers that should nominate new candidate terms. This has to be done with care in order to always keep the rule of only one descriptor (or prescribed combination of descriptors) for one classification number or else the principle behind our approach is not working any more. A history note of the terms recording the changes those terms undergone would be recommendable.

- The display format also has to be as explicit as needed to meet the expectations of the end user. Help messages should be implemented so that the system can explicitly assist the user in his search behaviour. Interactivity is highly recommendable in this respect the suggestions of word associations as much as serendipity having a great potential in information discovery.

- The objectives of our approach are so close to those of the Audacious project: to use a system that would permit indexing and searching using a controlled natural language vocabulary of local choice but having also a table of equivalences between the UDC and the natural language vocabulary; this system would take advantage of the logical structure and hierarchical notation of the UDC without the users being aware of this; the UDC being the internal form of indexing, the users of the system would address their queries in natural language words without regard to the original indexing language used.

- The main reason for using thesaurus terms rather than classification codes in information retrieval is that words are friendlier to the user than numbers. As long as there is a reasonable high degree of compatibility between the two, it is advisable to switch in favour of the thesaurus terms for information search and retrieval. Having the descriptors as alternative indexing and retrieval device the friendliness of the retrieval system is much enhanced and, which is of utmost importance, the reliability of the information system itself is growing.

- The user-friendliness of information retrieval systems being a highly regarded imperative there are sceptical views in that the class numbers are not desirable any more and have no future. However, the numerical representation is advisable to be preserved in such systems in order "to preserve the original meaning and scope of a particular concept" (McIlwaine, 2000).

- As several paragraphs throughout this dissertation prove it is essential to get more languages in such projects in order to provide equal opportunities to information access regardless of the language barriers.

One last word: traditional resources and tools, especially the standard ones that proved their validity in time, being authorised by long lasting practical use should not be given up. On the contrary, their continuous updating is highly required as they were built on reliable fundamental principles. The dynamic evolution of knowledge and the new concepts evolving in each discipline/domain every day impose rapid and effective adjustment of knowledge organisation systems able to deal with change. A simple example would be clarifying: recently the world has witnessed a new type of tourism i.e. the first trip of a human that was not an astronaut in the extraterrestrial space. He was called a ‘space tourist’. A classification system like the UDC is perfectly able to represent such a topic given its flexible structure and relational capabilities. For a word system that operates with free text, introducing a new category like “space tourism” would be of little consequences. For a controlled vocabulary though, given its rather rigid structure, things would not be that simple. Adding new terms in such a structure implies updating work on a regular basis with all the consequences on the entirety of the system.

What is then recommendable to be used as knowledge organiser and information retrieval tool? Which is the system that brings the fastest and the most efficient (or more precise) information in response to the (more and more demanding) user’s information need? Needless to repeat but our own answer to such questions is that the UDC is one of the best, as demonstrated in the present thesis, on condition that access with derived descriptors from a thesaurus in as many languages as possible is provided.

SAMENVATTING

Informatie is uitgebreid voorhanden. Niettemin beperken taalbarrières in de praktijk het algemeen gebruik ervan vaak.

De kern van dit onderzoek is het linguïstisch aspect van de informatietalen en meer speciaal de mogelijkheden tot een verbeterde toegang tot informatie te komen via een meertalige indexerings- en terugzoekinstrument gebaseerd op een universele classificatie.

Dit onderzoek wil aantonen dat UDC-notaties, die in een document behandelde onderwerpen voorstellen, gebruikersgerichter kunnen benaderd worden via op UDC gebaseerde descriptoren uit een meertalige thesaurus. Wanneer een bibliografische database die gebruik maakt van UDC, verbonden is met de Master Reference File van de **U**niversele **D**ecimale **C**lassificatie (UDC- MRF) en een op UDC gebaseerde thesaurus, kunnen de documenten uit deze database automatisch geïndexeerd worden via de thesaurusdescriptoren. Dit geeft een vlottere toegang tot de database en bijgevolg betere zoekresultaten. Een dergelijk geïntegreerd systeem (een bibliografische database, de officiële UDC-versie die als tussentaal optreedt en een op UDC gebaseerde veeltalige thesaurus) biedt voordelen zowel voor de indexeerder als voor de gebruiker.

Het onderzoek bestaat uit verschillende onderdelen.

1. Een taalkundige benadering van informatietalen vormt noodzakelijkerwijze het eerste deel. In die zin worden de belangrijkste kenmerken van de drie talen die in het onderzoek aan bod komen (het Engels, het Frans en het Roemeens) voorgesteld.
2. Vervolgens worden veeltaligheidsaspecten, zoals die zich voordoen in de opslag en het terugzoeken van informatie onderzocht. Een kort overzicht wordt gegeven van de specifieke kenmerken van de belangrijkste ‘ingangen’ in een veeltalige context.
3. Hoe de verschillende talen onderling met elkaar verbonden kunnen worden is een discussiepunt. Daarom worden in een volgend hoofdstuk compatibiliteit, omzetbaarheid en integratie van informatietalen behandeld.
4. Bij een overzicht van de huidige ‘trends’ in de meertalige toegang tot informatie werd speciaal aandacht besteed aan het werken met onderwerpstoegangen in verschillende alfabetten en letterschriften in meertalige landen.
5. De methodologische benadering van het bouwen van op UDC gebaseerde veeltalige thesauri wordt besproken. Deze benadering is gebaseerd op ervaringen met het opzetten van twee thesauri van onderscheiden specificiteit.
6. Verder worden de online toepassingen van de UDC-gebaseerde thesauri behandeld. Deze zijn gestoeld op testen met een experimentele database (in CDS/ISIS format) bestaande uit 3 onderdelen : een bibliografische deel, de UDC Master Reference File met meertalige descriptoren uit de twee thesauri, een verkorte MRF versie voor contextbenadering.
7. Vervolgens wordt de invloed van de factor specificiteit op de resultaten van het zoekproces onderzocht en meer bepaald de invloed op de vangst, de relevantie en de precisie. Het onderzoek in de experimentele database toont duidelijk de sterktes en de zwaktes van elk van de twee gebruikte thesauri aan.
8. Het afsluitend hoofdstuk beklemtoont opnieuw het doel van het onderzoek, beschrijft de gevolgde methodologie en formuleert conclusies. In het algemeen besluit worden de haalbaarheid en de effectiviteit van de benadering in onze studie toegelicht.

9. De drie bijlagen zijn bedoeld als leidraad bij de experimentele database. De eerste geeft informatie over de betekenis en het gebruik van de toegangscodes voor verschillende methoden ; de tweede en de derde geven voorbeelden uit de twee UDC-gebaseerde meertalige thesauri.

* * *

In feite bestaat deze verhandeling uit twee grote delen : een theoretische benadering van het onderwerp waarin de veeltaligheid bij het terugvinden van informatie vanuit verschillende oogpunten wordt benaderd en een voorstelling van het experiment waarin de vooronderstelde hypotheses aangetoond worden.

De hypothese van ons onderzoek is dat de inhoudelijke ontsluiting (en bijgevolg de toegang) van documenten op basis van UDC-notaties kan verbeterd worden door het automatisch toevoegen van meertalige descriptoren afgeleid van een op UDC gebaseerde thesaurus. Deze thesaurus bevat telkens de UDC code waarvan de descriptor is afgeleid. Zoals verderop zal blijken heeft deze functionaliteit interessante gevolgen.

Bij opzoekingen zijn thesaurustermen gebruiksvriendelijker dan classificatiecodes. Voor het terugzoeken van informatie is het, wanneer beide mogelijkheden beschikbaar zijn, interessanter thesaurustermen aan te bieden voor zover er een hoge mate van compatibiliteit is met het classificatiesysteem. Het automatisch toevoegen van descriptoren in verschillende talen betekent een tijdswinst bij het indexeren. Bij het zoeken van informatie is de intellectuele inspanning geringer. Daardoor wordt het aantal potentiële zoekers ook aanzienlijk uitgebreid. Dit systeem zal beter tegemoetkomen aan de behoeften van de onderzoeker inzake vangst, relevantie en precisie.

Dit onderzoek naar veeltalige toegang tot informatie in bestaande bibliografische databases begint met een vergelijking tussen documentaire talen en natuurlijke talen. Hierover bestaan verschillende semantische theorieën waaronder deze van Ferdinand de Saussure, Jakobson, Chomsky, Morris, Lyons en Hutchins. De voorstelling van deze theorieën laat ons toe de gebruikte concepten en betekenissen te verduidelijken en de processen van het informatieproces te expliciteren. Tevens identificeren wij de karakteristieken van de twee types talen evenals de onderlinge verschillen.

Beide ‘talen’ hebben een woordenschat die zich uit in vormen en betekenissen. Bij natuurlijke talen bevat de woordenschat elementen die gebruikt worden bij de woordelijke communicatie tussen mensen. Bij documentaire talen gaat het om voorstellingen van onderwerpen behandeld in documenten, zij geven een afgeleid beeld van de kennis opgeslagen in documenten.

In een natuurlijke taal is het een teken van rijkdom dat er woorden zijn met verschillende betekenissen. In een documentaire taal daarentegen is een term met verschillende betekenissen (notatie of descriptor) problematisch waarbij normalisering zich opdringt. Volgens Jacques Manniez moet de ideale documentaire taal streven naar één onderwerp per uitdrukking en één uitdrukking per onderwerp. Documentaire talen gebruiken specifieke notaties (UDC, DDC, LCC) of gestandaardiseerde en genormaliseerde woorden van natuurlijke talen (indexeertalen) om objecten of concepten te benoemen.

De informatietransfert tussen een zoeker en een informatiesysteem is een communicatieproces. In een verbaal communicatieproces hebben we, eenvoudig uitgedrukt, een zender, een ontvanger en een tussen hen uitgewisselde boodschap. De informatietransfert via een informatiesysteem verloopt over verschillende vertaalprocessen.

De zoekvraag wordt in een natuurlijke taal geformuleerd. Afhankelijk van de informatietaal kunnen deze woorden in meer of mindere mate overeenstemmen met de indexeertermen van het zoeksysteem.

De indexeerder vertaalt het onderwerp van het document en reduceert dit tot het essentiële. Een conceptuele vertaling zet de inhoud van het document om in indextermen. Hierbij zijn uiteraard linguïstische aspecten betrokken, de concepten die het onderwerp van het document voorstellen moeten uitgedrukt worden in indextermen zodanig dat de belangrijkste aspecten van het onderwerp nauwkeurig voorgesteld worden.

De paradigmatische structuur van de indexeertalen vertoont vier soorten relaties tussen de indextermen (cf. Hutchins): identiteit, substitutie, inclusie en associatie. Een vergelijking van de verschillende wijzen waarop naar dit onderwerp gekeken kan worden, helpt bij het formuleren van het belang en de functie van thesauri.

Wij vergelijken enkele controversiële aspecten tussen natuurlijke en documentaire talen: homoniemen, synoniemen en taaluniversalia. De wijze waarop deze specifieke linguïstische categorieën in documentaire talen behandeld worden wordt speciaal onderzocht.

Tenslotte worden de drie talen die in dit onderzoek betrokken worden (Roemeens, Engels en Frans) beknopt voorgesteld (geschiedenis en specifieke kenmerken). Meer specifiek wordt aandacht besteed aan het Latijns karakter van het Roemeens, het vrij recent doordringen van Engelse termen in het lexicon van verschillende talen, bepaalde contrasten tussen Roemeens en Engels en de belangrijkste semantische kenmerken van het Frans.

* * *

In de zoekmethodes in informatiesystemen en hun meertalige toepassingen kunnen wij twee situaties onderscheiden :

- zoeken van en toegang tot gekende documenten; documenten in een gekende taal worden gezocht;
- zoeken van en toegang tot onderwerpen; een zoekvraag wordt geformuleerd in de taal van de catalogus (door de zoeker gekend); het zoekresultaat stelt alle documenten voor die in de catalogus beschreven zijn (ongeacht van de taal).

Het zoeken op titels en titelwoorden kan volgende problemen bieden :

- spelling- of schrijffouten (zowel vanwege de indexeerder als van de zoeker;
- de titel is in een andere taal dan de inhoud;
- metaforische of misleidende titel;
- de verschillende uitgaven van hetzelfde werk kunnen verschillende titels hebben;
- verschillende taalvarianten van hetzelfde werk.

Voor de laatste twee gevallen biedt het systeem van de 'Uniforme titel' (o.m. voorgesteld in AACR2) de beste oplossing. Tot de verschillende titels is er op die manier toegang via één (gestandaardiseerde) titel.

Zoeken op auteursnamen kan problemen opleveren bij Latijnse en Griekse auteurs, bij getranscribeerde namen, bij namen met diakritische tekens en bij vertaalde namen. Zorgvuldig beheerde autoriteitsbestanden zijn de beste oplossing. Dit geldt ook voor namen van organisaties. Vele hebben een naam in verschillende talen, letterwoorden en afkortingen kunnen tot verwarring leiden.

Bij het zoeken op onderwerp heeft de ervaring aangetoond dat de ongecontroleerde informatietalen een grotere mogelijkheid bieden om meer documenten terug te vinden dan de gecontroleerde informatietalen. Voor zover de zoekterm voorspelbaar is en men het toeval een kans gunt, kan het zoeken doorheen de hele tekst nuttige informatie opleveren. Grasduinen ('browsing') doorheen een hele reeks documenten is evenwel tijdrovend. Anderzijds biedt het zoeken via classificatienotaties, hoewel deze als gebruiksonvriendelijk beschouwd worden, de minst frustrerende resultaten. Wij willen duidelijk onderzoeken in hoeverre deze 'gebruiksonvriendelijkheid' tot een voordeel kan omgeturnd worden.

Elk van de informatietalen heeft zijn eigen syntagmatische en paradigmatische structuur. Toch zijn ze onderling in zekere mate compatibel. Dit is zeker zo wanneer zij verwijzen naar dezelfde realiteit en zij dezelfde doelstellingen nastreven (het organiseren van kennis en het terugvinden mogelijk maken). Verschillende onderzoekers formuleerden theorieën over de compatibiliteit van informatietalen met het oog op practische toepassingen. Sommige van deze onderzoeken worden in het derde hoofdstuk voorgesteld met de bedoeling deze in ons onderzoek toe te passen.

De compatibiliteit (en gedeeltelijke compatibiliteit) wordt behandeld vanuit structurele en semantische eigenschappen van de informatietalen. Wanneer volledige compatibiliteit niet kan bereikt worden en zolang niet geraakt wordt aan het doel en de coherentie van het indexeren wordt de complementariteit van de informatietalen als compromis nagestreefd. Aan de hand van een aantal voorbeelden tonen wij aan hoe compatibiliteit tot een beter zoekresultaat kan leiden en tot een integratie van informatiebronnen. Een voorbeeld is de Unified Medical Language System (UMLS).

Wij illustreren de toepassing van compatibiliteit en integratie aan de hand van vijf specifieke vakthesauri die wij construeerden op basis van UDC-klassen. Eén ervan, klasse 8, Linguïstiek, is meertalig. Deze benadering laat ons toe bepaalde kenmerken van sommige klassen te beschrijven en neveneffecten en nadelen van onze methodologie te verantwoorden. Aan ambiguïteit en de methodes om dit teniet te doen wordt specifiek aandacht besteed.

Ten slotte, maar niet in het minst hebben wij het over de controle op het vocabularium. Het probleem van de ambiguïteit is ernstiger voor encyclopedische databases. Onderzoek naar de context biedt oplossingen. Lancaster stelt middelen voor om de ambiguïteit op te lossen. Er is een gelijkenis met de middelen die wij toepassen voor de op UDC gebaseerde thesaurus en die wij voorstellen in een volgende hoofdstuk.

In hoofdstuk vier stellen wij verschillende systemen voor die een veeltalige toegang tot documentverzamelingen bieden. Hiermee willen wij een status questionis bieden inzake veeltalige information retrieval. Sommige zijn projecten die nog uitgetest worden (vb. MACS), andere werden om financiële redenen opgegeven (Expo 2000).

Toch zijn er systemen die (met vallen en opstaan) sinds lang functioneren. Dit is o.m. het geval voor het welbekende ETHICSsysteem, toegepast in de ETH bibliotheek in Zürich in een meertalig land. Behalve in Zwitserland wordt ook in Finland en Israël intens aandacht besteed aan de toegang tot onderwerpen in meerdere talen en letterschriften.

Tenslotte betekenen de "*cross-language retrieval*" bijdragen (CLIR) van de Text Retrieval Conferences (TREC) een belangrijke stap voorwaarts door het toepassen van taaltechnologische procedures voor meertalige information retrieval.

De experimenten en de trends in dit hoofdstuk voorgesteld, worden gekenmerkt door functionele principes die ze werkzaam maken zoals :

1. In overeenstemming brengen van termen en meervoudige hoofdingen via verbindingen tussen de verschillende onderwerpssystemen (respectievelijk RAMEAU, SWD/RSWK en LCSH) in het MACS-project ;
2. Meertalige toegang op basis van de UDC-tabellen, gedeeltelijk aangepast om de doelstellingen van het systeem te realiseren (ETHICS) ;
3. Ontwerpen van een ontologie door de mogelijkheid van de moderne talen om nieuwe en nog niet gekende concepten uit te drukken op basis van erkende taaluniversalia. Het

algemeen indexeringssysteem is één van de toepassingen van het Basic Semantic Reference Structure ontwikkeld voor EXPO 2000. Het interesseert ons ten zeerste, omdat het in staat is plaats te bieden voor een thesaurus format en om data uitwisseling te bevorderen ;.

4. Verschillende inspanningen worden ondernomen om nauwgezet goed uitgebouwde autoriteitsbestanden te ontwikkelen. Op die manier kunnen problemen bij de toegang tot onderwerpen in een omgeving met meerdere talen en/of letterschriften in landen zoals Zwitserland, Finland en Israël aangepakt worden ;
5. Automatische vertaling van de zoekvragen, gesteld in een taal die verschilt van de verschillende talen van de verzameling documenten, naar de verschillende talen van het TREC project.

In al de beschreven systemen valt telkens weer de vereiste van samenwerking en werkverdeling op. Deze projecten stoelen noodgedwongen op samenwerking. De authenticiteit en de culturele verschillen bemoeilijken het vertalen. De financiële ondersteuning is daarenboven ontzettend belangrijk in zulke ondernemingen. Op het einde van dit hoofdstuk pleiten wij ervoor zoveel mogelijk talen bij deze projecten te betrekken. Op die manier alleen is een gelijke toegang tot informatie mogelijk ongeacht de taal.

In hoofdstuk 5 wordt nogmaals onderstreept dat UDC perfect in staat is om als tussentaal op te treden. De troeven die UDC hiertoe heeft, werden in dit onderzoek ten volle gebruikt. UDC biedt perfect de mogelijkheid om over te schakelen van classificatienotaties naar thesaurustermen enerzijds en van de ene naar de andere taal anderzijds.

De twee meertalige thesauri die wij opgezet en gebouwd hebben, worden in hoofdstuk 5 ten tonele gevoerd. LTHES is gebaseerd op een sterk verkorte Roemeense uitgaven van de UDC (ongeveer 12.000 notaties), PTHES is gebaseerd op de pocketeditie van de UDC (ongeveer 4.000 notaties). Op basis van ons onderzoek naar het in overeenstemming brengen van de UDC-structuur en de thesaurusstructuur werd de haalbaarheid van meertalige thesauri die alle klassen van UDC omvat bewezen. Ons onderzoek wijst uit dat het aantal ingangstermen (descriptoren en non-descriptoren) de doelmatigheid ervan aantoont, niettegenstaande de reeds bestaande lijst van descriptoren en de lage graad van specificiteit.

Een kort overzicht van de mogelijkheden van UDC als een krachtig instrument om kennis te structureren geeft inzicht in haar relationele structuur in vergelijking met die van een thesaurus.

In datzelfde vijfde hoofdstuk behandelen wij ook de problemen van vertalingen en de mogelijke oplossingen. De lexicale en syntactische gelijkenis tussen het Roemeens en het Frans (beiden van dezelfde familie) wordt nogmaals voorgesteld. Vertaalbaarheid en verwoording zijn eenvoudiger binnen eenzelfde taalfamilie dan tussen talen van verschillende taalfamilies. Problemen i.v.m. homoniemen, polysemie en gelijkwaardigheid tussen termen van de verschillende talen komen aan bod. Voorbeelden van taxonomische voorstellingen en alfabetische indexen worden voor de verschillende talen gegeven. In twee bijlagen geven wij voorbeelden van de alfabetische en systematische (taxonomische) voorstellingen van zowel LTHES en PTHES.

De lage specificiteit van de descriptoren in een van bovenvermelde thesauri kan tot op zekere hoogte kritisch zijn.De classificatie van een database en de specificiteit van de op UDC gebaseerde thesaurus moeten een grotendeels gelijk niveau van specificiteit hebben om het systeem goed te kunnen evalueren. Als aan deze vereiste niet voldaan wordt, is er een probleem van compatibiliteit. Een direct gevolg is dat er ook verschillende niveaus van vangst en precisie optreden en tot op zekere hoogte is er informatieverlies.

Een belangrijke aanbeveling is dat in het zoeksysteem ruimte voorzien wordt waar de gebruikers suggesties kunnen doen voor nieuwe descriptoren of veranderingen aan bestaande descriptoren. Op die manier kan de thesaurus beter afgestemd worden op de behoeften van de gebruikers waardoor de kwaliteit ervan verbeterd wordt. Gebruikers zijn zowel indexeerders als zoekers. Het spreekt vanzelf dat het principe van één classificatienotatie voor één descriptor (of voorgeschreven combinatie van descriptoren) onaantastbaar is. De historiek van veranderingen kan in een annotatie opgeslagen worden.

Wij beschrijven de methodes waarop UDC-nummers geactualiseerd worden en waarop tekst en descriptoren toegevoegd worden aan bibliografische beschrijvingen. Hierdoor blijft onze database voortdurend aangepast. De methode waardoor het terugvinden van informatie mogelijk gemaakt wordt door het toevoegen van tekst en van UDC-notaties afgeleide descriptoren heeft zijn nut bewezen. In hoofdstuk 6 tonen wij aan hoe degelijk beide systemen op mekaar ingespeeld zijn. Onze doelstelling is aldus bereikt.

Daarenboven voegt de veeltaligheid een extra voordeel toe aan het zoekproces. Zodra de descriptoren veeltalig zijn en er een verbinding bestaat met de UDC-notaties, zijn deze descriptoren in de verschillende talen beschikbaar voor een groter aantal gebruikers.

Door een voorstelling van de verschillende stappen in onze gevalstudie geven wij een inzicht in onze methodologie :

1. Eerst moeten de fouten uit onze databank verwijderd worden. Hierbij gaat het voornamelijk over fouten betreffende het behandelde onderwerp. De meest voorkomende fouten zijn : typefouten, fouten in de UDC-notatie, fouten door een gebrekkige actualiseren van het E&C niveau in de MRF. Wij hebben de identificatiemethodes om een verkeerde UDC-notatie te detecteren beschreven en met voorbeelden geïllustreerd. (wat volgt is niet relevant in een samenvatting).
2. Eens de verkeerde notatie ontdekt moet deze verbeterd worden : de afzonderlijke, individuele verbetering is moeizaam en tijdrovend. (de details zijn niet relevant voor een samenvatting).
3. Globale verbeteringen zijn makkelijker te hanteren. Hiertoe werd een vlotte methode ontwikkeld.
4. Het zoeken in de databank kan gebeuren zowel via een specifieke vraag (het onderwerp is bvb. gekend) of via een algemene nog niet gespecificeerde benadering. In dit laatste geval wil de zoeker eerder browsen en informatie bijeenbrengen rond een algemeen thema.

   Grosso modo zijn er twee groepen zoekcodes (zie bijlage 1). De eerste groep betreft formele aspecten uit de bibliografische beschrijving, de tweede groep betreft de onderwerpen.

   Voorbeelden uit de eerste groep zijn : TI (de titel van het document), KW (een woord uit de bibliografische beschrijving, de titel inbegrepen). Voorbeelden uit de tweede groep zijn : DU (woorden uit de UDC teksten), DE (descriptoren die vooraf werden toegekend).

   Meertalige toegang kan op twee manieren gebeuren, afhankelijk van de gekozen thesaurustermen :

   - De zoekcodes voor automatisch toegekende descriptoren uit de meertalige thesaurus zijn afhankelijk van de taal van de zoekformule en de verwachte specificiteitsgraad. Deze zijn toegankelijk via de indexen van de databank;
   - Een andere mogelijkheid is het formuleren van enkelvoudige vragen of gecombineerde vragen volgens de regels van de Booleaanse logica, waarbij de Booleaanse operatoren OR (+), AND (*) en NOT (^) het zoekresultaat kunnen wijzigen. Via verschillende zoekvragen (elk onderdelen van een zoekstrategie) kan men informatie uit verschillende velden en subvelden van een record bijeenbrengen en wordt op die manier complex zoeken mogelijk. De geldigheid

van een zoekresultaat via een specifieke zoeksleutel kan getest worden door de zoekactie te herhalen met een andere, logisch gelijkwaardige zoeksleutel.

5. De volledige descriptorindex kan bekeken worden. De descriptoren uit de verschillende talen worden in een doorlopende rangschikking getoond.

In hoofdstuk zeven tonen wij aan in welke mate het terugzoeken van informatie beïnvloed wordt door de specificiteit van de gebruikte informatietaal. Hiertoe verwijzen wij nog naar de 'oude' *Rules for a Dictionary Catalogue* van Cutter (1876) waarin hij het belang van de specifieke zoeksleutel benadrukt.

Aan de hand van vele geteste zoekformules konden wij het effect aantonen van verschillende zoektalen en verschillende niveaus van specificiteit op het zoekproces binnen hetzelfde type van informatietalen. Hierbij betrokken wij belangrijke concepten als vangst, precisie en relevantie. Wij wilden aantonen dat hoe specifieker de informatietaal is, hoe precieser en relevanter het zoekresultaat is. Wij kwamen tot de volgende conclusies :

1. De uitgebreidheid van de collectie en de specificiteitsgraad van de informatietaal bepalen het resultaat van het zoekproces. Salton en McGill beklemtonen dat een correct evenwicht tussen deze beide factoren belangrijk is. Precisie is van belang voor uitgebreide verzamelingen. Het verband tussen de grootte van de collectie en de specificiteit van de informatietaal bepaalt de performantie van het informatiesysteem als geheel.

2. De door ons samengestelde en uitgeteste gecombineerde informatietaal is gebaseerd op een selectie van UDC-notaties die enigszins verschilt van die uit de MRF editie die aangewend werd in de classificatiecatalogus waaruit wij onze experimentele database samenstelden. Wij ontwikkelden twee meertalige thesauri met een verschillende specificiteitsgraad en testten hun performantie. Hieruit is gebleken dat er of een direct verband is tussen het specificiteitsniveau van voornoemde catalogus en dat van de UDC selectie waarop de thesaurus gebaseerd is of, de precisie van het zoekproces lager wordt. LTHES, de'brede' thesaurus is te algemeen voor de specifieke classificatiecatalogus. De brede descriptoren kunnen de relevante documenten niet onderscheiden van de niet-relevante. De hoge vangst brengt mee dat de tijdrovende selectie van echt relevante records daarbovenop nog moet gebeuren. Het zoeken in de tweede thesaurus (PTHES) gaf een hogere precisie en bijgevolg waren meer van de gevonden documenten relevant.

3. De procedure van automatisch indexeren die wij in ons onderzoek toepasten leidt tot een grotere precisie en relevanter zoekresultaten. Deze procedure is immers gebaseerd op een sterke onderzoektaal : rechtlijnig indexeren en controle van de indextermen als gevolg van de één-op-één overeenstemming tussen de UDC-notatie en de ervan afgeleide descriptor. Bij het manuele indexeren treedt vaak onbeslistheid of verwarring op. Ambiguïteit is in onze procedure uitgesloten. Zolang de toegekende UDC-notatie voorkomt zowel in de MRF als in de selectie van UDC-notaties waarop de thesaurus gebaseerd is, is automatisch indexeren haalbaar.De nauwkeurigheid van het terugzoeken van informatie is in hoge mate afhankelijk van het exact toekennen van classificatienotaties. De regels uitgebouwd om complexe UDC-notaties samen te stellen moeten gerespecteerd worden. Hierdoor wordt de kwaliteit van het automatisch indexeren verhoogd.

4. De resultaten uit de drie zoekmogelijkheden [PTHES, LTHES, de alternatieve 700-veld mogelijkheid (gebruikt als controlemogelijkheid, vermits het de betekenis van de UDC-notatie weergeeft)] zijn quasi gelijklopend. De meeste verschillen zijn het gevolg van verkeerd toegekende UDC-notaties, typefouten of lange cijferreeksen die moeilijk te beheren zijn.

Het is duidelijk dat een informatietaal met een hoge graad van specificiteit enkel voordelen biedt. Hoe hoger de specificiteit, hoe precieser en relevanter de zoekresultaten. Wanneer classificatienotaties aan de grondslag liggen van de indextermen (descriptoren) heeft dit interessante voordelen voor het rechtlijnig indexeren en de controle van de indextermen. De meertaligheid verbreedt daarenboven de zoekmogelijkheden.

Onze methode vermijdt een vervelend werk van herindexeren. Wanneer onze methode stapsgewijze toegepast wordt kunnen oude UDC-notaties gemakkelijk geactualiseerd worden. Voor zover de UDC-notaties voorkomen in de MRF, worden descriptoren onmiddellijk toegevoegd. Deze zijn dan beschikbaar voor het zoekproces. Het blijft uiteraard belangrijk dat de classificatienotaties correct en consequent toegekend worden. Classificatienotaties die correct volgens de regels van de UDC-grammatica opgebouwd zijn, vormen een voorwaarde voor een vlot verlopend zoekproces. Om hiertoe te komen zijn competentie en vaardigheid in de verschillende vakgebieden vereist. J.E. Mai had het volkomen bij het rechte eind toen hij beweerde : '*classification is an art that goes beyond any formal rules of logic and science'*.

Een zwak punt is dat de relaties tussen thesaurustermen niet getoond worden in de bibliografische database. Het 'ontdekken' van informatie is bovenop de traditionele 'information retrieval' heel belangrijk geworden. Een goed uitgebouwd relationeel netwerk van thesaurustermen helpt hierbij. Het is hierbij noodzakelijk in een informatietaal zoveel mogelijk synoniemen, quasi synoniemen en verwante termen aan te bieden. Navigeren (eerder dan browsen) is een functionaliteit die een breder zicht geeft op de in het systeem aanwezige relevante informatie. Suggesties voor woordassociaties en serendipiteit zijn succesvol in het 'ontdekken van informatie. Een helpfunctie in het systeem kan de gebruiker expliciet bijstaan in het zoekproces.

Onze benadering heeft nog een ander ontegensprekelijk voordeel : de in de database ingebouwde thesaurus is in zijn totale structuur voor de zoeker altijd beschikbaar. Daartegenover biedt de aanwezigheid van de UDC-MRF altijd de mogelijkheid enerzijds de één-op-één overeenstemming tussen beide te controleren en anderzijds de UDC-notaties te actualiseren.

Ten slotte zou het nuttig zijn mochten in ons project meer talen betrokken worden waardoor de taalbarrières verder zouden afnemen.

BIBLIOGRAPHY

1. AACR2R (1998). *Anglo-American Cataloguing Rules*. 2nd ed., 1998 revision. American Library Association
2. Achiri, Monica (1999). *Tezaur bazat pe CZU. Clasa 2 – Religie*. Constanta: Ex-Ponto
3. Adler, Elhanan (2000). *Multilingual and Multiscript Subject Access: the Case of Israel*. Available at: http://www.ifla.org/IV/ifla66/papers/035-130e.htm . Accessed May, 2001
4. Aitchison, J. & Gilchrist, A. (1987). Thesaurus Construction: a Practical Manual. 2nd ed. London: Aslib
5. Beghtol, Clare (1998). Knowledge Domains: Multidisciplinarity and Bibliographic Classification Systems. In: Knowledge Organization, 25 (1/2), pp. 1-12
6. Beghtol, Clare (2000). *A Whole, Its Kinds, and Its Parts*. 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 J. Williamson. Würtzburg, ERGON Verlag, pp. 313-319
7. BIE (1991). *Thésaurus de l'Éducation Unesco : liste par facettes de termes destinés à la recherche des documents et données relatifs à l'éducation, avec leurs équivalents anglais et espagnols* / préparé par le Bureau International d'Éducation. - 15e éd. rév. et augmentée. – Paris, Organisation des Nation Unies pour l'Éducation, la Science et la Culture, XII, 137p.
8. Blair, D.C. (1990). *Language and representation in information retrieval*. Amsterdam: Elsevier Science
9. BSI, (1999). *Universal Decimal Classification Pocket Edition, PD 1000*. London, British Standard Institution, 288 p.
10. Buchanan, B. (1979). *Theory of Library Classification*. London : Clive Bingley
11. Buxton, A. *Online applications*. In: McIlwaine, I. C. (1993). *Guide to the Use of UDC: an Introductory Guide to the Use and Applications of the Universal Decimal Classification*. The Hague, FID
12. Buyssens, M. E. (1943). *Les langages et le discours*. Bruxelles : Lebègue
13. Chan, Lois Mai (1994). *Cataloguing and Classification : an Introduction*. New York : McGraw-Hill
14. Chmielska-Gorczyca, Ewa (1997). *Polyhierarchy as a Means of Knowledge Representation*. In: McIlwaine (1997), pp. 105-112
15. Chomsky, Noam (1965). *Current Issues in Linguistic Theory*. The Hague, Mouton
16. Chomsky, Noam (1965a). *Aspects of the Theory of Syntax*. Cambridge, Mass.
17. Chomsky, Noam (1968). *Language and mind*. New York, Harcourt, Brace & World
18. Chomsky, Noam (1991). *Linguistics and Cognitive Science: Problems and Mysteries. The Chomskyan Turn*. Cambrigde, Mass., Blackwell
19. *Classification décimale universelle: ed. moyenne internationale*. 2e édition. Liège : Editions du CLPCF, 1990
20. Classification Research Group (1957). *The need for a faceted classification as the basis of all methods of information retrieval*. In: Proceedings of the International Study Conference on Classification for Information Retrieval. Dorking, 1957. London, Aslib, pp. 137-147

21. Clavel-Merrin, Geneviève & Lehtinen, Riitta (1995). *Multilingual and Multicharacter Set Data in Library Systems and Networks: Experiences and Perspectives from Switzerland and Finland*. Paper presented in a meeting of the Professional Group on Cataloguing at the 61st IFLA General Conference. Available at: http://www.ifla.org/IV/ifla61/papers/95.htm
22. Clavel-Merrin, Genevieve (1999). *The need for co-operation in creating and maintaining multilingual subject authority files*. Available at http://www.ifla.org/IV/ifla65/papers/080-155e.htm
23. CLEF (2000). *Workshop On Cross-Language Information Retrieval And Evaluation*. Available at: http://www.iei.pi.cnr.it/DELOS/CLEF/workshop00.html Accessed 20.11.00. Revised Papers are available in: Carol Peters (Ed.): Cross-Language Information Retrieval and Evaluation, Workshop of Cross-Language Evaluation Forum, CLEF 2000, Lisbon, Portugal, September 21-22, 2000. Springer Verlag, 2001
24. Cochrane, Pauline A. (1985). Redesign of Catalogs and Indexes for Improved Online Subject Access. Phoenix, Arizona, The Oryx Press
25. Cochrane, Pauline A. (1994). Elsinore revisited. In: *Knowledge Organisation and Quality Management*: proceedings of the 3rd International ISKO Conference, 20-24 June 1994, Copenhagen, Denmark. Frankfurt, Indeks Verlag, pp. 11-15
26. Collins Cobuild (1992). *Collins COBUILD English Language Dictionary*. London; Glasgow, Collins
27. *Compatibility and Integration of Order Systems* (1996). Research Seminar proceedings of the TIP/ISKO meeting. Warsaw, 13-15 September, 1995. Warsawa: Wydaw, 242 p.
28. Coseriu, E. & Geckler, H. (1974). *Linguistics and Semantics*. In: Sebeok, T.A. (ed.). Current Trends in Linguistics, vol. 12. The Hague: Mouton, 1974, pp. 103-171
29. Cruse, D. A. (1986). *Lexical semantics*. Cambridge, Cambridge University Press, XIV, 310p.
30. Cutter, Ch. A. (1904). *Rules for a Dictionary Catalogue*. Washington, DC, Government Printing Office. 4th ed.
31. Dahlberg, Ingetraut (1975). *The UDC as an Ideal Indexing Language*. In : Proceedings of the International Symposium : "The UDC in relation with other indexing languages", Herceg Novi, June 28-July 1, 1971. Beograd, Yugoslav Centre for Technical and Scientific Documentation, pp. 1-25
32. D'Haenens, L. & Lorphèvre, G. (1974). *Rapport sur la rédaction d'index thésaurifiés de la Classification Décimale Universelle*. Bruxelles
33. Dinu, Mihai. Personalitatea limbii române: fizionomia vocabularului. Bucuresti, Cartea Româneasca, 367 p.
34. Drăgoi, Elena (1999). *Tezaur bazat pe CZU : Clasa 1 – Filosofie*. Constanta: Ex-Ponto
35. Dumitrăşconiu, C. (1999). *Tezaur bazat pe CZU : Clasa 02 – Biblioteconomie*. Constanta: Ex-Ponto
36. Foskett, A. C. (1971). *Misogynists all: a Study of Critical Classification*. In: *Library Resources & Technical Services*, 15, 2, pp. 117-121
37. Foskett, A. C. (1982). *The Subject Approach to Information*. London, Clive Bingley, Linnet Books, Hamden. XVI, 574 p.
38. Francu, 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
39. Francu, Victoria (1997). *Language Barriers and Bridges: a Comparative Study on Three UDC Editions*. In: McIlwaine (1997), pp. 72-77

40. Frâncu, Victoria (1999a). *Tezaur bazat pe CZU: Clasa 8 - Lingvistică Literatură.* Constanta: Ex-Ponto, IX, 127p.

41. Frâncu, Victoria (1999b). *A Universal Classification System going through Changes*. In: Albrechtsen, H. and Mai, J.-E. (Eds.) *Proceedings of the 10th ASIS SIG/CR Classification Research Workshop,* October 31, 1999, held at the 62nd ASIS Annual Meeting, Washington, D.C., pp. 65-86

42. 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.

43. 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 María J. López-Huertas. Würtzburg: Ergon Verlag, pp. 404-411.

44. Freeman, Robert R. & Cochrane, Pauline Atherton (1968). *Audacious – an Experiment with an Online Interactive Reference Retrieval System Using the Universal Decimal Classification as the Index Language in the Field of Nuclear Science*. In: Cochrane (1985), p. 325-370

45. Frege, G. (1892).'Über Sinn und Bedeutung'. *Zeitschr. f. Philosophie und philosoph. Kritik*. English translation: 'On sense and reference'. In Geach, P. & Black, M. (eds.). Translations from the Philosophical Writings of Gottlob Frege. Oxford: Blackwell

46. Fugmann, Robert (1993*). Subject Analysis and Indexing: Theoretical Foundations and Practical Advice*. Frankfurt/Main, Indeks Verlag. XVI, 250 p.

47. Fugmann, Robert (1997). *Bridging the Gap Between Database Indexing and Book Indexing*. In: *Knowledge Organisation*, 24(1997), 4, pp. 205-212

48. Fugmann, Robert (1999). The Empirical Approach in the Evaluation of Information Systems. In: *Knowledge Organisation*, 26(1999), 1, pp. 3-9

49. Gilchrist, Alan (1972). *Intermediate Languages for Switching and Control*. In: *ASLIB Proceedings*, Vol. 24, No. 7, pp. 387-399

50. Gillman, Peter (1997). *Thesauri to aid retrieval from very large text bases: subject term retrieval from large text resources, and the problem of ambiguity*. In: McIlwaine (1997), pp. 113-119

51. Glushkov et al. (1978). *The evaluation of the degree of compatibility of information retrieval languages in the information retrieval systems of documents*. In: Nauchno-Technicheskaya Informatsya, Series 2(1), pp. 14-19

52. Goossens, Paula. *Language Barriers in the Exchange of Bibliographic Records: Analysis and Solution*. In: Plassard, (1993), pp. 45-46

53. Greenberg, Joseph (ed.) (1963). *Universals of Language*. Cambridge, Mass. M.I.T. Press

54. Grünewald, Franz (1994). *Anwendung UDK in ETHICS*. In: G. Riesthuis et al. *Bang voor onderwerpsontsluiting*. Antwerpen, VVBAD, pp.46-52

55. Hanga-Calciu, Rodica. *Homonymy in English: a synchronic and diachronic approach with special reference to homophone-heterographs*. Bucharest, University of Bucharest

56. Hoppe, Stephan (1996). *The UMLS: A Model for Knowledge Integration in a Subject Field*. In: *Compatibility and Integration of Order Systems* (1996). Research Seminar proceedings of the TIP/ISKO meeting. Warsaw, 13-15 September 1995. Warsawa: Wydaw, pp. 97-110

57. Hornby, A. S. (1989). *Advanced Learner's Dictionary of Current English*. 4th ed. Oxford, Oxford University Press

58. Hudon, Michèle (1997). *Multilingual thesaurus construction: integrating the views of different cultures in one gateway to knowledge and concepts*. In: Knowledge Organization 24 (2), pp.84-91
59. Hunter, Eric J. (1994). *The Missing Link: the Role of Classification in OPAC's*. In: Riesthuis et al. *Bang voor onderwerpsontsluiting*. Antwerpen: VVBAD, pp. 23-45
60. Hutchins, W. J. (1975). *Languages of indexing and classification: a linguistic study of structure and functions*. Stevenage, Peter Peregrinus, 148 p.
61. IFLA (1998). http://ifla.inist.fr/VII/s29/pubs/ci18.htm
62. Iivonen, M. (1996). *Selection of search terms as a meeting place of different discourses*. 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. 224-230
63. International Organisation for Standardisation (1985). ISO 5964: *Guidelines for the establishment and development of multilingual thesauri*. Geneva: ISO
64. International Organisation for Standardisation (1986). ISO 2788: *Guidelines for the establishment and development of monolingual thesauri*. 2nd ed. Geneva: ISO
65. Jacob, Elin K. & Priss, Uta (1999). Application of Faceted Classification Structures in Electronic Knowledge Resources. In: Albrechtsen, H. and Mai, J.-E. (Eds.) *Proceedings of the 10th ASIS SIG/CR Classification Research Workshop,* October 31, 1999, held at the 62nd ASIS Annual Meeting, Washington, DC.
66. Jacob, Elin K. & Albrechtsen, Hanne (1997). *Constructing Reality: the Role of Dialogue in the Development of Clasificatory Structures.* In: McIlwaine (1997), pp. 42-50
67. Jackobson, Roman (1963). Implications of language universals for linguistics. In: Greenberg (1963)
68. Kunz, Martin (2002). *Subject retrieval in distributed resources: a short review of recent developments*. Available at: http://www.ifla.org/IV/ifla68/papers/007-122e.pdf accessed October 17, 2002
69. Lancaster, F. W. (1986). *Vocabulary control for information retrieval.* 2nd ed. Arlington, VA: Information Resources Press, 270 p.
70. Lancaster, F. W. (1998). *Indexing and abstracting in theory and practice*. 2nd ed. London, Library Association Publishing, XVI, 412 p.
71. Landry, Patrice (2000). *The MACS Project: Multilingual Access to Subjects (LCSH, RAMEAU, SWD)*. Available at: http://www.ifla.org/IV/ifla66/papers/165-181e.pdf
72. Lansing, J. (1995). Genus, Species, and Vantages. In: *Language and the Cognitive Construal of the World.* Ed. by J. R. Taylor and R. E. MacLaury. NY: Mouton de Gruyter, pp. 365-375
73. Lazinger, S & Adler, E. (1998). *Cataloguing Hebrew Materials in the Online Environment: a Comparative Study of American and Israeli Approaches*. Ed. by S. S. Intner. Englewood, Colorado: Libraries Unlimited
74. Lloyd, G. A. (1972). *The Universal Decimal Classification as an International Switching Language*. In: Subject retrieval in the seventies: proceedings of an international symposium. University of Maryland, May, 14-15, 1971, Westport, Greenwood, pp. 116-123
75. López Carreño, R et al. (1999). *El tesauro como herramienta en la optimación de la gestión de la documentación administrativa*. In: La Representación y la Organización del Conocimiento in sus distintas perspectivas: su influencia en la Recuperación de la Información : actes del IV Congreso ISKO-España EOCONSID '99, Granada. Ed. por María José López-Huertas, Juan Carlos Fernandes Molina. Granada, pp. 109-116
76. Lyons, J. (1968). *Introduction to Theoretical Linguistics*. Cambridge, Cambridge University Press

77. Lyons, J. (ed.) (1970). *New horizons in linguistics*. Harmondsworth, Penguin Books, 367 p.
78. Lyons, J. (1977). *Semantics*. Cambridge. Cambridge University Press, 2 vol.
79. Lyons, J. (1995). *Linguistic semantics*. Cambridge, Cambridge University Press
80. Mai, Jens-Erik (2000). *Likeness: a Pragmatic Approach*. 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, pp. 23-27
81. Maneca, Constant (1966). *Consideratii asupra frecventei cuvintelor în limba românа literara contemporana*. In: Studii si cercetari Lingvistice, XVII, 6, pp.623-633
82. Maniez, Jacques (1997). *Database Merging and the Compatibility of Indexing Languages*. In: *Knowledge Organization*, 24(1997), 4, pp. 213-224
83. Marcella, R. & Newton, R. (1994). *A new manual of classification*. Aldershot, Brookfield: Gower. XII, 287 p.
84. Marcu, Florin, Maneca, Constant (1978). *Dictionar de neologisme*. Bucuresti, Editura Academiei
85. McCrum, Robert et al. (1987). *The Story of English*. Penguin Books
86. McIlwaine, I. C. & Buxton, A. (1993). *Guide to the Use of UDC: an Introductory Guide to the Use and Applications of the Universal Decimal Classification*. The Hague, FID
87. McIlwaine, I. C. & Williamson, N. J. (1995). *Future revision of UDC: progress report on a feasibility study for restructuring*. In: Extensions and Corrections to the UDC, Vol. 17, pp. 11-17
88. McIlwaine, I. C. & Williamson, N. J. (1997). *Class 61 – Medicine: Restructuring Continued*. In: Extensions and Corrections to the UDC, Vol. 19, pp. 30-40
89. McIlwaine, I. C. (ed.) (1997). *Knowledge Organisation for Information Retrieval*: Proceedings of the Sixth International Study Conference on Classification Research, held at University College London, 16-18 June 1997. The Hague, FID, IX, 206 p.
90. McIlwaine, I. C. & Williamson, N. J. (1999). *International Trends in Subject Analysis Research*. In: Knowledge Organization, 26 (1), pp. 23-29
91. McIlwaine, I. C. (2000). *The Universal Decimal Classification: a guide to its use*. The Hague, UDC Consortium
92. MDA Archaeological Objects Thesaurus (1997). Available at: http://www.mds.org.uk/archobj/archcon.htm. Accessed September 2002.
93. Möller, G. et al. (1999). *Automatic classification of the World Wide Web using the Universal Decimal Classification*: Proceedings of the 23rd International Online Information Meeting, London, 7-9 December 1999. Oxford, Learned Information Europe, pp. 231-237
94. Morris, C. W. (1971). *Writings on the General Theory of signs*. The Hague, Mouton
95. National Information Standards Organization (1994). *Guidelines for the Construction, Format, and Management of Monolingual Thesauri*. ANSI/NISO Z39.19-1994. New York: National Information Standards Organization
96. Olson, Hope (1997). *Feminist Locales in Dewey's Landscape: Mapping a Marginalized Knowledge Domain*. In: McIlwaine (1997), pp.129-133
97. Panman, Otto (1982). *Homonymy and Polysemy*. North Holland Publishing Company
98. Pei, Mario (1976). *The Story of Latin and the Romance Languages*. London
99. Perreault, J. M. (1969). *Towards a Theory for UDC*. London, Bingley
100. Plassard, Marie France & McLean Brooking, Diana (eds.) (1993). *UNIMARC/CCF*: Proceedings of the Workshop held in Florence, 5-7 June 1991. München [etc.], K. G. Saur
101. Popescu, Liliana (1999). *Tezaur bazat pe CZU: Clasa 57/59*. Constanta: Ex-Ponto

102. Pottier, B. (1963). *Recherches sur l'analyse sémantique en linguistique et en traduction mécanique*. Nancy

103. Quantz, Joachim J. (1995). *Preferential Disambiguation in Natural Language Processing*. Berlin

104. Quirk, R. et al. (1985). *A Comprehensive Grammar of English*. London, Longman

105. Richardson, E. C. (1935). *Classification*. New York: H. W. Wilson

106. Riesthuis, G.J.A. & Bliedung, S. (1990). Thesaurification of UDC: preliminary report. In: Gilchrist, A. and Strachan D. (Eds.). *The UDC: essays for a new decade*. London: Aslib, pp. 85-97

107. Riesthuis, G.J.A. (1996). Theory of Compatibility of Information Languages. In: *Compatibility and Integration of Order Systems* (1996). Research Seminar proceedings of the TIP/ISKO meeting. Warsaw, 13-15 September 1995. Warsawa: Wydaw, pp. 23-31

108. Riesthuis, G.J.A. (1997). *Decomposition of complex UDC notations*. In: McIlwaine (1997), pp. 139-143

109. Riesthuis, G.J.A. (1998). *Zoeken met woorden: hergebruik van onderwerpsontsluiting*. Amsterdam: Universiteit van Amsterdam. Leerstoelgroep Boek-, Archief- en Informatiewetenschap. VIII, 186 p.

110. 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

111. Rowley, Jenifer E. (1998). *Organising knowledge*. Aldershot, Gower. XVIII, 454 p.

112. Russell, Bertrand (1948). *Human Knowledge: its Scope and Limits*. New York, Simon & Schuster

113. Salton, Gerard & McGill, Michael J. (1983). *Introduction to Modern Information Retrieval*. Auckland [etc.], McGraw-Hill. XV, 448 p.

114. Santoro, Michele (1995). Per una riflessione sulla istoria, il ruolo e le prospettive della Classificazione Decimale Universale. In: Biblioteche oggi, 13 (8), p. 48-57, also available at: http://www.aib.it/aib/contr/santoro1.htm

115. Santoro, Michele (1996). *International exchange languages: the case of the Universal Decimal Classification*. In: Extensions and Corrections to the UDC, Vol. 18, pp. 81-91

116. Saussure, Ferdinand de (1964). *Cours de linguistique générale*. Paris, Payot

117. Schäuble, Peter & Sheridan, Paraic (1998). *Cross-Language Information Retrieval (CLIR) Track Overview.* In: Voorhees, E. M. & Harman, D. K. (eds.) (1998). *Information Technology: the Sixth Text REtrieval Conference (TREC-6)*. Washington: National Institute of Science and Technology (NIST)

118. Schmitz-Esser, Winfried (1996). *Language of General Communication and Concept Compatibility*. In: *Compatibility and Integration of Order Systems*: Research Seminar Proceedings of the TIP/ISKO Meeting, Warsaw, 13-15 September 1995. Warsawa, Wydaw, pp. 11-22

119. Schmitz-Esser, Winfried (1998). *Defining the Conceptual Space for a World Exhibition – First Experiences.* In: W. Mustafa el Hadi, J. Maniez and S. Pollitt (eds.). *Structure and Relations in Knowledge Organization*: Proceedings of the Fifth International ISKO Conference, 25-28 August 1998, Lille, France. Würtzburg, ERGON Verlag, pp. 146-152

120. Schmitz-Esser, Winfried (1999). *Thesaurus and Beyond: an Advanced Formula for Linguistic Engineering and Information Retrieval*. In: Knowledge Organization, 26 (1), pp. 10-22

121. Scibor, E. (1997). *UDC as a tool for information retrieval: general considerations*. In: McIlwaine (1997), pp. 200-204

122. Scott, M. L. (1993). *Conversion Tables: LC-Dewey, Dewey-LC*. Englewood: Libraries Unlimited, 365p.

123. Seymour, Chaim (2000). *A Time to Build – Israeli Cataloging in Transition*. Available at: http://www.ifla.org/IV/ifla66/papers/011-164e.htm . Accessed May, 2001

124. Soergel, Dagobert (1985). *Organising Information: Principles of Database and Retrieval Systems*. Orlando, Fl., Academic Press

125. Stati, Sorin (1979). *La semantique des adjectives en langues romanes*. Paris, Jean-Favard

126. Suteu, Valeriu (1959). *Observatii asupra frecventei cuvintelor în operele unor scriitori români*. In: *Studii si Cercetari Lingvistice*, X, 3, pp. 419-445

127. Svenonius, Elaine (1997). *Definitional Approaches in the Design of Classification and Thesauri and their Implications for Retrieval and for Automatic Classification*. In: McIlwaine (1997), pp. 11-16

128. TREC-7 (1998). [http://trec.nist.gov/pubs/trec7/t7_proceedings.html]

129. TREC-8 (1999). [http://trec.nist.gov/pubs/trec8/t8_proceedings.html]

130. Tzigara-Samurcas, Alexandru (1933). *Fundatiunea Universitara "Carol I"*. Bucuresti: Socec, 168 p.

131. UDC (1985). *Universal Decimal Classification International Medium Edition, English text.* London, British Standard Institution

132. Ullmann, Stephen (1952). *Précis de sémantique française*. Paris, Presses Universitaires de France, 334 p.

133. UMLS (1998). Available at: http://www.nlm.gov Accessed 10.09.98

134. Vickery, Brian (1961). *La Classification Décimale Universelle et l'indexage de la documentation technique*. In *: Bulletin de l'Unesco à l'intention des bibliothèques*, 15 (2), pp. 57-75

135. Vickery, Brian (1971). *Document Description and Representation*. In: *Annual Review of Information Science and Technology*, 6, pp. 113-140

136. Vickery, Brian (1997). *Issues in Knowledge Representation*. In: McIlwaine (1997), pp. 180-182

137. Voorhees, E. M. & Harman, D. K. (eds.) (1998). *Information Technology: the Sixth Text REtrieval Conference (TREC-6)*. Washington: National Institute of Science and Technology (NIST)

138. Webster's (1989). *Webster's Dictionary of English Usage*. Springfield, Mass., Merriam-Webster, Inc., 978 p.

139. Weinreich, (1963). *On the Semantic Structure of Language*. In: Greenberg (1963)

140. Whaley, Lindsay J (1997). *Introduction to Typology: the Unity and Diversity of Language*. Thousand Oaks, Sage, XVI, 323 p.

141. Wittgenstein, Ludwig (1958). *Philosophical investigations*. New York, Macmillan Publishing

142. Yancey, Trish & Clarke, Dave (1999). *Lexicography without limits – a Web-Based Solution*. Available at: www.synaptica.com

143. Yee, Martha M. & Layne, Sara Shatford (1998). *Improving Online Public Access Catalogues*. Chicago, London, ALA, 220 p.

APPENDIX 1
USER INSTRUCTIONS

With a view to improve the user-friendliness of our system some user instructions are mentioned below. In the first place the user needs to get familiar with the search codes used and their correspondence with the database fields.

For this purpose and in order to have a more clear image of the database we clustered the fields of the experimental database in three major sections: the first includes *the bibliographic data and the subject representation data as they were found initially in the classified catalogue,* the second contains several categories of *fields resulting from various automatic subject analysis procedures* and the third, *fields in the UDC Master Reference File (MRF)* as follows:

I. Fields in the bibliographic part of the experimental database (as they initially existed)

1. Fields in the descriptive part of the record

They have little relevance at this point in the development of our study although their importance in information retrieval has already been demonstrated (see the 'Bank' example in Chapter 6). These are:

- field 200 for title
- field 210 for author
- field 220 for publisher
- field 230 for number of pages
- field 240 for series title
- field 250 for standard number
- field 300 for annotations

They are all grouped under the tag TITEL in the display format of the experimental database. For information retrieval one can use as search codes 'TI=' for the title (or words of the title) of the document and 'KW=' for words in the whole body of the bibliographic description to mention the most important and easy-to-use ones.

2. Fields in the subject analysis part of the bibliographic record

They can hold textual or alphabetical data and then we deal with Romanian descriptors assigned to bibliographic records by manual indexing such as:

- field 610 for topical subject headings
- field 611 for geographical subject headings

Both these fields are labelled with the tag BDES in the display format of the experimental database and in order to retrieve them the prefix used as search code is 'DE=' followed by the Romanian subject heading.

Along with them and in the first and most important place there is a field that holds numerical and alphanumerical data in the classified catalogue that our whole study is based on:

- field 675 for the UDC notations

This is tagged UDC in the database and can be accessed by using the field number as a search code i.e. '675=' followed by the UDC notation.

II. Fields holding textual data resulted from automatic subject analysis procedures

In the same manner as the first category of fields, these fields are divided into 2 main groups:

1. fields that hold UDC text corresponding to the component parts of the UDC notations in the classified catalogue:

- field 701for auxiliaries of language, tagged as LANG in the database;
- field 702 for auxiliaries of form, tagged as FORM;
- field 703 for auxiliaries of place, tagged as PLAC;
- field 704 for auxiliaries of ethnic groups, tagged as ETHN;
- field 705 for auxiliaries of time, tagged as TIME;
- field 706 for alphabetical addition, tagged as TEXT;
- field 707 for auxiliaries of point of view tagged VIEW;
- field 708 for auxiliaries of materials and persons, tagged as CHAR;
- field 709 for UDC main numbers, tagged as MAIN.

All the above fields can be accessed by the same search code 'DU=' followed by the description (caption) of the kind of UDC number they represent. This is enabling one of the most comprehensive search methods.

2. fields that hold the two kinds of multilingual descriptors – from LTHES and PTHES – plus their related non-descriptors such as:

- field 711 for PTHES descriptors, tagged in the database as PDES
- field 712 for PTHES non-descriptors, tagged as PNDES
- field 713 for LTHES descriptors, tagged in the database as LDES
- field 712 for LTHES non-descriptors, tagged as LNDES

These fields can be accessed as follows:

- the PTHES multilingual descriptors by search codes PMBE=/PMBF=/PMBR= according to the language needed
- the PTHES non-descriptors by search codes PNBE=/PNBF=/PNBR=
- the LTHES multilingual descriptors by search codes LMBE=/LMBF=/LMBR=
- the LTHES non-descriptors by search codes LNBE=/LNBF=/LNBR=

In addition to these there is a set of search codes that are capable to search for truncated numbers. In case the UDC numbers given as classification notations are not found in exactly that particular form among the selection of UDC numbers used in building the two multilingual thesauri. Therefore they are truncated till the notation and the appropriate thesaurus term are found. These search codes are:

- PMDE=/PMDF=/PMDR= and PNDE=/PNDF=/PNDR= for PTHES terms
- LMDE=/LMDF=/LMDR= and LNDE=/LNDF=/LNDR= for LTHES terms

It is recommendable that both sets of search codes are used in information retrieval procedures for an all-inclusive search result.

III. Fields in the UDC Master Reference File (MRF)

The fields listed below are found in the MRF part of the experimental database. Out of these, the most frequently used are those holding *information on the UDC number* and its description (fields 1 and 2) and those holding *descriptors and non-descriptors* related to that particular UDC number (160, 165 and 170,175 respectively). Here they are:

- field 001 for UDC number
- field 002 for table
- field 100 for description (caption)
- field 105 for verbal examples
- field 160 for PTHES descriptor
- field 165 for PTHES non-descriptors
- field 170 for LTHES descriptor
- field 175 for LTHES non-descriptors

The UDC MRF numbers can be accessed by search code 'NC=' followed by the number. The retrieved record contains all the information connected with that number.

A special mention can be made at this point about the *context provider* i.e. about the 'short MRF' as it was called. This information is placed in a separate section of the experimental database and it is used to provide context to the UDC notations in the bibliographic records (*subfield ^x* in the 70- fields). The 175 records in this section have a limited number of fields and they fix the domain of the subject denoted by the UDC number:

- field 001 for UDC number
- field 002 for table
- field 100 for description

The short MRF can be accessed by using 'NK=' followed by the UDC class number as a search code. It is not likely that this part of the database is frequently used in searching but its utility for the context of the subject is highly appreciated particularly for concepts that occur in more disciplines (see subfield ^x of any 70- field in all the bibliographic records given as example in the foregoing).

An outline of all these fields, tags and search codes is given in the following table:

BCUB database	Field no.	Field description/tag	Search codes	Meaning of the code
MRF part of the experimental database	001	UDC number	NC=	UDC MRF number
			NK=	UDC number in the short MRF or the context provider
	002	UDC table		
	100	UDC description (caption)	UE=	UDC caption
	105	Verbal examples		
	160	PTHES descriptor	DPE/F/R=	PTHES multilingual descriptors in 3 languages
	165	PTHES non-descriptors	NPE/F/R=	PTHES non-descriptors in 3 languages
	170	LTHES descriptor	DLE/F/R=	LTHES multilingual descriptors in 3 languages
	175	LTHES non-descriptors	NLE/F/R=	LTHES non-descriptors in 3 languages
Bibliographic part of the experimental database	200 200/ 300	title of the work/TITEL bibliographic description in the BCUB database/TITEL	TI= TW= KW=	title title words word from the whole body of the bibliographic record (except for field 300)
	610 611	Subject headings/BDES Geographical headings/BDES	DE=	descriptors manually assigned in the indexing process
	675	UDC notations in the record/UDC	675=	UDC notation in the classified catalogue
	701	auxiliaries of language/LANG	DU=	- textual counterpart of the kind of UDC number they represent; - the text corresponds to the component part of the decomposed UDC notation
	702	auxiliaries of form/FORM		
	703	auxiliaries of place/PLAC		
	704	auxiliaries of ethnic groups/ETHN		
	705	auxiliaries of time/TIME		
	706	alphabetical additions/TEXT		
	707	auxiliaries of point of view/VIEW		
	708	auxiliaries of materials/pers/CHAR		
	709	UDC main numbers/MAIN		
	711	PTHES descriptors/PDES	PMBE/F/R= PMDE/F/R=	PTHES multilingual descriptors in 3 languages
	712	PTHES non-descriptors/PNDES	PNBE/F/R= PNDE/F/R=	PTHES non-descriptors in 3 languages
	713	LTHES descriptors/LDES	LMBE/F/R= LMDE/F/R=	LTHES multilingual descriptors in 3 languages
	714	LTHES non-descriptors/LNDES	LNBE/F/R= LMDE/P/R=	LTHES non-descriptors in 3 languages

APPENDIX 2
SAMPLE OF THE MULTILINGUAL THESAURUS BASED ON AN ABRIDGED EDITION OF THE UDC (LTHES) WITH ENTRY TERMS IN ENGLISH

FIELD CROPS
F: Plantes de culture
R: Plante de câmp
UDC: 633
UF : Aromatic plants
Beverage plants
Cereals
Condiment plants
Edible roots and tubers
Forage grasses
Industrial plants
Leguminosae
Medicinal plants
Oleaginous plants
Plants yielding stimulants
Sugar plants
Tanning plants
Textile plants
BT : Agriculture

Film
use: CINEMA

Film presentations
use: PERFORMANCES

FILM PRODUCTION
F: Production du film
R: Producţia de filme
UDC: 791.44
BT : Public entertainments

Film studio buildings
use: STUDIO BUILDINGS

FINANCE
F: Finances
R: Finanţe
UDC: 336
UF : Financial policy
BT : Economics
NT : Banking
Public expenditure
Public finance
Public revenue

Financial law
use: COMMERCIAL LAW

Financial legislation
use: ACTIVITIES OF PUBLIC ADMINISTRATION

Financial need
use: SOCIAL PROBLEMS REQUIRING ASSISTANCE

Financial plans
use: BUDGETS

Financial policy
use: FINANCE

FINISH
F: Finois
R: Finlandeză
UDC: =511.111
BT : Uralic languages

FINISH LANGUAGE
F: Langue finnoise
R: Limba finlandeză
UDC: 811.511.111
BT : Ural-Altaic languages

FINISH LITERATURE
F: Littérature finnoise
R: Literatură finlandeză
UDC: 821.511.111
BT : Ural-Altaic literatures

FINISHING AND DECORATING TRADES
F: Finition et décoration
R: Finisare şi decorare
UDC: 698
BT : Building trade

Finland
use: SCANDINAVIAN STATES

FINNS
F: Finlandais
R: Finlandezi
UDC: (=511.111)

FIRE HAZARDS
F: Risques d'incendie
R: Incendii
UDC: 614.84
UF : Fire prevention
Fires
BT : Accidents

Fire prevention
use: FIRE HAZARDS

Firearms
use: HAND WEAPONS

Fires
use: FIRE HAZARDS

Fireworks
use: EXPLOSIVES

Firms
use: BUSINESS HOUSES

Fish breeding
use: PISCICULTURE

Fishes
use: VERTEBRATA

FISHING
F: Pêche
R: Pescuit
UDC: 639.2
RT : Fishing and hunting
Pisciculture

FISHING AND HUNTING
F: Pêche et chasse sportive
R: Pescuit şi vânătoare sportive
UDC: 799
BT : Sport
RT : Fishing
Hunting

Fixing devices
use: FASTENING DEVICES

FLAGS
F: Drapeaux
R: Drapele
UDC: 929.9
UF : Banners

FLEXIBLE TRANSMISSIONS
F: Transmissions flexibles
R: Transmisii flexibile
UDC: 621.85
UF : Chain drives
BT : Motive power engineering

Floods
use: NATURAL INLAND WATERS

Floors
use: STRUCTURAL PARTS OF BUILDINGS
Flora
use: GEOGRAPHIC BOTANY

Floral arts
use: DECORATIVE ARTS

Flour and corn milling
use: PROCESSING OF CEREAL GRAINS

Flour confectionery
use: PROCESSING OF CEREAL GRAINS

FLUID MECHANICS
F: Mécanique des fluides
R: Mecanica fluidelor
UDC: 532
UF : Hydromechanics
BT : Physics
NT : Hydrodynamics

FLUIDS HANDLING TECHNIQUES
F: Technique de manipulation des fluides
R: Instalaţii pentru transportul fluidelor
UDC: 621.6
BT : Mechanical engineering
NT : Blowers
Devices for transporting of liquids
Fans
Fluids storage and distribution installations
Pumps and pumping

FLUIDS STORAGE AND DISTRIBUTION INSTALLATIONS
F: Installation de stockage et de distribution des fluides
R: Instalaţii de depozitare şi distribuţie a fluidelor
UDC: 621.64
UF : Conduits
Pipes
BT : Fluids handling techniques

Fluorine
use: PRODUCTION OF HALOGENS

Folk literature
use: FOLKLORE

FOLKLORE
F: Folklore
R: Folclor
UDC: 398
UF : Folk literature
Popular beliefs
Popular traditions
Traditional songs
RT : Ethnography

FOOD
F: Aliments
R: Alimente
UDC: 641/642
UF : Cooking
Cutlery
Meals
Preparation of foodstuffs
Preservation of foodstuffs
Table decoration
Tableware
BT : Home economics

Football
use: BALL GAMES

FOOTWEAR
F: Chaussures
R: Incălţăminte
UDC: 685.3
BT : Leatherware

Forage grasses
use: FIELD CROPS

FOREIGN LANGUAGE
F: Langue etrangère
R: Limbă străină
UDC: 81'243
BT : Practical knowledge of languages

FOREIGN POLICY
F: Politique extérieure
R: Politică externă
UDC: 327
UF : Diplomatic relations
International affairs
International blocs
International relations
World politics
BT : Politics

FOREIGN TRADE
F: Commerce extérieur
R: Comerţ exterior
UDC: 339.5
UF : Export
Import
BT : Trade
RT : International economy
International finance
Market analysis

Forest engineering
use: FORESTRY

FORESTRY
F: Sylviculture
R: Silvicultură
UDC: 630
UF : Forest engineering
RT : Agriculture

Forge equipment
use: FORGE WORK

FORGE WORK
F: Travail de la forge
R: Forjă
UDC: 621.73
UF : Forge equipment
BT : Mechanical technology

Forged ironware
use: ARTICLES OF IRON AND STEEL

FORGERY
F: Falsification
R: Falsuri în artă
UDC: 7.061
UF : Counterfeiting
Plagiarism

Forged steelware
use: ARTICLES OF IRON AND STEEL

FORMS OF ENTERPRISE
F: Formes d'entreprises
R: Intreprinderi
UDC: 658.1/.2
BT : Business management
RT : Forms of organization in the economy

FORMS OF ORGANIZATION IN THE ECONOMY
F: Formes d'organisation dans l'activité économique
R: Forme de organizare economică
UDC: 334
UF : Cartels
Economic cooperation
Economic organization
Multinational enterprises
Trusts
BT : Economics
RT : Forms of enterprise

Forms of political organization
use: STATE

FORTIFICATIONS
F: Fortifications
R: Fortificaţii
UDC: 623.1/.3
UF : Defence works
Fortresses
Land minefields
BT : Military engineering

Fortresses
use: FORTIFICATIONS

Fossils
use: PALAEONTOLOGY

FOUNDATION ENGINEERING
F: Ingénierie des fondations
R: Fundaţii
UDC: 624.15
UF : Foundations in water
Pile foundations
Substructure work
BT : Infrastructures

Foundations in water
use: FOUNDATION ENGINEERING

FOUNDATIONS OF MATHEMATICS
F: Fondements des mathématiques
R: Bazele matematicii
UDC: 510
BT : Mathematics
NT : Algorithms
Mathematical logic
Set theory

Foundry equipment
use: FOUNDRY WORK

FOUNDRY WORK
F: Travail de fonderie
R: Turnătorie
UDC: 621.74
UF : Foundry equipment
BT : Mechanical technology

FRANCE
F: France
R: Franţa
UDC: (44)

Freight cars
use: TRAIN CARS

FRENCH
F: Français
R: Franceză
UDC: =133.1

FRENCH LANGUAGE
F: Langue française
R: Limba franceză
UDC: 811.133.1
BT : Romance languages

FRENCH LITERATURE
F: Littérature française
R: Literatură franceză
UDC: 821.133.1
BT : Romance literatures

FRENCH-SPEAKING PEOPLES
F: Peuple français
R: Francezi
UDC: (=133.1)
BT : Roman peoples

FREQUENCY
F: Fréquence
R: Frecvenţă
UDC: 621.3.029
BT : Electrical characteristics

FREQUENCY MULTIPLIERS AND DIVIDERS
F: Multiplicateurs et diviseurs de fréquence
R: Multiplicatoare şi divizoare de frecvenţă
UDC: 621.374
BT : Technique of electromagnetic waves

Friction
use: INTERMOLECULAR FORCES

Friendship
use: ETHICS AND SOCIETY

FRUIT GROWING
F: Cultures fruitières
R: Cultura fructelor
UDC: 634.2/.7
UF : Berries
Citrous fruits
Nuts
Stone fruits
Tropical and subtropical fruits
BT : Agriculture

Fruit juices
use: SOFT DRINKS

FRUIT WINES
F: Vins de fruits
R: Vinuri din fructe
UDC: 663.3
UF : Cider
Perry
BT : Industrial microbiology
RT : Wine

Fuel economy
use: FUELS

FUEL TECHNOLOGY
F: Technologie des combustibles
R: Tehnologia combustibililor
UDC: 662.7
UF : Processed fuels
BT : Explosives and fuels
NT : Fuel technology of wood

Fuel technology of hard coal
Gaseous fuels
Liquid fuels
Technology of low-grade fuels

FUEL TECHNOLOGY OF WOOD
F: Combustibles derivés du bois
R: Tehnologia combustibililor lemnoşi
UDC: 662.71
BT : Fuel technology
RT : Wood as fuel

FUEL TEHNOLOGY OF HARD COAL
F: Charbon dur bitumineux
R: Prelucrarea huilei
UDC: 662.74
BT : Fuel technology
RT : Hard coal as fuel

FUELS
F: Combustibles
R: Combustibili
UDC: 662.6/.9
UF : Fuel economy
Natural fuels
BT : Explosives and fuels
NT : Combustion
Hard coal as fuel
Low-grade fuels
Wood as fuel

Function of libraries
use: DEVELOPMENT OF LIBRARIES

Functional equations
use: DIFFERENTIAL EQUATIONS

Fundamental rights
use: HUMAN RIGHTS

FUNERARY ARCHITECTURE
F: Architecture funéraire
R: Arhitectură funerară
UDC: 726.8
UF : Sepulchral monuments
BT : Religious architecture

Fur and imitation leather
use: LEATHER INDUSTRY

FUR PRODUCTS
F: Produits des fourrures
R: Blănărie
UDC: 675.6
BT : Leather industry

FURNACE ENGINEERING
F: Ingénierie des chaudières
R: Tehnica încălzitului industrial
UDC: 662.9
UF : Combustion engineering
Smoke consumption
BT : Explosives and fuels
NT : Heat recovery

FURNITURE FINISHING
F: Finition de meuble
R: Finisarea mobilei
UDC: 684.6
UF : Inlay
Marquetry
Veneering
BT : Furniture industry

FURNITURE INDUSTRY
F: Industrie du meuble
R: Industria mobilei
UDC: 684
UF : Furniture manufacture
BT : Industries and trades
NT : Furniture finishing
Upholstery

Furniture manufacture
use: FURNITURE INDUSTRY

Further education
use: EDUCATION AND TRAINING OUT OF SCHOOL

GAMES OF CHANCE
F: Jeux de hasard
R: Jocuri de noroc
UDC: 794.9
BT : Games of thought, skill and chance

GAMES OF MOTION AND SKILL
F: Jeux de mouvement et habilité
R: Jocuri de îndemânare
UDC: 796.2
BT : Sport

GAMES OF THOUGHT, SKILL AND CHANCE
F: Jeux exigeant réflexion, habilité, mémoire
R: Jocuri de gândire, memorie şi noroc
UDC: 794
BT : Recreation and sport
NT : Card games
Chess
Games of chance

Garbage
use: URBAN HYGIENE

Garden plants
use: GARDENING

GARDENING
F: Jardinage
R: Grădinărit
UDC: 635
UF : Garden plants
BT : Agriculture
NT : Ornamental gardening
RT : Horticulture
Planning of landscape

SYSTEMATIC DISPLAY

6
  Applied sciences

61
  Medical sciences

611
  Anatomy
    UF: Human anatomy
        Systematic anatomy

612
  Physiology
    UF: Human physiology

613
  Hygiene
    UF: Dietetics
        Hygiene of dwellings
        Personal health
        Sexual hygiene

614
  Public health

614.4
  Prevention of epidemics
    UF: Disinfection

614.8
  Accidents
    UF: Accident prevention

614.84
  Fire hazards
    UF: Fire prevention
        Fires

615
  Pharmacology
    UF: Therapeutics

615.2/.3
  Medicaments
    UF: Drugs

615.8
  Physiotherapy
    UF: Massage
        Radiotherapy

615.85
  Psychotherapy
    UF: Psychoanalysis

615.9
  Toxicology

616
  Pathology
    UF: Clinical medicine

616-07
  Diagnosis
    UF: Semeiology
        Symptomatology

616.1
  Cardiovascular complaints
    UF: Heart
        Pathology of the circulatory system

616.2
  Respiratory complaints
    UF: Otorhinolaryngology
        Pathology of the respiratory system
        Pulmonary complaints

616.3
  Digestive complaints

616.5
  Clinical dermatology
    UF: Cutaneous complaints
        Skin diseases

616.6
  Urogenital complaints
    UF: Genital complaints
        Pathology of the urogenital system
        Sexual complaints

616.7
  Pathology of the organs of locomotion
    UF: Clinical myology
        Clinical osteology
        Pathology of the locomotion system

616.8
  Neurology
    UF: Neuropathology
        Pathology of the nervous system
        Psychiatry
        Psychoses

617
  Surgery

617.3
  Orthopaedics

617.7
  Ophtalmology
    UF: Eye disorders

618
  Gynaecology and obstetrics
    UF: Childbirth
        Pathology of parturition
        Pathology of the female
        Physiology of pregnancy
        Pregnancy

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

801.6
  Prosody
    UF: Metre
        Rime

808
  Rhetoric

808.1
  Literary technique
    UF: Art of writing

81
  Linguistics

81'1
General linguistics

81'22
Semiotics
UF: Semiology

81'221
Nonverbal communication

81'23
Psycholinguistics
UF: Psychology of language

81'24
Practical knowledge of Languages

81'242
Native language
UF: Mother tongue

81'243
Foreign language

81'246
Multilingualism

81'25
Theory of translation

81'27
Sociolinguistics
UF: Correct usage of language

81'276
Social dialects
UF: Sociolects

81'276.4
Jargon
UF: Slang

81'276.5
Occupational slang

81'28
Dialectology

81'32
Mathematical linguistics

81'34
Phonetics
UF: Phonology

81'35
Orthography

81'36
Grammar

81'366
Morphology

81'367
Syntax

81'37
Semantics

81'373
Lexicology

81'374
Lexicography

81'38
Stylistics

81'42
Discourse analysis

81-11
Linguistic schools
UF: Comparative linguistics
Historical linguistics
Structural linguistics
Synchronic linguistics

81-2
Caracteristic features of languages

811
Individual languages

811.11
Germanic languages

811.111
English language

811.111(73)
American English

811.112.5
Dutch language
German language

811.113.4
Danish language
Norwegian language
Swedish language

811.124
Latin language

811.124'02
Classical Latin
Modern Latin

811.13
Romance languages

811.131.1
Italian language

811.133.1
French language

811.134.3
Portuguese language
Spanish language

811.135.1
Romanian language

811.14
Greek language

811.14'02
Classical Greek
Modern Greek

811.15
Celtic languages

811.16
Slavic languages

811.161.1
Russian language
Ukrainian language

811.162.3
Czech language
Polish language
Slovakian language

811.163.2
Bulgarian language
Serbo-Croatian language

811.17
Baltic languages

811.18
Albanian language

811.19
Armenian language

811.21
Indian languages

811.22
Iranian languages

811.29
Indo-European dead languages

811.34
Dead languages of unknown origin

811.35
Caucasian languages

811.361
Basque language

811.41
Hamito-Semitic languages

811.411.21
Arabic language
Hebrew language

811.42/.45
Black African languages

811.51
Ural-Altaic languages

811.511
Finish language

811.511.113
Estonian language

811.511.141
Hungarian language

811.511.161
Turkish language

811.512.1
Turkic languages

811.512.145
Tatar language

811.521
Japanese language

811.531
Corean language

811.58
Sino-Tibetan languages

811.581
Chinese language

811.8
American Indian languages

811.92
Human artificial languages
UF: Esperanto
Volapk

82
Literature

82-1
Poetry
UF: Poems

82-1/-9
Literary genres

82-2
Plays
UF: Comedies
Tragedies

82-3
Prose

82-31
Novels

82-32
Short stories

82-34
Tales

82-36
Anecdotes

82-4
Essays

82-5
Speeches

82-6
Letters

82-7
Prose satire

82-82
Anthologies

82-83
Dialogues

82-84
Aphorisms
UF: Maxims

82-91
Popular literature

82-92
Periodical literature

82-93
Children's literature

82-94
Biographies

82-95
Literary criticism

82-992
Descriptions of travels

82.02
Literary trends

82.09
Literary history
UF: Literary studies

82.091
Comparative literature

821.11
Germanic literatures

821.111
English literature

821.111(73)
American literature

821.111(94)
Australian literature

821.112.2(436)
Austrian literature

821.112.2
German literature

821.112.5
Dutch literature

821.113.4
Danish literature

821.113.5
Norwegian literature

821.113.6
Swedish literature

821.124
Latin literature

821.124'02
Classical Latin literature

821.13
Romance literatures

821.131.1
Italian literature

821.133.1
French literature

821.134.2
Spanish literature

821.134.3
Portuguese literature

821.135.1
Romanian literature

821.14
Greek literature

821.14'02
Classical Greek literature

821.15
Celtic literatures

821.16
Slavic literatures

821.161.1
Russian literature

821.161.2
Ukrainian literature

821.162.1
Polish literature

821.162.3
Czech literature

821.162.4
Slovakian literature

821.163.2
Bulgarian literature

821.163.4
Serbo-Croatian literature

821.17
Baltic literatures
UF: Latvian literature
Lithuanian literature

821.18
Albanian literature

821.19
Armenian literature

821.21
Indian literatures

821.22
Iranian literatures

821.35
Caucasian literatures

821.361
Basque literature

821.41
Hamito-Semitic literatures

821.411.16
Hebrew literature
UF: Jewish literature

821.411.21
Arabic literature

821.42/.45
African literatures

821.51
Ural-Altaic literatures

821.511.111
Finish literature

821.511.113
Estonian literature

821.511.141
Hungarian literature

821.512.161
Turkish literature

821.521
Japanese literature

821.531
Korean literature

821.58
Sino-Tibetan literatures

821.581
Chinese literature

821.8
American Indian literatures

APPENDIX 3

SAMPLE OF THE MULTILINGUAL THESAURUS BASED ON THE POCKET EDITION OF THE UDC (PTHES) WITH ENTRY TERMS IN ENGLISH

ALPHABETICAL DISPLAY

FIELD CROPS
F: Plantes de culture
R: Plante de cultură
UDC: 633
NT : Aromatic plants
Cereals
Edible roots and
tubers
Forage grasses
Forage plants
Industrial plants
Plants yielding
stimulants
Sugar plants
Textile plants

FIELDS
F: Plaines
R: Câmpii
UDC: (251)
UF : Pampas
Prairies
Savannas
Steppes
BT : Natural flat
ground

FIJI
F: Iles Fiji
R: Insulele Fiji
UDC: (961.1)
BT : Polynesia

Files
use: DATA

FINANCE
F: Finances
R: Finanţe
UDC: 336
BT : Economics
NT : Money
Public expenditure
Public finance
Public revenue
RT : Commercial law
International finance
Trade

Financial assistance
use: MATERIAL ASSISTANCE

Financial law
use: COMMERCIAL LAW

Financial need
use: PROBLEMS REQUIRING
DISTINCT FORMS OF
RELIEF

FINANCING OF ACADEMIC
STUDIES
F: Financement des
études universitaires
R: Finanţarea studiilor
universitare
UDC: 378.3
UF : Bursaries
Endowments (higher
education)
Grants (higher
education)
Scholarships
(higher education)
Subsidies (higher
education)
BT : Higher education

Fingerprinting
use: CRIMINALISTICS

Finitness and infinitness
use: FINITY AND
INFINITY

FINITY AND INFINITY
F: Fini et infini
R: Finit şi infinit
UDC: 125
UF : Finitness and
infinitness
Infinite and
boundless
Universe
(Metaphysics)
BT : Special metaphysics

FINLAND
F: Finlande
R: Finlanda
UDC: (480)
UF : Suomi
BT : Scandinavian
States

FINNISH
F: Finnois
R: Finlandeză
UDC: =511.111
BT : Finno-Ugric
languages

FINNO-UGRIC LANGUAGES
F: Langues finno-
ougriennes
R: Limbi fino-ugrice
UDC: =511.1
BT : Uralic langages
NT : Estonian
Finnish
Karelian
Lappic
Ugric languages

FIRE HAZARDS
F: Risques d'incendie
R: Incendii
UDC: 614.84
UF : Firefighting
Fires
BT : Accidents

Fireclay
use: Earthenware

Firefighting
use: Fire hazards

Fires
use: Fire hazards

FIRST AID
F: Premiers secours
R: Prim ajutor
UDC: 614.88
UF : Casualty and
ambulance services
BT : Accidents
Public health
and hygiene

First cause
use: NATURE OF GOD

First lessons
use: ELEMENTARY
EDUCATION

Fiscal practice
use: PUBLIC REVENUE

Fittings
use: SCHOOL BUILDINGS
AND EQUIPMENT

Fixed and relative
locations
use: HANDLING,
TREATMENT,

SHELVING OF BOOKS

Flat rate
use: FORMS OF PAYMENT

Flattery
use: SINCERITY

Fleet air arm
use: NAVAL AVIATOR
CORPS

Flemish
use: DUTCH

Flex time
use: HOURS OF WORK

Flirrtation
use: ENGAGEMENT

Flirtation
use: LOVE

FLOODED LAND
F: Terre inondée
R: Teren inundat
UDC: (255)
UF : Polders
BT : Natural flat
ground

FLOWING WATERS
F: Cours d'eau
R: Ape curgătoare
UDC: (282)
UF : Rivers
BT : Inland waters
NT : Waterfalls

Flying corps
use: MILITARY AVIATION

Flying saucers
use: UNIDENTIFIED
FLYING OBJECTS

Folk festivals
use: NATIONAL FESTIVALS

Folk humour
use: FOLK TALES

FOLK LITERATURE AND DRAMA
F: Littérature populaire
R: Literatură populară
UDC: 398.5
UF : Folk plays
Mumming
BT : Folklore
RT : Popular theatre

Folk plays
use: FOLK LITERATURE
AND DRAMA

Folk sayings
use: POPULAR WISDOM

FOLK TALES
F: Contes populaires
R: Poveşti populare
UDC: 398.2
UF : Fairy stories
Folk humour
Narations
Stories
BT : Folklore

Folk wisdom
use: FOLKLORE

FOLKLORE
F: Folklore
R: Folclor
UDC: 398
UF : Folk wisdom
Old wives' tales
Popular traditions
Superstitions
BT : Ethnology
NT : Dream books
Folk literature
and drama
Folk tales
Popular beliefs
and customs
Popular wisdom
Supernatural

Foot soldiers
use: Infantry

FORAGE GRASSES
F: Herbes fourragères
R: Ierburi furajere
UDC: 633.2
UF : Meadow and pasture
grasses
BT : Field crops

FORAGE PLANTS
F: Plantes fourragères
R: Plante furajere
UDC: 633.3
BT : Field crops

FORCE
F: Force
R: Forţă
UDC: 118
UF : Energy
BT : Cosmology

FORCED CHANGE
F: Fluctuations forcées
R: Fluctuaţii forţate
UDC: 314.045
BT : Population change

FORCED MIGRATION
F: Migrations forcées
R: Migraţii forţate
UDC: 314.7.045
BT : Migrations

Forecasting
use: FUTURE OF
KNOWLEDGE

FOREIGN
F: Etranger
R: Străin
UDC: (1-87)
SN : Elsewhere than
one's own country
UF : Abroad
RT : Foreigners

Foreign contingents
use: SPECIAL CORPS

Foreign language
use: PRACTICAL KNOWLEDGE
OF LANGUAGES

Foreign Office
use: MINISTRY OF
FOREIGN AFFAIRS

FOREIGN POLICY
F: Politique extérieure
R: Politică externă
UDC: 327
UF : International
affairs
International
relations
World politics
BT : Politics
NT : Imperialism
International
blocs
Internationalism
Movements for
integration

Political influence on other states

FOREIGN TRADE
F: Commerce extérieur
R: Comerţ exterior
UDC: 339.5
UF : Customs (trade)
External trade
Free trade
International trade
BT : Trade

FOREIGNERS
F: Résidents étrangers
R: Străini
UDC: -054.6
UF : Non-nationals
BT : Persons according to ethnic characteristics
RT : Foreign

FORENSIC MEDICINE
F: Médecine légale
R: Medicină legală
UDC: 340.6:61
BT : Auxiliary legal sciences

Forensic sciences
use: AUXILIARY LEGAL SCIENCES

Forgery
use: OFFENCES AGAINST PUBLIC CREDIT, MORALITY

Formal cause
use: MATTER

FORMER ASIATIC USSR
F: Partie asiatique de l'ex URSS
R: Partea asiatică a fostei URSS
UDC: (57)
BT : Former Union of Soviet Socialist Republics
NT : Kazakhstan
Kyrgyzstan
Russian Federation in Asia
Tajikistan
Turkmenistan
Uzbekistan
RT : Former European USSR

FORMER CONTINENTS
F: Continents primitifs
R: Continente primitive
UDC: (217)
BT : Land areas
RT : Legendary countries

FORMER EUROPEAN USSR
F: Partie européenne de l'ex URSS
R: Partea europeană a fostei URSS
UDC: (47)
BT : Former Union of Soviet Socialist Republics
RT : Former Asiatic USSR

FORMER UNION OF SOVIET SOCIALIST REPUBLICS
F: Ex Union des Républiques Socialistes Sovietiques
R: Fosta Uniune a Republicilor Sovietice Socialiste
UDC: (47+57)
UF : USSR
NT : Former Asiatic USSR
Former European USSR

Forms of business organization
use: ECONOMIC ALLIANCES

Forms of class warfare
use: CLASS WAR

Forms of gouvernment
use: SUPREME AUTHORITY

Forms of instruction
use: ORGANIZATION OF STUDY AND TUITION

FORMS OF PAYMENT
F: Formes de salaires
R: Forme de plată
UDC: 331.23
UF : Flat rate
Piece rate
Time-based rate
BT : Salaries

FORMS OF POLITICAL ORGANIZATION
F: Formes d'organisation politique
R: Forme de organizare politică
UDC: 321
BT : Politics
NT : Historical forms of government
Modern forms of government
RT : Sociography

FORMS OF PRE-SCHOOL EDUCATION
F: Formes de l'enseignement préscolaire
R: Forme de învăţământ preşcolar
UDC: 373.2
UF : Crèches
Day nurseries
Kindergartens
Nursery schools
BT : Schools providing general education

Forms of worship
use: RELIGIOUS USAGE

FORTIFIED COUNTRY
F: Terres fortifiées
R: Terenuri fortificate
UDC: (258)
BT : Natural flat ground

FOSSIL RESINS
F: Résines fossiles
R: Răşini fosilizate
UDC: -032.38
BT : Carbonaceous and hydrocarbon minerals

FRANCE
F: France
R: Franţa
UDC: (44)

Frankness
use: SINCERITY

Franks
use: REGIONS OF GERMANIC TRIBES

Fraud
use: HONESTY

Fraud (law)
use: OFFENCES AGAINST
PUBLIC CREDIT,
MORALITY

FREE CHURCHES IN BRITAIN
F: Eglises protestantes
britanniques
R: Biserica liberă în
Anglia
UDC: 285/288
UF : Nonconformists
BT : Christian churches
NT : Methodists
Puritans
Unitarianism

Free love
use: POLYGAMY AND
MONOGAMY (ETHICS)

Free newsrooms
use: FREE READING ROOMS

FREE PUBLIC LIBRARIES
F: Bibliothèques
publiques libres
R: Biblioteci publice cu
intrare liberă
UDC: 027.4
SN : Libraries
established and
supported by
individuals or
institutions
BT : General libraries

FREE READING ROOMS
F: Salles de lecture de
fréquentation gratuite
R: Săli de lectură cu
acces gratuit
UDC: 027.9
UF : Free newsrooms
BT : General libraries

Free religion
use: LAICISM

Free thinking
use: LAICISM

Free trade
use: FOREIGN TRADE

FREEDOM
F: Liberté
R: Libertate
UDC: 123.1
UF : Indeterminism
BT : Freedom and
necessity

FREEDOM AND NECESSITY
F: Liberté et nécessité
R: Libertate şi necesitate
UDC: 123
BT : Special
metaphysics
NT : Freedom
Necessity

Freedom of religion
use: CHURCH AND CIVIL
AUTHORITIES

Freedom of will
use: WILL

FRENCH
F: Français
R: Franceză
UDC: =133.1
BT : Romance languages

FRENCH GUYANA (FRANCE)
F: Guyane Française
R: Guiana franceză
UDC: (882)
BT : Guiana
territories

FRENCH POLYNESIA
F: Polinésie Française
R: Polinezia franceză
UDC: (963)
BT : Polynesia

FRENCH SWITZERLAND
F: Suisse romane
R: Elveţia franceză
UDC: (494.4)
BT : Switzerland

Friendliness
use: VIRTUES AND
QUALITIES

Friendship (ethics)
use: LOVE

FRIENDSHIP (ETHNOLOGY)
F: Amitié (ethnologie)
R: Prietenie (etnologie)
UDC: 392.7
UF : Feuds
Hospitality
Vendettas
BT : Customs in
private life

Frigid regions
use: POLAR REGIONS

Friulian
use: RHAETO-ROMANCE
LANGUAGES

Frontiers
use: PERSONS AND THINGS
IN INTERNATIONAL
LAW

Full-length stories
use: NOVELS

Function of the skin
use: MOTOR FUNCTIONS
(PHYSIOLOGY)

FUNCTION, VALUE, UTILITY,
CREATION, DEVELOPMENT OF
LIBRARIES
F: Fonction, valeur,
création et
développement des
bibliothèques
R: Funcţia, valoarea,
utilitatea, crearea
şi dezvoltarea
bibliotecilor
UDC: 021
BT : Librarianship

Functions with more
variables
use: LOGIC OF
RELATIONS

Functions with one
variable
use: LOGIC OF
CONCEPTS

FUNDAMENTAL TYPES AND
PRINCIPLES OF EDUCATION
F: Types et principes
fondamentales de
l'éducation
R: Tipuri si principii
fundamentale ale
educaţiei

6
Applied sciences

61
Medical sciences

611
Anatomy
UF: Human and compared anatomy

611.1
Angiology
UF: Arteries
Blood vessels
Cardiovascular system (anatomy)
Heart
Veins

611.1/.8
Systematic anatomy
UF: Organs

611.2
Respiratory system
UF: Lungs
Windpipe

611.3
Digestive system
UF: Alimentary canal
Intestines
Mouth
Stomach
Teeth
Throat

611.4
Lymphatic system
UF: Ductless glands
Endocrine organs
Haematopoietic organs

611.6
Urogenital system
UF: Bladder
Genital organs
Kidneys
Reproductive organs
Urinary and sexual organs

611.7
Skeletal, locomotor and integumentary systems
UF: Dermatology: skin
Myology: muscles
Osteology: bones

611.8
Nervous system (anatomy)
UF: Brain
Ears
Eyes
Nose
Sensory organs
Spinal cord

611.9
Anatomical topography
UF: Regional anatomy
Regions of the body
Somatology

612
Physiology
UF: Human and comparative physiology

612.1
Blood and its circulation

612.1/.8
Systematic physiology

612.2
Respiration

612.3
Alimentation
UF: Digestion
Eating (physiology)
Nutrition

612.4
Glandular functions
UF: Excretion
Secretion

612.5
Body temperature
UF: Animal heat
Effects of heat and cold
Hyperthermia
Hypothermia
Thermal processes (physiology)

612.6
Reproduction
UF: Ageing
Coitus
Development (physiology)
Growth (physiology)
Parturition (physiology)
Puberty (physiology)

612.7
Motor functions (physiology)
UF: Function of the skin
Locomotion (physiology)
Muscular actions
Voice

612.8
Nervous system (physiology)
UF: Cerebral functions
Hearing(physiology)
Olfaction (physiology)
Sensory organs (physiology)
Sight (physiology)
Smell (physiology)
Taste (physiology)
Touch (physiology)

613
Hygiene

613.1
Climatic factors

613.2
Dietetics

613.3
Drinks
UF: Curative drinks
Liquid diet
Medicinal waters

613.4
Personal hygiene
UF: Clothing (hygiene)

613.5
Hygiene of dwellings

613.6
Occupational health and hygiene

613.7
Health and hygiene of leisure

613.8
Health and hygiene of the nervous system
UF: Health and ethics
Sexual hygiene
Sexual life (hygiene)

613.9
Health and hygiene in relation to race, age, sex

614
Public health and hygiene

614.1
Population, depopulation

614.2
Public and professional organization of health
UF: Medical ethics
Regulation of medical profession

614.3
Sanitary inspection and control
UF: Inspection of foods
Inspection of medicines

614.39
National health services

614.4
Prevention of epidemics
UF: Quarantine

614.6
Cemetery hygiene
UF: Disposal of the dead

614.7
Hygiene of air, water, soil
UF: Pollution and its control

614.8
Accidents
UF: Hazards
Risks

614.8.084
Accident prevention

614.84
Fire hazards
UF: Fire fighting
Fires

614.88
First aid
UF: Casualty and ambulance services

615
Pharmacology
UF: Therapeutics

615.1
General and professional pharmacy

615.4
Pharmaceutical preparations
UF: Medical equipment
Medical material

615.8
Physiotherapy, radiotherapy
UF: Physical therapy

615.9
General toxicology
UF: Intoxication

616
Pathology
UF: Clinical medicine

616-001
Traumata
UF: Injuries
Wounds

616-006
Tumours
UF: Cancer
Neoplasms
Oncology

616-051
Medical staff

616-052
Medical patients

616-07
Diagnosis

616-08
Treatment

616-083
Nursing

616-089
Operative treatment
UF: Operative technique

616-089.5
Surgical anaesthesia

616-7
Medical instrumentation and equipment

616.1
Cardiovascular diseases
UF: Cardiac diseases
Cardiology

616.1/.9
Special pathology

616.2
Respiratory diseases
UF: Diseases of the respiratory organs
Otorhinolaringology (ear, nose and throat)
Pulmonary diseases

616.3
Gastroenterology
UF: Diseases of the mouth, stomach, intestine, liver

616.314
Stomatology
UF: Dentistry
Odontology
Orthodontics

616.314-7
Dental instruments and materials

616.4
Endocrinology

616.5
Cutaneous diseases
UF: Clinical dermatology

616.6
Urogenital diseases
UF: Diseases of the kidneys, bladder, male reproductive organs

616.7
Clinical osteology and myology
UF: Diseases of the bones, skeletal and locomotor systems
Diseases of the muscles

616.8
Neurology
UF: Neuropathology

616.89
Psychiatry
UF: Abnormal psychology
Morbid mental states
Psychoses

616.9
Communicable diseases
UF: AIDS
Contagious diseases
HIV infection
Infectious diseases
Sexually transmitted diseases

617
Surgery

617.3
Orthopaedics

617.7
Ophthalmology
UF: Eye disorders and treatment

618
Gynaecology and obstetrics

618.1
Gynaecology
UF: Pathology of the female

618.2
Obstetrics
UF: Cyesiology
Gravidity
Midwifery
Physiology of pregnancy
Pregnancy
Tocology

618.3
Pathology of pregnancy

618.4
Childbirth
UF: Delivery
Eutocia
Natural birth
Parturition (obstetrics)
Physiology of labour

618.5
Pathology of parturition
UF: Difficult birth
Dystocia
Pathology of labour

618.7
Pathology of the puerperium
UF: Post-partum period

633
Field crops

633.1
Cereals
UF: Grain crops

633.2
Forage grasses
UF: Meadow and pasture grasses

633.3
Forage plants

633.4
Edible roots and tubers

633.5
Textile plants

633.6
Sugar plants

633.7
Plants yielding stimulants
UF: Beverage plants
Narcotic plants
Tobacco

633.8
Aromatic plants
UF: Condiment plants
Medicinal plants
Oleaginous plants
Tanning plants

633.9
Industrial plants

8
Language, linguistics, literature

80
Philology

801.6
Prosody
UF: Metre
Rhythm
Rime
Verse pattern

801.7
Studies of philology

808
Rhetoric
UF: Effective usage of language
Oratory

808.1
Literary technique
UF: Authorship
Creative writing
Literary activity
Writing in publishable form

808.2
Editing

808.5
Rhetoric of speech
UF: Public speaking

81
Linguistics

81-11
Linguistic schools and trends

81-112
Historical linguistics
UF: Diachronic linguistics

81-114
Synchronic linguistics
UF: Static linguistics

81-2
Characteristic features of languages

81-23
Living languages

81-24
Dead languages
UF: Extinct languages

81-25
Spoken languages

81-26
Written languages
UF: Literary languages

811
Individual languages

81:1
Philosophy of linguistics

81:39
Ethnolinguistics

81`0
Origins and periods of languages
UF: Phases of development of languages

81`01
Old period
UF: Archaic period

81`02
Classical period

81`04
Middle period

81`06
Modern period

81`1
General linguistics

81`22
Semiotics
UF: Non-verbal communication
Semiology

81`23
Psycholinguistics
UF: Psychology of Language

81`24
Practical knowledge of languages
UF: Bilingualism
Foreign language
Monolingualism
Native language

81`25
Theory of translation
UF: Interpreting
Literal translation
Simultaneous translation

81`26
Standardization of language
UF: Control of language
Language planning

81`27
Sociolinguistics
UF: Correct and incorrect usage of language
Idioms
Parlances
Slangs
Usage of language

81`28
Dialectology
UF: Areal linguistics
Geographical linguistics

81`282
Dialects
UF: Contact languages
Language variants
Local and regional language
Vernaculars

81`32
Mathematical linguistics
UF: Computational linguistics

81`33
Applied linguistics

81`34
Phonetics
UF: Phonemics
Phonology

81`35
Orthography
UF: Graphemics
Pronunciation
Spelling

81`36
Grammar
UF: Morphology
Parts of speech
Syntax

81`37
Semantics

81`373
Lexicology
UF: Antonyms
Archaisms
Categories of words
Etymology
Homonyms
Loan words
Onomastics
Synonyms

81`374
Lexicography

81`38
Stylistics

81`42
Text linguistics
UF: Discourse analysis

81`44
Typological linguistics