Victoria Francu
The Central University Library—Bucharest

Building a Multilingual Thesaurus Based on UDC

Abstract: Once the library has been through a process of transition from traditional library procedures to automated ones, natural language searching became a necessity for both indexers and searchers. Therefore, aside from the precoordinated classified catalogue we started to build a dictionary of terms in order to make postcoordinated search possible in keeping with the UDC notations assigned to each bibliographic record. After a while we came to the conclusion that the dictionary needed a control of its terms so that synonymous concepts and semantic ambiguities be avoided. The project presented in this paper shows how reality imposed the improvement of the quality of indexing and hence of the searching possibilities. It also shows the reasons why we consider a multilingual thesaurus based on UDC an ideal indexing and searching device. The experiment applied on Class 8 of UDC illustrates the way the UDC tables can be quite successfully used in building a thesaurus due to their qualities and how their limitations can be overcome by a thesaurus. An appendix to the paper contains a sample of the multilingual thesaurus given in both alphabetical and systematic layouts.

1. Introduction

As an institution expected to meet the information needs of a particular category of people (that is to say students, university teaching staff and researchers) the Central University Library of Bucharest—Romania is increasingly confronted with the necessity to find the most suitable solutions to provide search tools able to cope with these needs. Indexing is providing an intermediary between the user and the document on which the information retrieval is largely dependent. It means all the same extracting the main subject matter from each document and representing it in a way that makes retrieval possible. Any piece of information that has not been included in the indexing system will be not identified when searching. Hence the importance of the way the indexing is made upon the quality and effectiveness of the index, and ultimately, upon the reliability of the database as a whole.

2. The Idea of a Word System for Subject Indexing

The preoccupations for an as accurate as possible indexing language was dictated by the growing rate of the library collection reflected in the database. Besides, the number of requests for a certain subject was greater than for works whose authors and titles were known. The existence of an OPAC offered the possibility of using uncontrolled keywords for searching, aside from UDC codes. A word system for information retrieval proved soon to be what our readers expected. We witness nowadays a general tendency to replace subject access with keywords. Because what relevance may have an answer like: "821.112.2'375.5" given to an inquirer interested in the German pronunciation? This piece of information is not able to fit into the structure of knowledge in his mind. Words belonging to natural language are more suitable to the user's demands in terms of precision. The same is true when speaking about currentness, a keyword for a new concept can be immediately used in indexing as access point (e.g., 'airbrush painting'). Still then, the vocabulary used for searching is necessary to be known by the indexer and the searcher as well. Therefore, we are already speaking about an agreement between the author of the book as a source of a message.
or information and the inquiry matching it so that it can be decoded by the searcher with the
least noise possible. We deal then with a vocabulary of keywords which is very flexible on
condition that the searcher has the adequate term in mind (e.g., 'entomology' used for the study
of insects). But, among other things, free indexing has the disadvantage of little control on the
synonyms and homographs, the consistency of terms being also not its very strong point.

So, we came to the conclusion that the best solution was to build a controlled vocabulary
that should be used in searching in order to avoid the vagueness and inaccuracies of the natural
language which result particularly from homonymy, synonymy and syntactic ambiguities
(Fugmann, 1993, 67-107). Our decision was that to each precoordinated UDC notation
should be attached one or more descriptors so that postcoordinated search was possible
(Riesthuis and Bliedung, 1991, 109). By means of a controlled vocabulary the scattering of
synonymous terms denoting the same concept (e.g., 'popular ballad' and 'folk ballad') in an
alphabetic list is avoided. The relations, both hierarchical and associative, are meant to make
retrieval more effective on condition that the thesaurus is well structured and the descriptors
are adequately assigned.

3. Why a Multilingual Thesaurus?

Having a word system to search with was just a stage in the evolution of the search devices
our library provides. It is obvious why this system was necessary to be in the mother-tongue
of the library users. But, for some years now, it has happened more than once that not only
Romanian people come to use our library for their information needs. Then the question
arises, how should we satisfy the demands of this category of users? So far, each bibliographic
record was assigned one (or more) UDC notation(s) and several corresponding subject head­
ing in Romanian as access points to the subject. Given these circumstances the problem still
remains since never-the-less the Universal Decimal Classification "speaks with every tongue" it
does not say too much to most of the readers.

Moreover the library has been recently connected to Internet, so our database, being
included in an international network can be reached from different places outside our country
and having only Romanian subject headings for search is not enough. That is how the idea of
a multilingual thesaurus was born in order to meet the present requirements of the library
users. We consider English and French as most suitable languages to this purpose, the
VUBIS system our library uses having a provision for them in the subject headings manage­
ment.

4. Why Still UDC?

As mentioned before the main strong points of UDC are that it is international and not
language dependent. The international retrieval and exchange of information is made easier by
a systematic approach based on symbols than by an alphabetical one based on one language.
Information retrieval systems based only on words do not enable the user to see the structure
of a subject and search up and down hierarchies or related topics that a classification scheme
permits (McIlwaine and Buxton, 1993, 103). Consistency and control of notations are two
other strong points of UDC that become more difficult to maintain in systems based on words
alone. The hierarchical nature of most of the notations makes it easy for indexers to create
verbal indexes to classified catalogues to be used in searching in an OPAC. As a matter of
facts this is the feature on which most of the project undertaken here is based. Synonymous
terms which are often given in the descriptions of UDC notations in the schedules prove to be
a valuable source of terminology for those interested in building thesauri. None-the-less, cross references are ready to be used for concepts having related meanings though in different domains.

According to McIlwaine and Buxton (1993, 110), the hierarchical structure of UDC gives a solution to multiple searches for broader and narrower terms in online catalogues. The truncation sign placed after a notation gives a comprehensive search for narrower terms belonging to a more general concept without any effort to find out each of these narrower terms separately. If a searcher is interested by the grammar of the Italian language in general he is free to choose between:

811.131.1'36

which will give him all items covering different aspects of the grammar of the Italian language and:

811.131.1'36?

which will give several other numbers such as:

811.131.1'362 Comparative grammar
811.131.1'364 Generative grammar
811.131.1'366 Morphology
811.131.1'367 Syntax
811.131.1'367.2 Sentences generally
811.131.1'367.3 Types of sentences

according to his interests.

The independence of language is an advantage in online systems because irrespective of the country a certain searcher belongs to he will find a UDC notation in the UDC edition in his language and the same number will be found by another one provided that he has a UDC version in his own language. The access to the UDC database being given both searchers have equal opportunities ending in equally good results, irrespective of the language they used in searching.

5. Building the Thesaurus

As Ray Prythearch writes in his Glossary (Harrod's, 1995, 635), a thesaurus may be defined either in terms of its function or of its structure. In terms of function it is a terminological control device used in translating from the natural language of documents into a more constrained system language. In terms of structure, a thesaurus is a controlled and dynamic vocabulary of systematically and generically related terms covering a specific domain of knowledge.

As mentioned before, the first stage of the subject indexing activities in our institution was the natural language or free indexing stage when each record was assigned keywords starting from the text of the UDC notations. This had advantages in more precision in indexing and better recall in retrieval. But disadvantages were soon to occur while the dictionary of terms was growing in dimensions. Most of the inconsistencies were due to various forms of terms assigned to define the same concept given either the singular-plural differences or synonyms or words whose meanings were somewhat overlapping with each other.

E.g., Ballads / Ballad / Popular ballads / Popular ballad
Rhetoric / Art of speaking
Public opinion / Social opinion
Etymology / Origins of the language
This way the information was scattered all over the alphabetical index without any possibility to control the use of terms. So we decided that the control of terms is obviously needed the way to reach it being provided by the use of the UDC tables themselves. That was the most valuable source for both the selection of terms and for their relationships, either hierarchical or associative. There were still some other problems related to the meanings of the words, some of them being very difficult to disseminate. A group of indexers was then established, each of the members being a subject specialist and we tried, and still try, by sharing responsibilities, to overcome the problems we had to face. Thus, we started 'cleaning' the dictionary of keywords having in mind the advantages offered by controlled vocabularies (Lancaster, 1986, 161):

- reducing the semantic ambiguities by distinguishing among various meanings of homographs;
- promoting consistency in the representation of the subject matter through control of synonyms;
- facilitating the comprehensive searches by some superimposed structure that links semantically related terms.

6. Problems and Solutions with Multilingual Thesaurification

The UNISIST Guidelines for the Establishment and Development of Multilingual Thesauri (1981) recognize three possible approaches to the construction of multilingual thesauri:

1. Begin from scratch and collect terms from all the languages involved, seeking equivalents after the terms are collected;
2. Take an existing thesaurus in one language and translate it into the other languages;
3. Take two or more existing thesauri, in different languages and merge them.

Considering the dictionary of indexing terms our library uses as an existing thesaurus, we may say we are in the second type of approach to the construction of a multilingual thesaurus. Therefore, the main activity to be undertaken is translating, or rather finding equivalents for the descriptors already used. Several kinds of equivalence will be met such as:

- perfect equivalence:
  E.g., Rhetoric = Rhétorique = Retorica (82.085)
- partial equivalence:
  E.g., Various other Genres littéraires = Alte genuri
genres = divers = literare (82-9)
- non-equivalence such as the case of the Romanian word 'Doina' defining a species of lyrical poem specific to the Romanian folklore expressing feelings of love, or revenge, or mourning.
- borrowings or linguistic calques:
  E.g., Computational = Linguistique = Lingvistica
  linguistics computationale computationala (81'322)

As much as the form of the terms is concerned, the ISO Guidelines (ISO 2788) recommends the use of nouns or syntagms and not adjectives. The most familiar or officially adopted term in a particular language should be chosen as descriptor, the synonyms or equivalents being mentioned in the thesaurus as nondescriptors

E.g., Art of writing = Art de l'écriture = Arta scrisului

UF Creative writing (82.085)

There are two distinctive situations in which homographs may occur. They may be either
- **intralingual homographs** - when the form of a term has more than one significance within the frame of one language:

  E.g., *Sonnet* as:
  
a) Italian verse form consisting of 14 lines with a formal rhyming scheme;
  
b) short poetic form.

or

- **interlingual homographs** - when the same form of a term is to be found in more than one language:

  E.g., *Rondel* - a French verse form and a short poetic form too has the same spelling in all three languages.

  Solutions:
  
  - qualifier that is attached to the term and is not a scope note;
  
  E.g., *Hermeneutics and Hermeneutics (Religion)*
  
  *Symbolism (Literature) and Symbolism (Art)*

  used for the trend or movement in literature and art at the end of the 19th and beginning of the 20th century but

  *Symbolism in literature*
  
  *Symbolism in art*
  
  *Symbolism in religion*

  used to denote the use and meanings of symbols in each of these domains

  - distinction by using in each situation another form (singular and plural);

  E.g., *Sonnet (801.675.2)*
  
  *Sonnets (82-193.3)*

  each of these being likely to be combined by means of the Boolean operator AND with Prosody and with Poetry respectively during the search.

  The descriptions of the UDC notations prove to be an excellent source both for terms and scope notes:

  E.g., *Philology (801)*
  
  *SN Used for general questions relating both to linguistics and literature*
  
  *Rhythm (801.65)*
  
  *SN Used for verse line and syllable patterns according to numerical characteristics*
  
  *Rhétorique (82.085)*
  
  *SN Art de parler en publique*

  When synonyms are given in the descriptions of UDC notations the preferred term becomes descriptor and the synonym (or quasi-synonym) nondescriptor:

  E.g., *Satirical and humorous poems (82-17) becomes:*

  *Satirical poems (82-17)*
  
  *UF: Humorous poems*

  The combinations recommended by Class 82 - Literature are treated as syntagms in the thesaurus:

  E.g., *Aesthetics of literature (82:111.852)*

  *F: Esthétique littéraire*
  
  *R: Estética literaria*
  
  *UF: Aesthetics (111.852) + Literature (82)*

  In general, the literary genres are no problem when translating the corresponding terms represented in the UDC tables by the special auxiliary subdivisions. The selection of terms has still to be carefully made in order to avoid inconsistencies and overlapping concepts.
The treatment of literatures of individual languages deserves a special attention when speaking about bilingual or multilingual countries like Canada and Switzerland, for instance. We may be tempted to use

- Canadian literature (English) 821.111(71)
- Canadian literature (French) 821.133.1(71)

The same is true for

- Swiss literature (French) 821.133.1(494)
- Swiss literature (Italian) 821.131.1(494)
- Swiss literature (German) 821.112.2(494)

Yet, in order to be consistent with the UDC notations, the qualifiers should be not language names but country names, this implying a change in the descriptors. Then we have:

- French literature (Switzerland)
- Italian literature (Switzerland)
- German literature (Switzerland)

One more example is

- Dutch literature 821.112.5
  and Dutch literature (Belgium) 821.112.5(493)
  UF Belgian literature (Dutch)

The same special attention should be paid when it comes to authors whose literary output is created in a different language (or country) than their native language (or country).

Thus we will have

- Romanian literature (France) 821.135.1(44)

and, in addition to that

- French literature + Romanian authors for writers like Eugene Ionesco and Emil Cioran.

7. Conclusions

It goes without saying that the experiment with Class 8 can be generalized for the other classes of the UDC. Provided that a set of common rules starting from the international standards are established, the duplication of work can be avoided and efficiency achieved to the benefit of our work.

Indexing with words from a multilingual thesaurus will enhance retrievability of bibliographic items by means of multiple search possibilities. The systematic thesaurus offering more control on the terms used in subject indexing may also be considered as a guarantee for consistency in the use of descriptors for documents with similar subjects. In addition to that it enables the user to make comprehensive searches by giving him the possibility to move up and down the hierarchical scale.

We may speak here about the compatibility of the information languages the way Prof. Riesthuis (1995) distinguishes several categories in that. He states that information languages that are enumerated in some parts and faceted in others are potentially more compatible the less precoordinated they are. Yet, full compatibility can hardly be reached, the experiment presented in this paper shows that there is a practical possibility to turn to account the tremendous experience and reliability of UDC with a view to using it in the multilingual exchange of information.

References


ALPHABETICAL THESAURUS—ENGLISH

ALLEGORICAL POEMS
F: Poèmes allégoriques
R: Poeme alegorice
UDC: 82-191
BT: Poetry
RT: Verse fables

ART OF WRITING
F: Art d’écriture
R: Arta scrisului
UDC: 82.081
UF: Creative writing
BT: Literary technique

BALLADS
F: Ballades
R: Balade
UDC: 82-144
BT: Poetry

CHILDREN LITERATURE
F: Littérature pour enfants
R: Literatura pentru copii
UDC: 82-93
BT: Literature

CLASSICAL STROPHES
F: Strophes classiques
R: Strofe clasice
UDC: 801.672
BT: Strophes
RT: Saphic strophes

Creative writing
USE: Art of writing

Diachronic linguistics
USE: Historical linguistics

EDITING
F: Edition
R: Editare
UDC: 82.083
BT: Literary technique

ENGLISH VERSE FORMS
F: Formes anglais de versification
R: Forme engleze de vers
UDC: 801.673.2
BT: Germanic verse forms

EPIC POEMS
F: Poèmes épiques
R: Poeme epice
UDC: 82-13
BT: Poetry

EPIGRAMS
F: Epigrammes
R: Epigrame
UDC: 82-193.2
BT: Poetry
Exegesis
USE: Hermeneutics

FRENCH VERSE FORMS
F: Formes françaises de versification
R: Forme franceze de vers
UDC: 801.675.1
BT: Romance verse forms
RT: Quatrain
Rondeau
Rondel

GENERAL LINGUISTICS
F: Linguistique générale
R: Lingvistica generala
UDC: 81'1
BT: Linguistic schools

GENERATIVE GRAMMAR
F: Grammaire générative
R: Gramatica generativa
UDC: 81'364
UF: Transformational grammar
RT: Generative linguistics
UF: Diachronic linguistics
BT: Linguistic schools
GENERATIVE LINGUISTICS
F: Linguistique générative
R: Lingvistica generativa
UDC: 81-116.6
BT: Structuralism
RT: Generative grammar
Generative semantics

GENERATIVE SEMANTICS
F: Semantique générative
R: Semantica generativa
UDC: 81’372
RT: Generative linguistics

GERMAN VERSE FORMS
F: Formes allemandes de versification
R: Forme germane de vers
UDC: 801.673.3
BT: Germanic verse forms
RT: Minnesänger stanza

GERMANIC VERSE FORMS
F: Formes germaniques de versification
R: Forme germanice de vers
UDC: 801.673
BT: Strophes
NT: English verse forms
German verse forms

HERMENEUTICS
F: Herméneutique
R: Hermeneutica
UDC: 801.73
SN: Used for textual criticism and interpretation
UF: Exegesis
BT: Studies of philology
RT: Hermeneutics (Religion)

HERMENEUTICS (RELIGION)
F: Herméneutique (Religion)
R: Hermeneutica (Religie)
UDC: 22.07
RT: Hermeneutics

HISTORICAL LINGUISTICS
F: Linguistique historique
R: Lingvistica istorica
UDC: 81-112

Humorous poems
USE: Satirical poems

ITALIAN VERSE FORMS
F: Formes italiennes de versification
R: Forme italienne de vers
UDC: 801.675.2
BT: Romance verse forms
NT: Madrigal
RT: Sonnet

LINGUISTIC SCHOOLS
F: Écoles linguistiques
R: Scoli lingvistice
UDC: 81-11
UF: Linguistic trends
NT: Historical linguistics
Structuralism
Synchronic linguistics
RT: General linguistics

Linguistic trends
USE: Linguistic schools

LITERARY FORMS AND GENRES
F: Formes et genres littéraires
R: Forme si genuri literare
UDC: 82/-9
NT: Various other genres

NT: Literary schools
USE: Literary trends

LITERARY TECHNIQUE
F: Technique littéraire
R: Tchnica literara
UDC: 82.0
BT: Literature
NT: Art of writing
Editing
Rhetoric
025.43
Multilingual thesauri
F: Thésauri multilingues
R: Tezaure multilingue
BT: Thesauri
RT: Multilingual dictionaries

22.07
Hermeneutics (Religion)
F: Herméneutique (Religion)
R: Hermeneutica (Religie)
RT: Hermeneutics

801
Philology
F: Philologie
R: Filologie
SN: Used for general questions relating to both linguistics and literature
NT: Prosody

801.6
Prosody
F: Prosodie
R: Prozodie
BT: Philology
NT: Metre
Rhyme
Rhythm
Strophes

801.63
Metre
F: Métrique
R: Metrica
BT: Prosody

801.65
Rhythm
F: Rythme
R: Ritm
SN: Used for verse line and syllable patterns according to numerical characteristics
BT: Prosody

801.66
Rhyme
F: Rime
R: Rima
BT: Prosody

801.67
Strophes
F: Strophes
R: Strofe
UF: Stanzas
BT: Prosody
NT: Classical strophes
Geramic verse forms
Nordic verse forms
Oriental verse forms
Romance verse forms

801.672
Classical strophes
F: Strophes classiques
R: Strofe clasice
BT: Strophes
RT: Saphic strophes

801.677
Oriental verse forms
F: Formes orientales de versification
R: Forme orientale de vers
BT: Strophes

801.7
Studies of philology
F: Étude de la philologie
R: Studii de filologie
NT: Hermeneutics
801.73
Hermeneutics
F: Herméneutique
R: Hermeneutica
SN: Used for textual criticism and interpretation
UF: Exegesis
BT: Studies of philology
RT: Hermeneutics (Religion)

81'1
General linguistics
F: Linguistique générale
R: Lingvistica generala
RT: Linguistic schools

81'28
Dialectology
F: Dialectologie
R: Dialectologie
UF: Areal linguistics
NT: Dialects

81'282
Dialects
F: Dialects
R: Dialecte
SN: Used for local and regional language
BT: Dialectology
NT: Aromanian dialect

81'364
Generative grammar
F: Grammaire générative
R: Grammatica generativa
UF: Transformational grammar
RT: Generative linguistics

81'372
Generative semantics
F: Semantique générative
R: Semantica generativa
RT: Generative linguistics

81'374
Lexicography
F: Lexicographie
R: Lexicografie

81'374.2
Special dictionaries
F: Dictionnaires spéciales
R: Dictionare de specialitate
SN: Used for special dictionaries according to the vocabulary they contain

81'374.3
Explanatory dictionaries
F: Dictionnaires explicatifs
R: Dictionare explicative

81'374.5
Phonetic dictionaries
F: Dictionnaires phonétiques
R: Dictionare fonetice

81'374.7
Dictionaries (Ordering)
F: Dictionnaires (Ordre)
R: Dictionare (Ordonare)