Discover Thesauri: State of the Art in Hungary

Rudolf Ungváry [1], Tamás Radnai [2]

[1] National Library, Budapest, Hungary, ungvary@oszk.hu
[2] Budapest Tech and Hungary.Network, Inc.
Szépvölgyi út 39, H-1037 Budapest, Hungary, radnait@hungary.com

Abstract – Role of Thesauri in the recent library, documentation and Internet retrieval practice is discussed. Similarities and differences between the English and Hungarian language Thesauri are compared and special features of Hungarian Thesauri are described. Directions of strategic development in order to extend the use of thesauri are explained. Basic properties of Thesauri as structured dictionaries and organizing systems are shown and several examples help to understand how to use Thesaurus entries correctly.

1 Historical Remarks

Suppose you type in the word “thesaurus” in any indexing service of the Web (e.g., Google or the Hungarian Heureka) or any Internet Catalogue (e.g., Yahoo or the German Metager), and then read several technical books on the subject; you will soon observe that it is something which has been used for many decades. The first Hungarian language Theasaurus was published 90 years ago, and they have been used for information retrieval purposes for about 30 years already.

A software package called Thesaurus is attached in the English version of Microsoft Word and it is called in Hungarian “tezaurusz” since the summer of 1989 as latest [1]. The first Theasaurus in Hungary has been published by a secondary school teacher named Ferenc Póra in 1906 and a second edition was also presented a year later. It was published next in 1991 in unchanged form [2], and since then even the copies of this edition became rarities. The textbook has been stolen from the libraries for long; they can be searched in vain; and this is the only dictionary of Hungarian language organized by topics. Its fate is similar to its famous English predecessor’s, the Roget Thesaurus. However, there is a significant difference between the “Póra” and “Roger” Thesauri; but this is neither in the language, since languages are equal; nor the in the quality, since „Póra” itself is excellent as well. The real difference is in the number of editions: „Póra” has been published three times during 94 years, while the number of publications of Roget Thesaurus is near to hundred since 1852, and there is a great abundance of its variations. Both the MS Word Thesaurus and Póra’s Thesaurus are of everyday language. Their aim is to support the composition and speech. In the world of traditional library and documentation practice and recently in the information technology and Internet search services special thesauri are built for aiming at information retrieval. Professor Karen Spark Jones (1935–) at Cambridge University, who is a pioneer of automatic textual analysis, clustering and automatic Thesaurus building has been written on purpose already decades ago in his paper about the history of thesauri that Theasaurus is a well known („familiar") product [3]. This statement is given not only by chance since as being an Englishman, Jones obviously could see Thesaurus from his childhood already just simply looking at his bookshelf. In the English speaking world it is natural to find the Roget, Webster or any other standard Thesaurus in many apartments along the Bible. In Hungary this is hard to say.

2 Life Cycle of Thesauri

2.1 Mortality of Hungarian Thesauri

It is possible that the above fact is why Hungarian Thesauri have so fragile fate. There have been 57 information retrieval Thesauri built or being built since 1970 in Hungary. [3]. For the time being only 15% of them is in use. The mortality ratio is incredibly high. A possible explanation is that if the Thesauri do not belong directly to any special community, organization or even to an individual person, they will exit soon. There will be nobody to know them, and not even those will remember them, who will have to prepare later on a new information retrieval Thesaurus often in the same subject. There is no other solution just rediscover Thesaurus.

This ever repeated rediscovery and admiration of the Thesauri is a very valuable event from both human and sensational point of view, but for rationalism and professional science it is infinitely prodigal. And it underrates the related Hungarian upper education institutions. This is how it can happen that somebody with a background of several university terms' education as a librarian prepares a thesaurus on Egyptology, driven by her own enthusiasm, working alone and for her pleasure; then she publishes it on the Web in a really professional way both as far as the Thesaurus itself and its HTML representation is concerned, but still she does not have any education on Thesauri, since the experience piled up in Hungary and world-wide was forgotten to be taught to her. As a result, she does not know that Thesauri have a Hungarian standard (MSZ 3718) since more than 20 years which is of equal quality with any International or other national standards. As a consequence, she prepared all relation symbols and the Thesaurus form according to ISO 2888 and in addition, in English [4]. Most of the university students do not even see a Hungarian Thesaurus, although the subject must trivially be part of the librarians' educational courses.

2.2 How a Thesaurus can Survive?

The Hungarian documentation and informatics experiences are also lacking a recognition that if a given Thesaurus has once been prepared, but is not applied institutionally for a long time, or it is not understood that any Thesaurus is never finished, and it should be permanently extended and modified; and that the new, modified or deleted lexical units have thousands of consequences on the relation network, which should be reconsidered each time otherwise the Thesaurus will fall apart; and mostly if there is no person in the institution who takes all these problems as a personal matter or matter of his profession (perhaps finding some joy in it as well) – i.e, there is no edition board set up with 1-2 people employed, nothing will remain from the Thesaurus. But even in case when a Thesaurus is saved in good conditions, its fate is strictly bound to the institution in which it is used. If the institution is closed, there is a little chance that any other institution will use the Thesaurus further on. The reason is that the fate of Thesauri is bound to the coverage of collection. Two Thesauri are never identical, and even if two Thesauri of the same subject are used in two different institutions, the Thesauri themselves are not kept in identical forms. This is analogous to the language itself: every person speaks it somewhat differently. The professionals in Hungary do not use more than 1-2 Thesauri and even these are of foreign origin. The always repeated discovery of the "genre" is related to several information retrieval Thesauri well known from international use. The number of less known Thesauri is of order of ten thousand, and they are run by dedicated institutions. These dictionaries have very sophisticated structures both by language and – necessarily – by thoughts. They comprise a huge pile of information, and therefore they require a very well developed, structured professional-spiritual infrastructure in order to survive. It happens often that the Thesaurus is in use for many years but the when changing the old ones the new colleagues often simply throw away the whole Thesaurus or truncate it significantly because it is "too complicated" and since than they use a primitive dictionary for indexing purposes without relations between words. Moreover, there is not known in Hungary among the professionals what is the difference between a "concept" and its "nomination". There are people who believe that Thesauri are concept vocabularies. Or even more: they don't know that it is not possible to prepare concept vocabularies at all, since it is obvious that the exclusive concept vocabularies is the mind itself. And what could be prepared is nothing else than dictionaries. One set of these dictionaries are the standard language dictionaries, a further set of which are the information retrieval language dictionaries, and one subset of the latter ones are the descriptor based information retrieval language dictionaries called Thesauri.

3 Basic Properties of Thesauri

3.1 Thesaurus is a Structured Dictionary

All Thesauri, either being of everyday language, linguistics, terminology, taxonomy, or information retrieval are so called *structured dictionaries*. They are structured because they list the most important semantic relations between the lexical units. Let us consider the example of "Dog". Its relations are the following: it is *Predatory*, and *Domestic* animal. Its species are e.g., *Greyhound, Puli* or *Poodle*. They roam in *Herds* when living free, and live in a *Kennel* when domesticated. Their function is e.g., *Watch the house* or *Lead a blind*. Their properties include *barking* and *canine madness*, and a synonym for "dog" is "*canis familiaris*". A standard Thesaurus dictionary entry based on the above example relations (without claiming completeness) is:

Dog

H canis familiaris
F Domestic,
Predatory
A Greyhound,
Puli,
Poodle
T Herds,
Kennel
R Watch the house,
Lead the blind
X Barking,
Madness.

Another example for programming languages (without claiming completeness again):

Algorithmic programming language

F high level programming language
A Basic,
C,
Pascal,

ANSI standard

F National standard
R ANSI standard language,

FORTRAN

F Compiled language
E ANSI standard language
X Modular software,

Compiled language

F High level programming Language
A Compiled Basic,
Fortran,

High level programming language

F Symbolic language
A Algorithmic programming language,
Interpreted language,
Compiled language,

National Standard

A ANSI,
MSZ

Standard

A National standard,
International standard
E Standardization
R Standard
X Quality assurance

No keywords can be found in Thesauri, but descriptors and non-descriptors. A keyword is a synonym for a secondary identifier, and as such, it is never found in any information retrieval language dictionary, but always present in an indexed document. In the mentioned cases the indexing has been performed by descriptors taken from the Thesauri and therefore a keyword has been formed from the descriptors. In an automatic indexing system keywords originate from textual words appearing in certain locations e.g., title, annotation etc.

3.2 A Thesaurus is an Organizing System

Thesauri, just like the Universal Decimal Classification system (UDC), are organizing systems according to their functionality. Organizing systems can be differentiated along various aspects. UDC is *monohierarchic* according to it sstructure (1 : N), *precoordinated* by its use, it is *based on artificial language* by its type of information retrieval language, *universal* by its coverage of collection, and a *generalizing (synthetic) classification system* by its role. On the contrary, most of the information retrieval languages are *polyhierarchic* (M : N), *post* *coordinated*, *natural language based*, *special*, *individualizing (analytical) subject name based* and *descriptor-based* information retrieval languages. A dictionary of all these is a Thesaurus.

Thesauri appeared for general access parallel to the computer science. A lots of professional, remote access data bases can be accessed by Thesauri. But this fact by far will not exclude the use of monohierarchic search dictionaries traditionally called *classification systems* (e.g., UDC). The two types of organizing systems will not substitute each other. On the contrary, they complete each other. Classification systems and information retrieval systems are used usually parallel in the modern documentation and information institutions and especailly on the Web as well and most often UDC is selected for classification system. One can say that nothing has been changed in essence, on this field, in spite of the enormous development of computer science. Even more, an advance is to be experienced at international level for the UDC system. That means that these organizing systems are part of the most important development trends along with their „rivals", the dictionaries of information retrieval languages (Thesauri and subject catalogues). Unfortunately, in Hungary UDC's are dying after 1990 since nobody took care of their maintenance due to financial reasons.

3.3 Secret of Thesaurus Building: Grab It and Make It!

Of course, one has to locate financial sources first. But once they are available, content analysis of any given data base is possible. It is a simple routine work, by using both of a special thesaurus and a hierarchic classification system. The first task is the preparation of these two information tools. Selecting the working language depends on the data base: obviously one has to build English language organizing systems for English data bases and Hungarian systems for Hungarian data bases. Accordingly, each natural language requires its language versions in classification and information retrieval systems. Special attention should be devoted to the mixed language data bases, but discussion of this subject is beyond the scope of this paper.

References

[1] S. János Petőfi: Present situation of Thesaurus building, with special care on scientific and technical-economical information. Budapest: OMKDK, 1969. p. 167 (Theory and practice of scientific information, 12, in Hungarian.)

[2] Ferenc Póra: Handbook of Hungarian words and sentences: contains thirty thousands of synonym words and sentences in eight hundred logical groups. Budapest: Gondolat, 1991. p. 452 (in Hungarian)

[3] Karen Spark Jones: Some thesauri's history. Aslib Processing, Vol. 24, No. 7, 1972. pp. 4000-411

[4] Rudolf Ungváry: Actual publications. Thesauri in Hungary. Budapest, 2000. <http://www.mvkkvar.hu/tezaurusz/tezkozt.html> (in Hungarian)