TESAURVAI: Extraction, Annotation and Term Organization Tool

3 authors:

Jesús Cardeñosa
Universidad Politécnica de Madrid

113 PUBLICATIONS   **196** CITATIONS

Carolina Gallardo
Universidad Politécnica de Madrid

37 PUBLICATIONS   **82** CITATIONS

Angeles Maldonado Martinez
Spanish National Research Council

47 PUBLICATIONS   **27** CITATIONS

TESAURVAI: Extraction, Annotation and Term Organization Tool

Jesús Cardeñosa
Univ. Politécnica de Madrid, Dpto. de Inteligencia Artificial
carde@opera.dia.fi.upm.es
Carolina Gallardo
Univ. Politécnica de Madrid, Dpto. de Inteligencia Artificial
carolina@opera.dia.fi.upm.es
Ángeles Maldonado-Martínez
Centro de Ciencias Sociales y Humanas (CCHS) – CSIC
urbasoc@cindoc.csic.es

1. Introduction: Tool Context

Each concrete field of disciplinary or thematic specializations makes use of its own terminology. The compilation, definition, and organization of terms used in a given domain are a basic task, because it becomes the base for the constitution of specialized terminology resources of great usefulness. Thesauri are a type of terminological resource of increasing relevance at the present time; frequently used in the recovery and localization of information in digital environments. The hierarchic organization of terms in a thesaurus helps to optimize searches both in close information systems and open ones like Internet.

TESAURVAI is a tool for the extraction, annotation and organization of specialized terms in concrete domains taken from digitized texts. TESAURVAI is one of the tools developed in the context of the project "Búsqueda documental sobre Patrimonio Cultural basada en recursos léxicos multilingües - *Patrilex*" (HUM2005-07260/FILO), sponsored by the I+D+I National Plan, National Program of Humanities of Spain, having as one of its objectives the creation of a methodology and the necessary tools for the creation of multilingual lexical resources, allowing to support a multilingual documentary search system. PATRILEX works in the concrete domain of Cultural Heritage, and as its source it uses the texts in the section dedicated to this subject in the Web of the Spanish Ministry of Culture.

2. Tool Functionality

The fundamental contribution of TESAURVAI is the integration of an *extractor of words and phrases* and of a *thesaurus manager* in a unique tool. TESAURVAI is able of extracting words and phrases from one or several texts, making possible their annotation and structuring according to the meaning of these terms. Currently, the existent tools carry out only one or another of these functionalities. In the remaining, the extractor and the thesaurus manager functionalities are described.

2.1 Term Extractor
TESAURVAI term extractor allows selecting and transferring words and extracted phrases from one or several digitized texts. In order to carry out such tasks the following actions are executed:

- o identification of words and phrases
- o automatic selection of words an phrases from a given number of repetitions
- o elimination of stop words and phrases containing such stop words
- o creation of a list with words and phrases automatically selected according to required criteria
- o user's manual selection of words and phrases from the previous presented listing
- o transfer of terms finally selected by the user to the management module

The extraction of words and phrases takes place from one or several digitized texts. The program can work with texts in Word or pdf. In the extractor configuration the user can select the files containing the working texts by indicating the path where they can be found. The user has also to define the permissiveness of the filter for both simple words and phrases, that is, he/she must indicate the maximum or minimum number of repetitions for a word or phrase within the selected texts in order that these can become part of the automatically selected term listing. In the case of the phrase listing, the user must also define the number or words he/she wants it to contain. Once these parameters are set up, the program automatically analyzes the selected texts and composes the required listing.

In each one of the listings, the selected words or phrases are visualized, as well as the number of repetitions obtained in the text or work texts. The user can organize this listing by alphabetical order or by frequency of appearance in the text or texts used as corpus. These automatic listings can be printed or exported.

Finally, for the *selection and transfer of terms to the management module*, the user selects those words and phrases of interest from the listing previously made in automatic form. Once this selection has been made, the tool stores the terms in the management module of the thesaurus in order to be able to work later with them.

In the TESAURVAI there is the possibility of using a *dictionary of stop words*. The program is provided with a dictionary of stop words, used by default in case the user does not provide an alternative one. In addition, the user can make as many dictionaries of stop words as he/she wishes and use them when he/she may consider it convenient. The words listed in the selected dictionary of stop words are eliminated from the listings both in the case of words or phrases. In this second case, the phrases beginning or containing the chosen empty words do not appear in the final listing.

Figure 1 Extractor Screen

2.2 Thesaurus Manager
For the TESAURVAI creation it has been followed the Norm UNITES 50106:90 ("Directives for the establishment and development of a monolingual thesaurus"), which presents the engineering specifications that a thesaurus must fulfill. The TESAURVAI functionalities are the following:

2.2.1 Creation and elimination of thesauri
The program can create and manage more than one thesaurus. The creation of a new thesaurus is simple, because it consists in giving it a name in order that all the program functionalities may be operative for it. It is possible to open two or more thesauri at the

same time and to work simultaneously with all of them. The elimination of an already existing thesaurus is equally simple, because the user only needs to select the option "Eliminate Thesaurus" from the file menu. Before total elimination a message of confirmation appears warning about such elimination.

2.2.2 Term management
Once a TESAURVAI thesaurus is created, it can carry out the basic functions of a standardized thesaurus manager. For term management this tool allows:

- *Creation, modification and elimination of terms*
Each thesaurus can be composed of an unlimited number of terms. Once a thesaurus is created the terms can be inserted in it, using three different procedures:

o extracting words and phrases from one or several selected texts (process already explained in the above section)
o introducing terms directly by the user in a manual form
o importing term listings already existing

When a new term is entered, the tool registers the following information: the denomination of the term, its status (Candidate, Approved, Not Valid), the term type (Preferred and Non-preferred) and date of insertion. In a thesaurus there are two types of different terms: preferred and non-preferred. These two kinds of terms differ in the types of relations that they can maintain with other terms. If the term registration takes place in a manual form, the user must specify each one of those data, except for the date of insertion which is automatically assigned. In the massive incorporation of terms, either by an import or by an extraction, the tool assigns by defect the label of "preferred" to all types of terms, and the label of "candidate" to all status of terms. Once the massive insertion is finished, the user can modify these labels in those cases where he/she considers it convenient.

Equally, the denomination of any term can be modified at any moment, and also the basic information existing about it and it is possible to eliminate all the terms as well.

- *Term annotation*
The program offers two different fields to introduce annotations on the terms. The field "scope note" must be used by the user to introduce the definition of a term, or an explanatory note about its meaning, when it has more than one. In the field "personal note" it is possible to write down all those incidences considered convenient: source of extraction of the term, number of repetitions in base corpus, etc.

- *Definition of categories and distribution of terms among them*
The terms of a thesaurus are usually grouped around great semantic categories. TESAURVAI allows to create any number of desired categories and to distribute all the terms composing a thesaurus among them. The program allows a term to belong to more than one category.

- *Creation and elimination of relations among terms*
TESAURVAI is able to support the basic relations established among the terms of a thesaurus:

o *Equivalence*: relation between two terms with the same meaning
o *Hierarchy:* relation between two terms in which there is a conceptual dependency of generality/specificity
o *Association:* relation between two terms maintaining a semantic link different to synonymy or hierarchy

Each one of these relations must fulfill a norm of distinct reciprocity, since all of these relations are established between two terms that the tool is able to sustain. At the same time, TESAURVAI does not allow contradictions between relations due to the fact that it has the necessary operative controls in order that they do not take place. In case that a contradictory relation is tried, the program is able of detecting it and of displaying a message about the corresponding error. Finally, it is necessary to say that the tool admits multihierarchies, that is, the fact that a specific term can have more than one generic term.

- *Term edition:*
The program offers the possibility of showing individualized information about all and each one of the terms composing the thesaurus. The information displayed is:

o *General information***:** type, status, date of insertion, approval, invalidation and modification
o *Category*
o *Personal Note*
o *Scope Note*
o *Relations*: term(s) equivalent(s), generic term(s), specific term(s), and associated term(s)

2.2.3 Visualization of the general listing of terms
The program visualizes by default the alphabetical listing of all the terms composing the thesaurus in the screen. The user can add information to this listing: relations, status, categories, and existence or not of personal note. It is possible to choose to visualize all the terms of the thesaurus or only the preferred ones. Also a typesetter differentiation between preferred and non-preferred terms can be applied.

2.2.4 Term search
A thesaurus can manage several thousands of terms, that is why it is important to be able to make powerful searches to select those of special interest in a given moment. TESAURVAI has two search options:

- *simple search* where the user has to choose between localizing terms starting by a specific character chain or containing those characters.
- *advanced search* to use more complex criteria in term localization. The considered search options are the following:

o search of terms belonging to a specific category
o search of terms whose status is one of three possible: Candidate, Approved or Rejected)
o search of words of phrases within the text of annotations

2.2.5 Generation of reports

TESAURVAI is able to make the following reports to be printed or exported later to a text file:

- *General listing*, alphabetic listing of all the terms composing the thesaurus with all the information available about them and of all other terms with which they maintain some kind of relationship.

- *Alphabetic listing*, where all the terms composing the thesaurus appear in alphabetic order with all the information available about each one of them that can be selected by the user. The user can choose to visualize: status, notes, dates of insertion and approval, and related terms (equivalents, generics, specific, or associated).

- *Hierarchic listing*, where the terms appear by a hierarchic structure. The user can choose from which term he/she wishes to reproduce such structure. The hierarchic structure is presented using indentation and dots to express the different hierarchic levels.

- *Alphabetic/hierarchic listing*, where the terms composing the thesaurus appear in alphabetic order, reproducing in addition the section of hierarchic representation where each one of them is inserted.

- *KWIC listing*, type of permuted listing where the alphabetic terms appear with all and each one of the significant words they contain. In the KWIC permuted listing type all the words object of alphabetization occupy the central column of the presented terms and they go accompanied by all those other words composing the terms where they appear.

- *KWOC listing*, type of permuted listing where the significant words composing the terms are alphabetized and accompanied by each one of the terms containing them.

- *Listing of terms by categories*, listing where the terms grouped by considered categories appear by alphabetic order.

- *Listing of orphan terms*, alphabetic listing of those terms not depending on a more general term.

All the generated reports go preceded by the name of the thesaurus and the date of impression.

Figure 3. Report Generation screen

2.2.6 Import and export of thesaurus

TESAURVAI facilitates the interchange of data between thesauri generated by the own tool or by other thesaurus managers because it allows both the import and export of terms and the information regarding them. Listings of terms in plain text and XML can be imported. In the same way, the thesauri generated by TESAURVAI can be exported into plain text and XML.

3. Implementation and Availability

TESAURVAI has been developed in JAVA v.6. The tool is supported by a data base. The tool is compatible with any database manager having a JDBC (*Java DataBase Connectivity*) file, as MySql or Postgres. At this moment, TESAURVAI is under a test phase. There is a version 1 in progress. It is possible that as soon as July 2008 a more advanced version will be available to the public.

4. Acknowledgements

The author would like to thank Jorge Vergara for the software programming of the TESAURVAI.