VERBS & TOPIC MAPS: A PROPOSAL FOR LEGAL DOCUMENTATION FROM THE DOCUMENT CONTENT ANALYSIS PERSPECTIVE

Miguel Ángel Marzal
Professor
Dpto. Biblioteconomía y Documentación. Universidad Carlos III de Madrid

Carmen Bolaños
Associate Professor
Departamento de Historia del Derecho y las Instituciones. UNED

Jorge Morato
Visiting Professor
Dpto. Informática. Universidad Carlos III de Madrid

Javier Calzada Prado
Research Grant
Dpto. Biblioteconomía y Documentación. Universidad Carlos III de Madrid

Keywords: Legal Documentation, Document Analysis, Verb Thesauri, Topic Maps.

Abstract: A final evaluation of the project *Development of a verb thesaurus for dynamic information environments. Implementation of the ISO/IEC 12350:1999 standard* is presented as a base for the use of its results on future research projects. First, the interest of the Library and Information Science field towards verbal structures is justified for its efficacy on the documental analysis of the movement for: a) a relevant indentification, description and classification of hypermedia, b) its sufficient analytic documental adaptation to the the dynamic transverse character of the hypertext contents and c) the required redefinition of the hierarchical structure of thesauri. Then the methodological evolution of these dynamic concepts is tackled as well as its implementation to the development of verb thesauri for the Spanish legal language. Main contributions of this paper are new research lines oriented to a more specific field of theoric formulation and instrumental implementation.

1 INTRODUCTION

Every investigative process implies, as a scientific imperative, the need to evaluate not only the results, but the process itself, with the aim of providing valid conclusions for a specialized scientific community. This is the objective which guides the present contribution, to evaluate the project *Development of a thesaurus of verbs for dynamic informational settings. Application of the ISO/IEC Standard: 12350:1999*[1], initiated in 2000 with financing from the CICYT. Nevertheless, a second objective emerged during the course of the study, which is that the evaluation should serve as the basis for the development of a new Scientific-Technical Report which allows for the soliciting of a new research project capable of applying the results from the project which has just ended.

The research project initiated in 2000 followed in the wake of prior analyses of our research group,

[1] Comisión Interministerial de Ciencia y Tecnología. Plan General del Conocimiento. TIC 2000-2003.

Ángel Marzal M., Ángel Marzal M., Bolaños C., Morato J. and Calzada Prado J. (2004).
VERBS & TOPIC MAPS: A PROPOSAL FOR LEGAL DOCUMENTATION FROM THE DOCUMENT CONTENT ANALYSIS PERSPECTIVE.
In *Proceedings of the Sixth International Conference on Enterprise Information Systems*, pages 63-68
DOI: 10.5220/0002629400630068

whose purpose was the automation of the construction of thesauri through the concurrence of certain nouns in a thesaurus with certain dynamic concepts, such that, once the *corpus* of the document was indexed by nouns, these could be interrelated with verbal structures (Díaz, 2001). Precedents existed in the scientific literature which support such an effort, in which current technological developments could enable its application to the knowledge domain.
Three stimuli guided our interest, from Documentation toward verbal structures:

- Its efficiency in document analysis in the movement for an adequate identification, description and classification of hypermedia and, in particular, materials in *Virtual Reality*. The term noun indexing offers a static conceptual representation which obligates complex searches by juxtaposition or development of increasingly specific noun phrases.
- Sufficient document analysis adaptation of the dynamic transverse character of the hypertext contents, as well as the priority of the users in the recovery of online information, generating a true research trend around the *users' thesauri*. The development of markup languages and metadata offer interesting prospects.
- The required redefinition of the hierarchical structure of thesauri. A study by the *American Library Association* (ALA) in 2000 specified up to nine new large subtypes of first level associative relations as necessary, which allowed D. Tudhope, H. Alani y C. Jones (2001) to propose an expansion of the relations in automated thesauri through a noticeable increase in associative relations.

2 METHODOLOGICAL EVOLUTION OF DYNAMIC CONCEPTS

We denote the verb as a *dynamic concept* in technological settings for the information retrieval, for its capacity to represent conceptual action. Nevertheless, the verb expresses a great complexity as a vector of recovery due to its great inflexive semantic richness in a documentary context, as well as its prototypic inflection (time, mode, number) according to the literary genre and structure, as shown by Karlgren (1994, 1998). At the same time, its use as an element of information retrieval demanded an organization and ordering for which the verbal classification of B. Levin (1993) wasn't sufficient in practical application because of its complexity.

The successive methodological phases of the project, defined in different publications and conference presentations by the research team, oriented us toward a two-fold conviction: foremost is the need to normalize the verb terminologically, to segment the sentences where it acted and determine its semantic function within a context; secondarily, but no less important, is the need to delimit textual domains and styles to conform to a *corpus* adequate for both our research objectives and the intention to organize the verbs semantically for information retrieval. Within this two-fold approach, there are two fruitful areas of collaboration.

2.1 Online linguistic instruments

Verbs appeared to behave with a visible recovery efficiency within certain settings where they were provided with a well-determined function. This was the case with *concept maps*, quickly steered toward their use as didactic digital resources by J. D. Novak (1995, 45-49) where the significant learning object concepts are represented by nouns, but are ordered on the map through relationships (arches) which are defined by verbs. This illustrative model confirms for us the recovery effectiveness of the verb, if its *normalized presence* in an information system is refined and its function is determined through the definition of the semantic role and meaning in the system. We have chosen two bases:

2.1.1 WordNet

This is classified within the group of manually constructed electronic dictionaries and, without adhering to any particular domain**,** covers the majority of English nouns, adjectives, verbs and adverbs, making it an ideal instrument for clarifying meanings, semantic labeling and information retrieval.

Its terms, whether simple or compound, are organized in groupings of synonyms known as synsets, where each of the synsets corresponds to a concept. Each term has a brief description and sometimes a related phrase which demonstrates it usage, through a glossary. What makes WordNet a truly useful tool is its lexical online structure, with nearly 30,000 types of relations, typified in equivalent topics relations, hierarchically between synsets (hyperonym/hyponym; meronym/holonym), but it is also responsible for two grammatical

categories unknown in document language: adjectives and verbs, for which three possibilities are recognized: *Entailment*, that is, a verb has this relationship with another if its existence depends on the other verb; *Troponym* or verb forms; *Causal Verb*.

Research which recognizes having used WordNet demonstrates its effectiveness in multilingual settings, in the extraction of textual and iconic document information, as well as in the identification of concepts in natural language through its use for clarification, for semantic distance and the expansion of the query**.** Even more significant for us, was its capacity in the extraction and categorization of documents through the extraction of semantics features by way of grammatical categorization of names, verbs and adjectives in WordNet and the prediction of the user's interest based on a hybrid model which considered the key words and the conceptual knowledge representation of WordNet[2]. Relying on this, S. M. Harabagiu (1998, 265-269) has presented a computational model to recognize the cohesive and coherent structures of texts, with the contribution of the lexical-semantic information of WordNet, whose objective is to construct association designs between phrases and coherence relations, as well as to discover lexical characteristics within coherence categories. WordNet, then, appeared as an auxiliary instrument for the design of semantic ontologies, immediately oriented for quality informational extraction on the web, which has led Keng Woei Tan, Hyoil-Han and R. Elmasri to present the prototype WebOntEx (Web Ontology Extraction) aimed at creating ontologies to describe semantically data from the web[3].

Judith Klavans & Min-Yen Kan (1998) have investigated the automatic determination of the genre of a document depending on the category of verb in WordNet used in the same.

2.1.2 Computerized linguistic resources

The analyses and tools used by researchers in Computational Linguistics have been particularly useful, especially since our objective was becoming the realization of a *semantic of the relationships*, an aspect in which it benefited from the linguistic applications for the administration of information contents. The computer applications of Computational Linguistics permit one to lemmatize**,** a voice by identifying its canonical form, its grammatical category and its inflexion, as well as obtain diverse forms from a single canonical form or inflexion. This capability allows one to recognize, generate and manipulate the morpholexical relations in a voice. Both products of two computational linguistics research groups have been useful to us:

- CLIC, a research group lead by Prof. María Antonia Martí[4]. Among the possibilities offered, in our effort to elaborate a system which generates thesauri automatically, a large role was played by: the *parser, generator and morphological clarifier* to identify morphological interpretations of a voice through the *inflexion generator* (which canonizes a voice as a *lema* (the canonical form) and refers to it all its associated forms), the *lemmatizer* (which gives morphological information to the *lema*) and the *tagger* (which labels the components of an oration); the *parser* to identify syntagmas in a sentence; and *EuroWordNet*, particularly for its capacity to define *meanings* according to the synsets, the *synonyms* and the *relations* between different meanings of words.
- GEDLC, of the Department of Computer Science and Systems in the University of Las Palmas in Gran Canaria[5], joins the functionality of the lemmatizer, inflexion generator, clarifier, morphological generator and morpholexical relations with a system for Text Analysis, a Computational System of Morphological Administration of Spanish, but specifically a Conjugator and Verbal Lemmatizer**.**

An adequate application of the principles of both instruments has allowed us to begin resolving the problems of the treatment of a base vocabulary of the *corpus*, the terminological normalization and clarification and a segmentation of units of information. We realized that it was both possible and necessary to widen the verbal morphology to periphrasis and verbal locutions, much richer in the

[2] G. Scheler, 1996, p. 499, suggested the grammatical categorization. INFOS was analyzed by K. J. Mock & V. R. Vemuri, 1997, pp. 633-644.

[3] Meersman, R. A., 1999, pp. 30-45, identified the ontological effectiveness of WordNet, demonstrated in WebOntEx, which Keng Woei Tan, Hyoil-Han and R. Elmasri present in 2000, pp. 11-18.

[4] Available from: <http://clic.fil.ub.es> [Cited 10/28/03].

[5] Available from: <http://gedlc.ulpgc.es> [Cited 10/28/03].

associative potential of the relation, such that we can begin to consider the *verbal phraseological units*.

2.2 Specialized languages

A field in which we converge with Terminology, the science which demonstrates the intrinsic and extrinsic characters of an object in a term, which remains fixed to an idea which is related exclusively to a single name. We had confirmed that for the *semantic of the relationships* one option should be cleared up: either we accept the dependence of an author, or the dependence of a knowledge domain. We prefer the second option, particularly due to the emergence of *specialized languages*, understood as a variant of the general *language* and defined as a group of linguistic instruments (lexical, morphological, syntactic) characteristic of a subject matter and used by specialists for an optimal understanding due to its exactitude, clarity and conciseness. Its positive impact has stimulated an avenue of investigation in some languages, particularly in the economic and legal realms (Alcaraz Varó, 2002), which is the case of EPA (Español Profesional y Académico), from which legal language has become a specialized language (Martín del Burgo y Marchán, 2000). In its classification, Legal Spanish, in effect, received a superstructure, that of *normative-legal*, which establishes the guidelines for legal action, a function which is particularly useful in accepting a verbal semantic applicability. From this superstructure Legal Spanish is further classified as *legislative*, *jurisprudential*, *administrative* and *notary*. In our research design, it became evident that the best application of thesauri based on verbs was produced in legislative and jurisprudential Legal Spanish.

3 THESAURI BASED ON VERBS AND LEGAL DOCUMENTATION

We believe, then, that we had discovered a field of scientific action sufficiently restricted and distant from an excessive arbitrariness to try to refine our own tool Indexer for generating thesauri, following the methodological procedure practiced by the research team in other domains:

3.1 Creation of the *corpus*

The *corpus* was generated from digital legal documentation, extracted as complete text from the databases of the BOE (Boletín Oficial del Estado), selecting the organic Laws in section I (General Regulations), together with sentences from the Supreme Court during the last five years.

We then proceed to the marking of the verbal structures, assigning a number according to whether it is a personal form (1), an impersonal form (2) or a periphrastic form (3), with particular attention being paid to this last form, susceptible to behaving as a phraseological unit with a strong capacity in the association of concepts for information retrieval.

3.2 Processing of the *corpus*

We decided to use the following *modus operandi* to systematize the processing:

- Verbal morphological normalization in formal rules, through lemmatizers, taggers and inflexion generators to comprise a list of index terms to which the different verbal inflexions are referred.
- Construction of links from the *verbal terms* toward the document *corpus*, to recover the context. In this way, the verbal structures were ordered according to a semantic criteria of role relevance, but with attention to each *document type*.
- Organization of the personal, impersonal and periphrastic forms within each type of document.

3.3 Classification

The extraordinary inflexive semantic richness of the verb due to its action within a context, has prevented us from undertaking a verbal hierarchical classification, even in such specific and regulated domains, which is more easily accomplished among nouns**.** For this reason, we have resorted to two initial criteria of categorization, in a simple outline, according to the recovery system desired for representation of the user:

3.3.1 Functional Categorization

A criteria which has allowed us, as a first approach, to propose these first categories:

- *Promulgative*
- *Coercive*
- *Procedural*: in origin, in development, instrumental, acting, destined.
- *Argumentative*: discursive, common law-jurisprudential, contextual, demonstrative, dialogical.
- *Documentary and Testimonial*
- *Sentencing law*
- *Case law.*

3.3.2 Categorization of associative relations for the verb

We follow the recommendations of A. Tudhope *et al*., stipulating seventeen categories of relations, to which we have assigned the verbal structures according to the type of document:

- Fields of study and objects studied.
- Instrumentation: agent, operation, process.
- Action and its product.
- Action and its passive product.
- Action and its passive subject with determination of the typical effect.
- Objects and concepts with properties and effects.
- Concepts related with their origins.
- Concepts linked by chance.
- Object and contrary agent.
- Concepts and units of measure.
- Concomitance: symptom, signal, symbol.
- Constituent materials.
- Conceptual proximity: pertaining to the same conceptual *family*.
- Similarity due to equivalent expression.
- Antonymy.
- Figured localization.
- Professionals and their field of study.

All these categories are suffering a process of contrast according to their effectiveness during information retrieval, but also in document extraction: verbal inflexion is particularly useful according to the type and constituent part of legal documents.

4 FUTURE WORK

The previous approach, useful as a field in the automatic generation of thesauri with a clear application in professional legal practice, like documentalists, nevertheless, has suggested to us an exciting field of research: the retrieval of legal information using the associative potential of online information resources, once classified based on the verbal phraseological units and the new tools of information retrieval by associative semantics, a process in which the development of markup languages metadata take on a decisive importance. Two elements have emerged as our referents:

- *Ontologies*, defined by Gruber as "a formal, explicit specification of a shared conceptualization", whose basic concepts, in our case, are: the *class* (concept) that defines a category, which includes *instances* (cases) and includes the *hierarchical class* and the *subclass*; *slots*, to describe properties and characteristics of the concept; *value*, attribute applied to a class or an instance and which determines it.
- *Topic maps*, metaindexes which unite index types with semantic interrelations, attempting to provide master indexes which can be administered independently and easily updated in a technical document with highly changing contents.

Topic maps currently comprise the starting point of our research interests. The objective is that the lawyer shouldn't have to search for the sources of information to issue a legal report, but rather, that they select the concepts which they want to use and interrelate independently of how the information sources are organized or of their format (paper, database, web pages, etc.) because each concept is directly associated by hyperlinks to the information sources on the particular concept. The *topic* is always the representation of the *subject* in XML language and is not defined solely by its denomination (*topic name)*, but also by its relations (*associations)* and its subject (*scope*). In this way, the ISO/IEC 13250 norm maximizes its importance through the richness of the groupings which allow for the establishment of *associations* among concepts, in accordance with the possibilities which logic and semantics offer the world. A particularly pertinent possibility in our research is that the *topic associations* can be classified according to an *association type*, which is defined by the verbal form which unites the *topics*, that is, the *association type* is that which defines the verb that unites the *topics* in each case.

Topic maps are most relevant in their possible application to commercial and corporate information resources, for which their online help facilitates access to mass information, especially for the information portals, since websites can be organized through maps, at the same time that they offer metainformation links with other websites, having

adequately combined the data modules, with which a simplification of the access to relevant information is achieved. *Topic maps* contribute, as well, one of the main proposals to visualize the semantic web (Le Grand and Soto, 2002, 203-225).The fact that a *topic map* can present thousands of associations of different typology (*association types, roles, occurrences*, etc.) has lead to the development of iconic representations, among which trees, *browsers* and graphics stand out. The interface is analogous to a normal web page, but in reality what we visualize is always the relation of indexes used and the index which we are using will be highlighted. The index is never lost, in which the following could appear: a) Laws b) Rules c) Statutes d) Legal Manuals e) Databases f) Electronic Documents (web pages, multimedia documents). What we are attempting to accomplish is that searches are not performed based on only a single term, without importance as to its ordering or format.

5 CONCLUSIONS

The trajectory of the research project has suggested new avenues oriented toward a more specific field of theoretical formulation and instrumental application:

- The need to efficiently categorize the verbal phraseological structures in the realm of Law, as well as improve the automatic administration of relations in thesauri in the field through Computational Linguistics.
- Achieve the representation of the user (the lawyer) in the information system, through a correct design of the associative functionality of the new tools in the processing of natural language.
- Design a manner to implement and apply topic maps in the representation and recovery of legal information, testing applied models in legal practices or jurisdictional instances of a different nature.

REFERENCES

ALCARAZ VARÓ, E., 2002. *El español jurídico.* Barcelona: Ariel.

DÍAZ, I. S., 2001. *Esquemas de representación de información basados en relaciones: aplicación a la generación automática de representaciones de dominios.* Unpublished Doctoral Thesis. Director: J. Llorens. Universidad Carlos III de Madrid: Departamento de Informática.

HARABAGIU, S. M., 1998. WordNet-based inference of contextual cohesion and coherence. In *Proceedings of the 11th International Florida Artificial Intelligence Research Symposium Conference.* Menlo Park (Ca): AAAI Press, pp. 265-269.

KARLGREN, J.; CUTTING, D., 1994. Recognizing Text Genres with simple metrics using discriminant analysis. In: *Proceedings of COLING 94*, Kyoto.

KARLGREN, J., 1998. Stylistic Experiments for Information Retrieval. In: Strzalkowski, T. (ed.) *Natural Language Information Retrieval*. Kluwer, Tomek.

KLAVANS, J. ; KAN, M. Y., 1998. Role of Verbs in Document Analysis. In *Proceedings of the Conference, COLING-ACL*. Canada: Université de Montreal, 1998.

LE GRAND, B.; SOTO, M., 2002. Visualisation of the semantic web: Topic Maps Visualisation. In: *Information Visualisation 6th International Conference. 10-12 July* pp. 203-225.

LEVIN, B., 1993. *English verb classes and Alterations: a preliminary investigation.* Chicago: University Chicago Press.

MARTÍN DEL BURGO Y MARCHÁN, A., 2000. *El lenguaje del derecho*. Barcelona: Bosch.

MEERSMAN, R. A., 1999. Semantic ontology tools in IS design. In *Proceedings Foundations of Intelligent Systems.11th International Symposium ISMIS´99.* Berlín: Springer-Verlag, pp. 30-45.

MEERSMAN, R. A., 2000. Web data cleansing and preparation of ontology extraction using WordNet. In *Proceedings of the 1st International Conference on Web Information Systems Engineering.* Los Alamitos (Ca): IEEE Computational Society, vol. 2, pp. 11-18.

MOCK, K. J.; VEMURI, V. R., 1997. Information filtering via hill climbing, WordNet and index patterns. *Information Processing and Management,* v. 33 (5), 1997, pp. 633-644.

NOVAK, J. D., 1991. Clarify with concepts maps: A tool for students and teachers alike. *The Science Teacher,* 8, (7), pp. 45-49.

SCHELER, G., 1996. Extracting semantic features from unrestricted text. In *WCNN´96.* Mahwah (NJ): L. Erlbaum, p. 499.

TUDHOPE, D.; ALANI, H.; JONES, C., 2001. Augmenting Thesaurus Relationships: Possibilities for Retrieval *Journal for Digital Information*. Retrieved October 28, 2003, from: <http://jodi.ecs.soton.ac.uk /Article/v01/i08/Tudhope>.