Legal Information Management, 11 (2011), pp. 268–271
© The Author(s) 2011. Published by British and Irish Association of Law Librarians
doi:10.1017/S1472669611000867

Concepts Mapping in the Legal Domain: Some Reflections

Abstract: This paper, by Ginevra Peruginelli, examines the general issue of mapping concepts of different legal orders, with a focus on legal terminology and semantics of the law. Interoperability between legal thesauri is a possible operative solution in so far as it meets two important functions: cross-collection retrieval and cross-language retrieval. At a practical level a feasibility study, carried out by the Institute of Legal Information Theory and Techniques of the National Research Council of Italy is described. The study aims at aligning a number of thesauri used in e-government services within the EU, while testing new methods of automatic mapping.
Keywords: information retrieval; thesauri; legal language

Towards knowledge sharing through concept mapping

The theme of interoperability of thesauri and the ability to create mappings between concepts expressed in information resources is currently a much-debated topic, addressing complex issues of a semantic and technical nature. Nowadays, more and more the implementation of information retrieval systems involves managing heterogeneous collections indexed using different controlled vocabularies. Not only this requires ad-hoc protocols and software tools, but also intellectual efforts to establish a sort of equivalence between terms and related concepts across vocabularies, languages, disciplines.

Identifying correspondences between words also concerns free-text searchable collections, where a comparison has to be sought between the terms of a vocabulary used for searching and those found in the collections.

It is no doubt that thesauri are tools designed through a cooperative work of experts for the creation of an established and coherent set of terms (with appropriate relations between them) concerning a specific domain or even multiple subject areas. However, they significantly differ in detail and coverage, thus making them interoperable is quite a difficult task.

There are important initiatives aimed at drawing up guidelines for the construction and interoperability of thesauri[1] and quite a number of projects have been developed for the solution of the problems connected with the alignment of controlled vocabularies[2]. Parallel to this, the development of metathesauri is an ongoing activity in an effort to merge terms of a number of thesauri providing links to the original sources[3].

Studies and experiments in this area show some similarities between the issues related to the management of multilingual thesauri with those arising from the need to match different thesauri of the same language: this is logical, given the differences existing in the cultural universe and more generally in the information world for identifying concepts, defining the many possible shades of meaning and using a specific terminology to express them.

Strategies to create and accept possible matches vary according to context and purpose of information access systems. In case an exact equivalent cannot be found, an option which is often used is to map to a more general term. The correspondence of a broader term to a more detailed one is generally less satisfactory, and preference should be given to mapping to similar terms, in the belief that often the best option might be creating a match to a term that is neither broader nor more detailed, but approximated in meaning.

As regards the degree of equivalence, it should be noted that exact equivalence between concepts expressed by terms is undoubtedly the ideal situation and this occurs when vocabularies are built with identical concepts. Inexact equivalence can be expressed, if this is useful, according to the degree of acceptable overlap according to the context in which the mapping is performed.

In case of vocabularies with a different structure, a one-to-many match is feasible, as it is the case of a target vocabulary containing one or more concepts

https://doi.org/10.1017/S1472669611000867 Published online by Cambridge University Press

corresponding to the source concept. When the levels of specificity of vocabularies differ, more concepts in the source vocabulary can be mapped to a concept in the target vocabulary, corresponding to the nearest equivalent.

Multilingual thesauri have similar traits but also specific characteristics. Cultural differences reflected in the language have a considerable influence on matching activities and it is not always easy to get to a situation where the various languages represented have identical status. Even in this case a no-exact equivalent can be accepted or, in case a concept is missing, the gap may be narrowed by a term borrowed from the original language or by coining a new word for the occasion, possibly accompanied by a note of explanation, useful to the user performing the search.

On the whole, a fairly broad consensus exists that there is no single correct solution for the creation of mappings between controlled vocabularies, given the diversity of contexts in which match is applied. Again, the opinion is shared that a fairly similar correspondence is acceptable if search results are satisfactory, based on the recognition that a strong demand exists for accessing heterogeneous resources by users who must be put in the position of searching in a language familiar to them.

The exchange of legal concepts worldwide

The system-bound nature of legal terminology has to be carefully taken into account to reach high quality mapping of concepts; this involves interoperability of thesauri in a way that two relevant functions be performed: cross-collection retrieval and cross-language retrieval. With reference to the first, a strong need is felt nowadays for communication between controlled vocabularies used for indexing heterogeneous collections so that users can, through a single search interface, access multiple information sources. In the second case interoperability of thesauri in different languages allows users to retrieve information resources in languages different from that used in the query.

It is worth mentioning here that a term consists of two basic elements: its linguistic form, which is the word in its alphabetical realisation, and content or concept in the form of a linguistic expression. In the field of law the basic concepts of legal science constitute, together with the linguistic expressions of them, the terminology of this discipline. Legal language is in fact a subset of the natural language, emerging as a real technical language. There is no doubt that legal language is full of formal expressions and the question arises whether this formalism facilitates or impedes communication and whether the simplification of legal language necessarily implies a diminishing precision.

This language coincides, in some circumstances, with that of politics, based on value judgments, full of ambiguity and oriented to persuasion. In some cases the complexity of this terminology excludes non-experts from communicating. Hence the so-called vagueness of legal language, bringing about two conflicting requirements: to eliminate or even to keep such vagueness. The first option derives from the doctrine of legal certainty, which requires the restriction of uncertainty or ambiguity and is linked to the scientific nature of legal language, while the second originates in the need to ensure suitable flexibility and adequacy of standards to the constantly changing reality.

Ambiguity can occur when the same term has more distinct meanings, or when their meanings are relevant to other aspects of the same reality or phenomena which are somehow connected. In this case it is essential to perform a sort of contextualization that is a process of identifying the scope of a specific concept. This is especially true in a multilingual environment, due to the fact that legal languages are intimately related to legal systems, hence the great difficulty of perfect and complete transposition of languages across legal cultures.

Despite the considerable obstacles encountered in translation, there is evidence that even in the field of law translation and transposition of concepts with reasonable accuracy is possible, and acceptable results can be obtained by adopting an approach referred to as functional equivalence between the source and the target text. Functional equivalence is the relation between two terms which, being distinct from one another in each constituent part, can somehow be matched by virtue of a common element. In comparative law the expression is used of "extensional concepts". In other words, a basic semantic equivalence is reached implying the prevalent use of terms belonging to the target language so to reach more familiarity and understanding[4].

Legal information retrieval systems should be based on the functional equivalence approach and on legal analysis to be carried out not so much at the level of positive law, rather of the general theory of law. This analysis is based on more general definitions, aiming at identifying similarities and differences between legal systems. General models can be distinguished: abstract concepts (genotypes) and individual manifestations of the positive law of each system (phenotypes)[5]. In fact, the ambiguity existing between an abstract model and single manifestations can be regarded as a fault of the legal language in so far as elasticity implies a lack of precision which is undoubtedly harmful to the language of a discipline. However, this lack of precision is, according to the legal translation science, one of the factors by which translation itself and transposition of concepts is made possible. It is only thanks to a general correspondence of legal concepts that it is possible, in a comparative perspective, to talk about contract, property, marriage and, ultimately, about law.

https://doi.org/10.1017/S1472669611000867 Published online by Cambridge University Press

A European feasibility study on alignment of thesauri for e-government services

Faced with a growing demand of multilingual access to legal information, thesauri, as organizational knowledge systems consisting of a collection of terms and semantic relationships between them (hierarchical, associative and equivalence) do represent a solution to ensure that a user query be matched with properly indexed documents. In multilingual thesauri, translation has occurred at the stage of their preparation, and facilities are implemented to exploit the existing links between terms across languages[6].

Experts strongly recommend when adopting a true multi-lingual approach in constructing thesauri not to privilege a source language so to ensure concepts be expressed in each language at the same level of specificity. In fact, a multilingual thesaurus is more than just putting together several monolingual thesauri[7]. It should be based on the principle of equality between languages, respecting the conceptual framework, terminology and semantic relations of each retained language. This is undoubtedly an extremely complex task that requires substantial resources.

In this regard, it is worth mentioning a European feasibility study, carried out under the programme EUROVOC Studies – LOT2: Interoperability Feasibility Study and sponsored by the Office of Publications of the European Union (Note to tender No 10118). Its objective is the alignment of thesauri for e-government services within the European Union. The study focuses on mapping five major thesauri of interest for the European Union: Eurovoc[8], ECLAS[9], UNESCO Thesaurus[10], European Training Thesaurus (ETT)[11] and General Multilingual Environmental Thesaurus (GEMET)[12]. While the first three are multidisciplinary, ETT and GEMET can be considered specialized thesauri.

The aim of the study, committed to the Institute of Legal Information Theory and Techniques of the National Research Council of Italy (CNR-ITTIG)[13] is to assess the interoperability of the multilingual Eurovoc thesaurus with the other thesauri. The choice of Eurovoc as a reference point has been dictated by the fact that the Office of Publications aims at providing its users with cross-collection retrieval functionality: starting from Eurovoc, allow extended search facilities and enhanced integration with other databases.

At a technological level this solution takes into account the recent recommendations and studies on open standards undertaken by the W3C community and the need for semantic web-oriented solutions. In particular, the SKOS (Simple Knowledge Organization System) standard has been adopted. to describe the concepts of each thesaurus together with their inner relations, as well as the relations between concepts of the different thesauri.

ITTIG also tested automating mapping methodologies supporting the intellectual work of experts. In particular, the proposed methodology for mapping between thesauri consists of the adoption of methods of a lexical type (similarity between strings), semantic type (calculation of the semantic similarity between vector representations of concepts) and structural type (identification of the structural similarity between graph representations of concepts and their relationships). The accuracy of this methodology has been tested by comparing the results of automatic mapping with a set of mappings established by legal experts (gold standard). This knowledge base is the result of intellectual work and represents the optimal set of mappings considered to be correct. This has been experimental in the field of labour law, which is a constantly changing discipline, including different aspects covered by the thesauri under treatment[14].

The gold standard, as well as the adopted mapping strategies, has been applied to descriptors, and English has been used as reference language (*pivot*). The relationships and their respective mappings have been described using the SKOS language, limited to the broad-match, exact-match, and narrow-match relations. For the creation of the gold standard the ITTIG legal experts have used a specific interface supporting their work. To maintain local control of data (this has been necessary due to the project's purpose and schedule), a specific application, Thesauri Alignment Environment (THALEN) has been developed that, through a relational database, provides a facility to establish relationships between terms. This software allows the selection of thesauri (simultaneous view), search by field and scrolling, creation of mappings (selection of terms and relationships), export into SKOS and thesaurus merging functionalities. This application can be developed on other platforms, creating a web version and integrating it with automated tools for mapping operations. As regards implementation in other domains, careful consideration is to be given to issues related to languages' specificity, and conceptual differences existing between different cultures and disciplines.

Final remarks

Based on the above considerations, no single criterion for establishing correspondences between concepts exists. Given that the accuracy of legal language is a prerequisite for the consequences that inevitably inaccuracy and sketchiness imply in any action where law is involved, in comparing concepts for the development of legal information retrieval systems, somewhat less rigid solutions may be accepted, as compared with those to be adopted in translating legal texts, where incorrect wording can have serious implications. Less rigour does not necessarily mean the abandonment of the principle of clarity and precision that always has to drive all comparison efforts between legal concepts belonging to different legal

https://doi.org/10.1017/S1472669611000867 Published online by Cambridge University Press

orders, rather it means adopting a flexible approach. Legal information retrieval systems should ensure knowledge communication through the provision of relevant resources, even by adopting a sort of compromise. This means, for example, in the matching process, presenting a more general expression of a concept as compared to the specificity of the original, possibly making systematic use of disambiguation techniques and present multiple results, through interaction with users by making them select the desired context.

The functional equivalence approach, to be considered as carrying out the same general function of the source concept, is likely to be the most effective way to provide users with relevant documents written in various languages and differently indexed.

Footnotes

[1]ISO 2788–1986: Guidelines for the establishment and development of monolingual thesauri; ISO 5964–1985: Guidelines for the establishment and development of multilingual thesauri; ANSI/NISO Z.39.19-1993 (R2003): Guidelines for the construction, format and management of monolingual thesauri; BS 8723: Structured Vocabularies for Information Retrieval. London, British Standard Institution, 2007

[2]Some examples of projects are: CARMEN, RENARDUS, MACS, HILT projects. For CARMEN: Schröder, Albert (2008) Cross concordances of classifications and thesauri, WP12 of the CARMEN Project, http://www.bibliothek.uni-regensburg.de/projects/carmen12/index.html.en; for RENARDUS: Renardus: Academic Subject Gateway Service Europe (2000) http://www.ukoln.ac.uk/metadata/renardus/; for MACS: Landry Patrice Providing multilingual subject access through linking of subject heading languages: the MACS approach, http://www.cacaoproject.eu/fileadmin/media/AT4DL/paper-09.pdf; for HILT: Wake, Susannah and Nicholson, Dennis (2001) HILT – High Level Thesaurus Project: building consensus for interoperable subject access across communities. *D-Lib Magazine*, 7(9), http://www.dlib.org/dlib/september01/wake/09wake.html

[3]An example is the Unified Medical Language System (UMLS). Unified Medical Language System – UMLS. http://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/

[4]Weston, Martin (1990) "Theoretical and practical approaches to translation: the linguistic issues", Part I, in *An English Reader's Guide to the French Legal System*, New York, Berg, 1990, 9–42

[5]Sacco, Rodolfo (1996) Riflessioni di un giurista sulla lingua uniforme (lingua del diritto uniforme e il diritto al servizio di una lingua uniforme), *Rivista di diritto civile* 1, 57–65

[6]Sadat, Fatiha; Yoshikawa, Masatoshi and Uemura, Shunsuke, (2006) Exploiting thesauri and hierarchical categories in cross-language information retrieval, *Lecture Notes in Computer Science* 2448, 139–146

[7]Hajdu Barát, Ágnes (2008) "Knowledge organization in the cross-cultural and multicultural society", in Clément Arsenault and Joseph T. Tennis (edited by), *Culture and identity in knowledge organization proceedings of the tenth International ISKO Conference*, 5–8 August 2008, Montréal, Canada, Würzburg, Ergon Verlag, 91–97

[8]Eurovoc is the multilingual, multidisciplinary thesaurus covering the activities of the EU, the European Parliament in particular. It contains terms in 22 EU languages. http://eurovoc.europa.eu/

[9]ECLAS. http://ec.europa.eu/eclas

[10]UNESCO. http://databases.unesco.org/thesaurus

[11]ETT. http://libserver.cedefop.europa.eu/ett/en/

[12]GEMET. http://www.eionet.europa.eu/gemet

[13]The ITTIG researches into and applies information and communication technology to the areas of law and legal language, legislative technique, legal decision-making, and the training of lawyers. It also researches into themes related to law and public information and computing policy. For further information: http://www.ittig.cnr.it/IndexEng.htm

[14]Francesconi, Enrico, Faro, Sebastiano and Marinai, Elisabetta (2008) "A framework for semantic mapping between thesauri", in T. Janowski, T.A. Pardo (edited by) *Proceedings of the 2nd International Conference on theory and practice of electronic governance (ICEGOV)*, Cairo, 1–4 December 2008, 251–257

Biography

Ginevra Peruginelli, graduated in law, obtained a Ph.D. in Telematics and Information Society at the University of Florence and received her Master's degree in Computer Science at the University of Northumbria, Newcastle. Since 2003 she has been entitled to practice as a lawyer. She has been working at the Institute of Theory and Techniques of Legal Information of CNR (ITTIG) first as a fellow and from 2007 as legal researcher. Since 2005 she has been a professor under contract on legal informatics at the Law Faculty of University of Perugia. Her current research interests include user legal information requirements, strategies for online legal information retrieval, analysis of legal language in a multilingual environment and evaluation criteria for legal web sites.

https://doi.org/10.1017/S1472669611000867 Published online by Cambridge University Press