TKE 2012

Terminology and Knowledge Engineering Conference

New frontiers in the constructive symbiosis of terminology and knowledge engineering

Workshop

Challenges to knowledge representation in multilingual contexts

Universidad Politécnica de Madrid
Escuela Técnica Superior de Ingenieros de Caminos, Canales y Puertos

Challenges to knowledge representation in multilingual contexts

Workshop Program

09:00 – 09:45

Strategies in automatic traversal of Wikipedia articles for mining multilingual resources

Andrés Domínguez Burgos, Koen Kerremans, Rita Temmerman

09:45 – 10:30

Cross-Cultural Concept Mapping of Standardized Datasets

Fumiko Kano Glückstad

10:30 – 11:00 Coffee Break

11:00 – 11:45

Acquisition, Representation, and Extension of Multilingual Labels of Financial Ontologies

Thierry Declerck, Hans-Ulrich Krieger, Dagmar Gromann

11:45 – 12:30

Supporting collaboration in multilingual ontology specification: the conceptME approach

Manuel Silva, António Lucas Soares, Rute Costa

12:30 - 13:00 discussion /debate

13:00 – 14:30 - Lunch

14:30 – 15:15

Translation Politics and Terminology in Legal Texts for better community networking.

Frieda Steurs, Hendrik J. Kockaert

15:15 – 16:00

Subject Librarians operating in a multi-cultural and multi-linguistic context: an ontology-based approach to semantic cataloguing and information retrieval

Deborah Grbac, Luca Losito, Andrea Sada, Paolo Sirito

16:00 – 16:30 discussion – closing session

Editors

Rute Costa	New University of Lisbon, CLUNL, Portugal
Manuel Silva	IPP/ISCAP and INESC TEC, Portugal
António Lucas Soares	University of Porto and INESC Porto, Portugal

Organizing Committee

António Lucas Soares	University of Porto and INESC Porto, Portugal
Rute Costa	New University of Lisbon, CLUNL, Portugal
Christophe Roche	University of Savoie, France
Frieda Steurs	Lessius /KULeuven, Belgium
Manuel Silva	IPP/ISCAP and INESC TEC, Portugal

Workshop Programme Committee

Alessandro Oltramari	Carnegie-Mellon University, USA
António Lucas Soares	University of Porto and INESC Porto, Portugal
Asuncion Gómez Pérez	Universidad Politécnica de Madrid, Spain
Carla Sofia Pereira	Polytechnic Institute of Porto and INESC Porto, Portugal
Cláudia Santos	University of Aveiro, Portugal
Christophe Roche	University of Savoie, France
Elena Montiel-Ponsoda	Universidad Politécnica de Madrid, Spain
Frieda Steurs	University of Lessius, Belgium
Guadalupe Aguado de Cea	Universidad Politécnica de Madrid, Spain
Gerhard Budin	University of Vienna, Austria
Margaret Rogers	University of Surrey, UK
Manuel Silva	Polytechnic Institute of Porto and INESC Porto, Portugal
Patrick Drouin	University of Montréal, Canada
Philipp Cimiano	Semantic Computing Group, CITEC - University of Bielefeld
Piek Vossen	VU University Amsterdam, Netherlands
Rute Costa	New University of Lisbon, CLUNL, Portugal
Thierry Declerck,	DFKI - Language Technology Lab, Germany
Wim Peters	University of Sheffield - Natural Language Processing group, UK

Table of contents

Author Index

Preface

To meet the increasing demands of the complex inter-organizational processes and the demand for continuous innovation and internationalization, it is evident that new forms of organisation are being adopted, fostering more intensive collaboration processes and sharing of resources, in what can be called collaborative networks (Camarinha-Matos, 2006:03). Information and knowledge are crucial resources in collaborative networks, being their management fundamental processes to optimize.

Knowledge organisation and collaboration systems are thus important instruments for the success of collaborative networks of organisations having been researched in the last decade in the areas of computer science, information science, management sciences, terminology and linguistics. Nevertheless, research in this area didn't give much attention to multilingual contexts of collaboration, which pose specific and challenging problems. It is then clear that access to and representation of knowledge will happen more and more on a multilingual setting which implies the overcoming of difficulties inherent to the presence of multiple languages, through the use of processes like localization of ontologies.

Although localization, like other processes that involve multilingualism, is a rather well-developed practice and its methodologies and tools fruitfully employed by the language industry in the development and adaptation of multilingual content, it has not yet been sufficiently explored as an element of support to the development of knowledge representations - in particular ontologies - expressed in more than one language. Multilingual knowledge representation is then an open research area calling for cross-contributions from knowledge engineering, terminology, ontology engineering, cognitive sciences, computational linguistics, natural language processing, and management sciences.

This workshop joined researchers interested in multilingual knowledge representation, in a multidisciplinary environment to debate the possibilities of cross-fertilization between knowledge engineering, terminology, ontology engineering, cognitive sciences, computational linguistics, natural language processing, and management sciences applied to contexts where multilingualism continuously creates new and demanding challenges to current knowledge representation methods and techniques.

In this workshop six papers dealing with different approaches to multilingual knowledge representation are presented, most of them describing tools, approaches and results obtained in the development of ongoing projects.

In the first case, Andrés Domínguez Burgos, Koen Kerremansa and Rita Temmerman present a software module that is part of a workbench for terminological and ontological mining, *Termontospider*, a wiki crawler that aims at optimally traverse Wikipedia in search of domain-specific texts for extracting terminological and ontological information. The crawler is part of a tool suite for automatically developing multilingual termontological databases, i.e. ontologically-underpinned multilingual terminological databases. In this paper the authors describe the basic principles behind the crawler and summarized the research setting in which the tool is currently tested.

In the second paper, Fumiko Kano presents a work comparing four feature-based similarity measures derived from cognitive sciences. The purpose of the comparative analysis presented by the author is to verify the potentially most effective model that can be applied for mapping

independent ontologies in a *culturally influenced domain*. For that, datasets based on standardized pre-defined feature dimensions and values, which are obtainable from the UNESCO Institute for Statistics (UIS) have been used for the comparative analysis of the similarity measures. The purpose of the comparison is to verify the similarity measures based on the objectively developed datasets. According to the author the results demonstrate that the Bayesian Model of Generalization provides for the most effective cognitive model for identifying the most similar corresponding concepts existing for a targeted socio-cultural community.

In another presentation, Thierry Declerck, Hans-Ulrich Krieger and Dagmar Gromann present an ongoing work and propose an approach to automatic extraction of information from multilingual financial Web resources, to provide candidate terms for building ontology elements or instances of ontology concepts. The authors present a complementary approach to the direct localization/translation of ontology labels, by acquiring terminologies through the access and harvesting of multilingual Web presences of structured information providers in the field of finance, leading to both the detection of candidate terms in various multilingual sources in the financial domain that can be used not only as labels of ontology classes and properties but also for the possible generation of (multilingual) domain ontologies themselves.

In the next paper, Manuel Silva, António Lucas Soares and Rute Costa claim that despite the availability of tools, resources and techniques aimed at the construction of ontological artifacts, developing a shared conceptualization of a given reality still raises questions about the principles and methods that support the initial phases of conceptualization. These questions become, according to the authors, more complex when the conceptualization occurs in a multilingual setting. To tackle these issues the authors present a collaborative platform – conceptME - where terminological and knowledge representation processes support domain experts throughout a conceptualization framework, allowing the inclusion of multilingual data as a way to promote knowledge sharing and enhance conceptualization and support a multilingual ontology specification.

In another presentation Frieda Steurs and Hendrik J. Kockaert present us TermWise, a large project dealing with legal terminology and phraseology for the Belgian public services, i.e. the translation office of the ministry of justice, a project which aims at developing an advanced tool including expert knowledge in the algorithms that extract specialized language from textual data (legal documents) and whose outcome is a knowledge database including Dutch/French equivalents for legal concepts, enriched with the phraseology related to the terms under discussion.

Finally, Deborah Grbac, Luca Losito, Andrea Sada and Paolo Sirito report on the preliminary results of a pilot project currently ongoing at UCSC Central Library, where they propose to adapt to subject librarians, employed in large and multilingual Academic Institutions, the model used by translators working within European Union Institutions. The authors are using User Experience (UX) Analysis in order to provide subject librarians with a visual support, by means of "ontology tables" depicting conceptual linking and connections of words with concepts presented according to their semantic and linguistic meaning.

The organizers hope that the selection of papers presented here will be of interest to a broad audience, and will be a starting point for further discussion and cooperation.

The Editors
Rute Costa
Manuel Silva
António Lucas Soares

Strategies in automatic traversal of Wikipedia articles for mining multilingual resources

Andrés Domínguez Burgos, Koen Kerremans, Rita Temmerman

CVC, Erasmushogeschool Brussel, Belgium
{andres.dominguez.burgos, koen.kerremans; rita.temmerman}@ehb.be

Abstract. In this article we present Termontospider, a wiki crawler that optimally traverses Wikipedia in search of domain-specific texts for extracting terminological and ontological information. The crawler is part of a tool suite for automatically developing multilingual termontological databases, i.e. ontologically-underpinned multilingual terminological databases. The focus is on analyzing the best value for internal links, categories and other metadata to assign weights and search mechanisms in network traversal .

Keywords: data mining, terminology, ontology engineering, Wikipedia

1 Introduction

The parallel working methods and mutual interests of ontology engineers and terminologists have caused an important shift in the development of terminological resources. The notion of terminological knowledge base was introduced by [1] to denote a type of terminological resource that provides the means to explicitly encode ontological information. The creation of a terminological knowledge base or termontological database [2] involves studying terms as they are used in texts and discovering the relationships that exist between them. It has been shown how ontological and linguistic information extracted from specialised texts can be reorganised and presented in different ways when constructing specialised dictionaries [3], terminological knowledge bases [4], thesauri [5] or ontologies [6–8].

In this article we present the Termontospider tool, a wiki crawler that traverses Wikipedia in search of domain-specific texts for extracting relevant terminological and ontological information. The crawler is part of a tool suite for automatically developing multilingual termontological databases. The focus is on analyzing the best value for internal links, categories and other metadata to assign weights and search mechanisms in network traversal.

Wikipedia has been used as a primary text mining source for many years now. Section 2 briefly summarises some related studies with respect to mining Wikipedia for terminology and knowledge engineering purposes. Section 3 provides a general description of the Termontospider tool. Section 4 describes an experiment that is carried out with the Termontospider in the framework of a research project that aims at de-

veloping a multilingual termontological resource of cultural events. Finally, section 5 summarises the work ahead.

2 Related work

Wikipedia has been exploited for the identification of definitions [9] or for the recognition and semantic disambiguation of Named Entities [10]. It has been used to determine semantic relatedness between words in different languages by exploiting the inter-language links available between Wikipedia versions in multiple languages [11]. It is also used for testing several automatic extractors of semantic relations [12–14]. [15] explored the use of graph structures based on Wikipedia category sets as well as useful tags for determining key semantic relations. [16] worked on strategies for building graphs out of Wikipedia entries to identify semantic relatedness. [17] showed how category labels found in Wikipedia tend to agree with labels produced by humans for word clusters in ways that seem better than labels produced by "purely statistical methods".

3 Termontospider

The Termontospider is a software module that is part of a workbench for terminological and ontological mining (Termontominer). The crawler aims at automating the search and extraction of domain-specific terms and relationships by exploring Wikipedia. Departing from a very limited set of seed Wikipedia entries carefully selected by a domain expert, the tool analyses Wikipedia entries to assign a relevance for the domain at hand. After validation, these entries are then used as input by other modules in the Termontominer workbench for extracting terminological and ontological data and for representing these data in a (multilingual) termontological database. The Termontospider traverses Wikipedia by selecting links based on several parameters we have identified as useful for determining relevance and stops its search once a certain amount of features is no longer valid. This will be further explained below. Figure 1 shows an example of some of the items that need to be taken into account for a proper Wikipedia traversal in search of domain-specific data. Wikipedia entries are represented as rectangles. The rectangles in red (*Belgian Comic Strip Center* and *Cantillon Brewery*) are the manually selected seed entries relevant to events in Brussels. The rectangle in yellow (*Brussels International Fantastic Film Festival*) is one of the entries the crawler should identify as relevant to this domain whereas the rectangles in white are entries the crawler should identify as irrelevant or less relevant. The ellipses stand for Wikipedia categories. Categories, in spite of inconsistencies, provide relevant clues with respect to the semantic relatedness between Wikipedia articles [15]. The Termontospider should be able to infer that the categories in grey are more relevant than those in white to determine the relevance of each Wikipedia entry to the domain of cultural events in Brussels. Knowledge about the relevance of each category is obtained by measuring the connectivity between them and the entries.

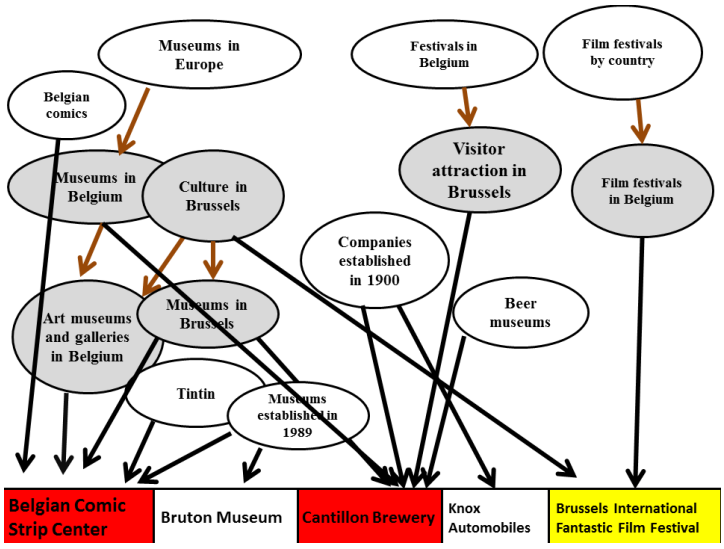

Fig. 1. Example of Wikipedia structures pointing at domain-relevant entries - in red and yellow. Most relevant categories are shown in grey.

Given the seed entries Belgian Comic Strip Center and Cantillon Brewery, we the crawler needs to look for other entries that may belong to the same domain, entries such as Brussels International Fantastic Film Festival, but not the Bruton Museum, which is in Britain, or Knox Automobiles, which was a car manufacturer set up at the same time as the Cantillon Brewery.

To this end, our crawler first calculates the tf*idf weight within the seed entries – with reference statistics based on a general corpus for the language we are using – and takes into account both metadata and rules for phrase construction to determine possible multi-word units. The most relevant lexical units of every new visited entry will be compared to a vector of the most relevant units in the seed entries. The system then records the categories assigned to the seeds and identifies all subcategories for the seed categories and recursively visits subcategories and all entries tagged with these subcategories plus upper categories. Next, the system visits entries directly hanging from upper categories and subcategories of the seed categories and verifies closeness based on comparing top terms. After that, the system visits entries directly departing from the seed entries. It verifies then whether those new entries link back. It assigns a higher relevant to those new entries with common categories to the seed entries or with categories that are subcategories for the seed entries. Categories that seem to have a one-to-one mapping in the other languages are given a higher weight. Categories that link to entries with no common top lexical units to the lexical units found in the seed entries get a lower weight. The system stops as default after a path of three nodes has been done from the seed entries.

The software also keeps track of the connectivity between entries of different languages to detect possible differences in conceptualization.

4 Implementation in a research setting

The Termontospider is currently being tested in the research project *Open Semantic Cloud for Brussels* (OSCB). This project aims to implement a framework of structured and interlinked information elements (so-called *Linked Data*) produced by "atomizing" a collection of databases and other resources that interoperate with each other to provide in a unified fashion information on the Brussels-Capital Region. Linguistic, semantic and visual information are processed to deliver requests for different users in three languages. In order to demonstrate the advantages and potential uses of such a framework, a series of use cases have been worked out for several domains. One such use case is a test application that retrieves information on cultural events in Brussels from several linked resources starting from natural language queries in either English, French or Dutch. To achieve this purpose, the application should map the natural language query to a semantic, language independent query that a semantic reasoner can use. The passage from language to semantic queries requires the use of a multilingual database connecting to an ontology. As initial material for building the linguistic and semantic databases we used a) textual data found in databases and sites of cultural organizations in Brussels, b) Wikipedia entries and c) possible utterances, queries that users may formulate in order to request information related to cultural events. The purpose of the Termontospider is to automate as much as possible the selection of documents in the three languages and discover issues regarding conceptualization based on Wikipedia's structure.

We initially selected a set of 10 seed words that have corresponding unambiguous entries in Wikipedia. We assume that a human expert should be able to select that number of entries representing typical items of the domain (s)he wants to analyze. The entries should represent instances or concepts of different types: devices, institutions, processes. We also selected a set of 100 further entries – our initial control group - that were manually identified as containing terminological – linguistic – and ontological material relevant for the OSCB use case with a set of 400 entries that are at most 2 links away from the initial seed entries but that are not relevant or are only marginally relevant. We are running Termontospider on the seeds and checking the percentage of control entries selected. We are subsequently calibrating the weights assigned for traversing to lower or upper categories. Initial results show that 5 to 10 seed entries can be enough to identify a large majority of relevant entries for these domains. The most reliable parameters to measure costs on new entries are common categories between the seeds, amount of paths between those categories and distance from source (seed) entry.

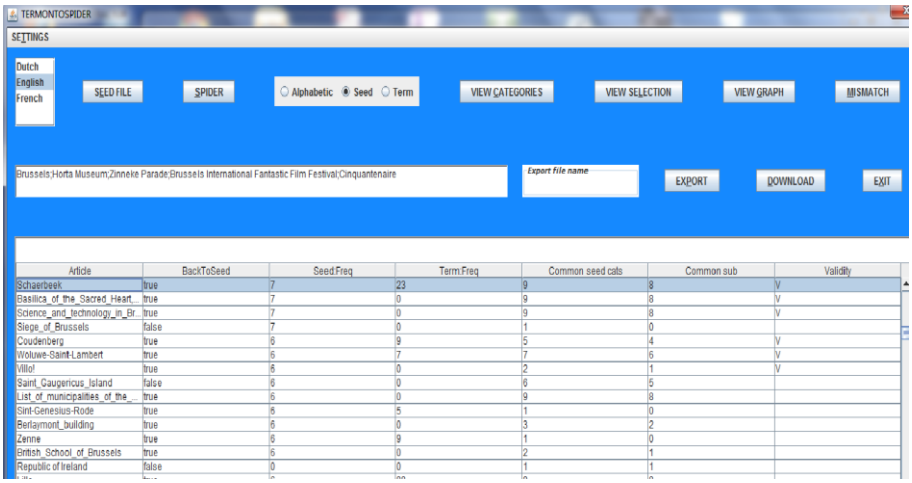

Fig. 2. A view of the results when user enters 5 seed entries for the English Wikipedia. Different parameters are shown that serve to determine whether the entries are closer to the wanted domain.

An indicator of the stability of Wikipedia concepts is the connectivity between entries across languages and how similar the categories are. Wikipedia projects often use robots that automatically link back across languages; connections between Wikipedia entries can then be seen as complete graphs whereby each node is connected to every other node once. Still, there are cases when this does not happen. In other cases, there seems to be a one-to-one mapping but intra-Wikipedia links not necessarily marked with disambiguation markers do contain related entries/concepts in one language alone.

Figure 3 shows different conceptualizations for Dutch (in yellow), English (in blue) and French (in red). "Ancienne Belgique" – a cultural building and organisation in Brussels- receives a separate entry as institution in Dutch alone. The metadata linking the entries is per se not enough to infer what kind of relationship there is between Ancienne Belgique as building or organisation. The second case shows how the competition "the Greatest Belgian" is conceptualized different and this is explicit through the inter-lingual – translation- links.

In order to determine the translation equivalence and the conceptual overlap between terms in the three languages, the Termontospider currently checks whether the language correspondences for a given entry can form a complete graph. That would require that if an initial entry 'X' in language D has translation entries Y and Z for languages E and F respectively, Y points back to X and Z, and Z points back to X and Y. This verification procedure may also reveal differences in semantic interpretations. The tool allows users to verify possible conceptual mismatches between terms in the different languages based not on the translation connections but on non-official markers such as "See also". Consider, for instance, the example of *Ancienne Belgique*, which can refer to the building *Ancienne Belgique*, for which we found Wikipedia

entries in the three languages, or the institution *Ancienne Belgique*, for which we currently an article in Dutch alone and not in the other languages. Possibly related concepts (such as *Ancienne Belgique* as institution or as building) are automatically retrieved. The possible relationship between a given entry and an entry linked to it under the "See also" section may very well have a much weaker connection, as in "Siege of Brussels" in the article about Brussels. How relevant the "Siege of Brussels" is for our crawling process has, obviously, more to do with what that entry has in common to the seed entries and their categories than to the article "Brussels".

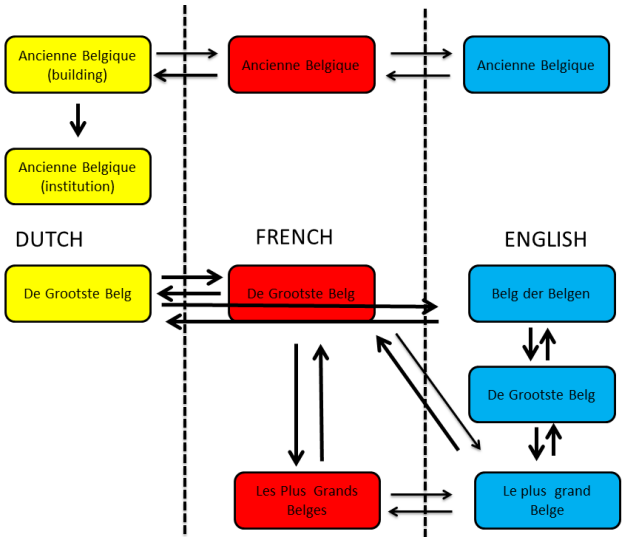

Fig. 3. The system also checks whether translation equivalences get mirrored across all languages compared. In the case of "de Grootste Belg" (the greatest Belgian) a difference can be discovered in this way. In the case of "Ancienne Belgique" the greater specificity in Dutch is harder to detect, it is implicit in a reference link under "See also".

5 Conclusions and work ahead

In this article, we have presented the Termontospider, a wiki crawler that optimally traverses Wikipedia in search of multilingual domain-specific texts. The tool is implemented as an independent module that is part of a workbench supporting automatic terminology and knowledge engineering processes. We have described the basic principles behind the crawler and summarized the research setting in which the tool is currently tested.

The experiment carried out for the cultural events use case shows that Wikipedia categories and their connectivity can effectively be used for directing the traversal of a crawler in search of Wikipedia texts and metadata to mine. The approach can be used for the three languages we are working with and can in principle be applied to the other Wikipedia languages. More testing and calibration still needs to be carried out to determine the most stable, scalable approaches for selecting categories. One of

the steps ahead is to try to link general Wikipedia categories to concepts within an ontology and use a semantic reasoner to influence the way in which categories can be considered relevant. For instance: if a seed entry is somehow connected to a high level Wikipedia category linked to the concept of location and if it is also linked to a new entry that has attached to itself a category that is also connected to location, the system would need to verify where the implied locations are mutually exclusive. If two of the seed entries share the same location and the new entry does not seem to refer to it, it is likely to belong to another domain. Assigning or linking Wikipedia categories to possible semantic facts or axioms for a semantic reasoner means – at least at this stage - some manual work. For scalability reasons, we need to keep to a minimum the amount of semantic axioms needed for the reasoned, in particular if are going to use this mechanism for crawling in any Wikipedia language project.

We are also working now on methods to utilize more strongly the hints given by categories in one language of Wikipedia across other languages. Works such as those by [18] have shown ways to increase the retrieval of translation candidates from Wikipedia. The authors used "quasi-morphological approaches" (deletion, addition of endings) to identify more candidate "translations" – in their case for terms. These approaches can be extended to categories. In order to obtain more (relevant) results to link terms and categories across languages, we will add proper stemmers to the Termontospider to verify possible correspondences and get hints from categories in other languages for entries not directly connected.

Finally, we are adding indexation capabilities for the selective crawling to verify content relevance by partially examining offline material only. This becomes particularly relevant when the domain to be mined is going through particularly large entries with many entries. We believe the software should optimally switch between offline checkups and run-time crawling depending on different parameters – estimated degree of relatedness based on sense similarity, links back and so on. A systematic analysis of the best criteria for this switching needs to be carried out.

Acknowledgements. This research is carried out in the framework of the Open Semantic Cloud for Brussels project (http://www.oscb.be), which is financed by the Brussels Institute for Research and Innovation (Innoviris).

6 References

1. Meyer, I., Skuce, D., Bowker, L., Eck, K.: Towards a new generation of terminological resources: an experiment in building a terminological knowledge base. Proceedings of the 14th conference on Computational linguistics - Volume 3. pp. 956–960. Association for Computational Linguistics, Stroudsburg, PA, USA (1992).
2. Kerremans, K., Desmeytere, I., Temmerman, R., Wille, P.: Application-oriented terminography in financial forensics. Terminology. 11, 83–106 (2005).
3. Leonardi, N.: Knowledge organisation in LSP texts and dictionaries: A case study. LSP Journal. 1, 81–98 (2010).
4. Tercedor Sánchez, M.I., López Rodríguez, C.I.: Integrating corpus data in dynamic knowledge bases: The Puertoterm project. Terminology. 14, 159–182 (2008).

5. Rychlý, P., Kilgarriff, A.: An efficient algorithm for building a distributional thesaurus (and other Sketch Engine developments). Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions. pp. 41–44. Association for Computational Linguistics, Prague, Czech Republic (2007).
6. Gillam, L., Tariq, M., Ahmad, K.: Terminology and the construction of ontology. Terminology. 11, 55–81 (2005).
7. Buitelaar, P., Cimiano, P.: Ontology learning and population: bridging the gap between text and knowledge. IOS Press (2008).
8. Poon, H., Domingos, P.: Unsupervised Ontology Induction from Text. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. pp. 296–305. Association for Computational Linguistics, Uppsala, Sweden (2010).
9. Navigli, R., Velardi, P.: Learning Word-Class Lattices for Definition and Hypernym Extraction. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. pp. 1318–1327. Association for Computational Linguistics, Uppsala, Sweden (2010).
10. Richman, A.E., Schone, P.: Mining Wiki resources for multilingual named entity recognition. Presented at the ACL-08: HLT. 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies , Columbus (2008).
11. Hassan, S., Mihalcea, R.: Cross-lingual Semantic Relatedness Using Encyclopedic Knowledge. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. pp. 1192–1201. Association for Computational Linguistics, Singapore (2009).
12. Barbu, E., Poesio, M.: Unsupervised Knowledge Extraction for Taxonomies of Concepts from Wikipedia. Proceedings of the International Conference RANLP-2009. pp. 28–32. Association for Computational Linguistics, Borovets, Bulgaria (2009).
13. Yamada, I., Torisawa, K., Kazama, J., Kuroda, K., Murata, M., De Saeger, S., Bond, F., Sumida, A.: Hypernym Discovery Based on Distributional Similarity and Hierarchical Structures. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. pp. 929–937. Association for Computational Linguistics, Singapore (2009).
14. Wu, F., Weld, D.S.: Open Information Extraction Using Wikipedia. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. pp. 118–127. Association for Computational Linguistics, Uppsala, Sweden (2010).
15. Milne, D., Medelyan, O., Witten, I.H.: Mining Domain-Specific Thesauri from Wikipedia: A Case Study. Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence. pp. 442–448. IEEE Computer Society, Washington, DC, USA (2006).
16. Yeh, E., Ramage, D., Manning, C.D., Agirre, E., Soroa, A.: WikiWalk: random walks on Wikipedia for semantic relatedness. Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing. pp. 41–49. Association for Computational Linguistics, Stroudsburg, PA, USA (2009).
17. Carmel, D., Roitman, H., Zwerdling, N.: Enhancing cluster labeling using wikipedia. Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval. pp. 139–146. ACM, New York, NY, USA (2009).
18. Niehues, J., Waibel, A.: Using Wikipedia to Translate Domain-specific Terms in SMT. Proceedings of the International Workshop on Spoken Language Translation 2011. pp. 230–237., San Francisco (2011).

Cross-Cultural Concept Mapping of Standardized Datasets

Fumiko Kano Glückstad

Copenhagen Business School, Dept. of International Business Communication
Dalgas Have 15, DK-2000 Frederiksberg, Denmark
fkg.ibc@cbs.dk

Keywords: feature-based similarity, cross-cultural communication, multilinguality, Bayesian model of generalization, categorization

Abstract. This work compares four feature-based similarity measures derived from cognitive sciences. The purpose of the comparative analysis is to verify the potentially most effective model that can be applied for mapping *independent ontologies* in a *culturally influenced domain* [1]. Here, datasets based on standardized pre-defined feature dimensions and values, which are obtainable from the UNESCO Institute for Statistics (UIS) have been used for the comparative analysis of the similarity measures.

1 Introduction

The recent internet revolution and its globalization impact has brought about new possibilities for people located at opposite sides of the globe to real-time dynamically communicate with each other. Although we most often use English as a common communication code, misunderstandings are almost unavoidable in such cross-cultural communications. This implies that multilinguality is a highly increasing demand that can correctly link concepts existing in diverse socio-cultural communities. This work challenges these multilingual issues based on the following pragmatic- and cognitive theories: the Relevance Theory of Communication [2] and the Knowledge Effects involved in category-based inductions [3]. A key point in these models is that a symmetric choice of code and context is not plausible in a cross-cultural communication scenario because the two communicating parties are unlikely to share an identical cognitive environment [2]. If e.g. a new object existing in a Source Language (SL) culture is introduced to a person in a Target Language (TL) culture, the TL reader will compare this new object with something he/she knows in advance (prior knowledge). This implies that feature-based asymmetric similarity measures play a key role for the communicating human cognitive mind.

In the ontology research domain, ref [1] compares several multilingual ontology frameworks such as the KYOTO project [4] and the MONNET project [5] based on a number of dimensions used in categorizing different types of ontology localization projects [6]. These dimensions are: *International (standardized) vs. culturally influ-*

enced domains; functional vs. documental localization; and interoperable vs. independent ontology. In this paper, potentially applicable asymmetric similarity measures that can be used for mapping *independent ontologies* in a *culturally influenced domain* are compared based on qualitative analyses. To increase the objectivity of the comparative analysis of the four different feature-based similarity measures, datasets based on standardized pre-defined feature dimensions and values, which are obtainable from the UNESCO Institute for Statistics (UIS) have been employed.

In the following, Section 2 describes the experimental settings of this work followed by a summary of results in Section 3, and summarizing with concluding remarks in Section 4.

2 Experimental Settings

2.1 Datasets

Datasets used in this experiment has been obtained from UIS who collected data from UNESCO Member States on an individual basis. The purpose of collecting data, according to UIS is to *map the Member States' national education systems according to the International Classification of Education (ISCED)*. UIS aims for *Member States to report their data in an internationally comparative framework.* These datasets from all over the world are downloadable from UIS' web-site[1]. Here, Japanese and Danish datasets have been used for the analysis. Each dataset consists of educational terms defined by several pre-defined feature dimensions such as ISCED level, programme destination and orientation, starting age, cumulative duration of education, and entrance requirements. Most feature dimension values are pre-defined, i.e. for the programme destination dimension, values are pre-defined as [general | pre-vocational | vocational].

One of the challenges of using these datasets is how to map the numeric feature values of dimensions such as "starting age" and "cumulative duration of education." For example, in the Danish educational system, the starting age of upper secondary school is defined as "16-17 years old" and its cumulative years of education is "12-13 years". On the other hand, the Japanese educational system is a so called "single-track system" meaning that the starting age of upper secondary school is exactly defined as "15 years old" and its cumulative years of education is "12 years". To handle this difficulty in an objective and systematic manner, the following procedure has been implemented: **1)** If a feature value in one country is completely included in a feature value in the other country (e.g. a feature "6-12 y.o." in Japan is completely included in a feature "6-17 y.o." in Denmark), a term possessing the feature that includes the other feature (a term possessing "6-17 y.o.") should also possess "6-12 y.o.", and **2)** If two features from the respective countries are partly overlapping (e.g. "13-15 y.o." in Japan and "14-17 y.o." in Denmark), a dummy feature referring to the exact overlapping range (i.e. "14-15 y.o.") is created. In this example, a Japanese term that possesses "13-15 y.o." should also possess the dummy feature "14-15 y.o." In the same

[1] http://www.uis.unesco.org/education/ISCEDmappings/Pages/default.aspx

way, a Danish term that possesses "14-17 y.o." should also possess the dummy feature "14-15 y.o.".

In order to objectively assess feature-based similarity measures, simpler datasets that do not contain these ambiguous feature dimensions/values have been prepared as control data. It means that these simpler datasets only contain the standardized feature dimensions/values defined by UIS. Based on these, similarity scores are computed by applying the four feature-based similarity measures described in the following.

2.2 Similarity computation

In this work, the first three similarity algorithms defined below based on Tversky´s Ratio Model are considered as baseline algorithms [7]:

$$sim\,(y,x) = 1/[\,1 + \frac{\alpha * f(Y-X) + \beta * f(X-Y)}{f(Y \cap X)}] \quad (1)$$

Equation (1) computes the degree to which object *y* is similar to *x*, when objects *x* and *y*, respectively, consist of feature sets *X* and *Y*. In here, object *x* is considered as referent and object *y* as subject of comparison according to the definitions of [7]. In equation (1) f is considered as additive function and α and β as free parameters. ($Y \cap X$) represents common features present in both *Y* and *X*, (*Y*-*X*) denotes distinctive features existing in *Y* but not in *X*, and (*X*-*Y*) in *X* but not in *Y*. In [9], three algorithms were defined based on different parameter settings: **i)** α=1 and β=1: which corresponds to the Jaccard Similarity Coefficient representing a symmetric similarity relationship between objects *x* and *y*; **ii)** α=1 and β=0: which only computes distinctive features present in *Y*, not in *X*; and **iii)** α=0 and β=1: which only computes distinctive features present in *X*, not in *Y*.

Here, a referent object *x* should be defined as an SL concept and a subject object *y* that is to be compared with *x* should be defined as a TL concept according to [7]. This definition should be applied to all three algorithms defined above. Keeping this definition in mind, an additional key point is that Tenenbaum & Griffiths [8] argue that the third algorithm is formally corresponding to the following equation (2) of the Bayesian Model of Generalization (BMG), which computes *the conditional probability that y falls under C (Consequential region) given the observation of the example x* [8]. Here, the consequential region *C* indicates the categorical region to where a subject *y* belongs.

$$P\,(y \in C|x) = \; 1/[1 + \frac{\sum_{h:x \in h, y \notin h} p(h,x)}{\sum_{h:x,y \in h} p(h,x)}] \quad (2)$$

In equation (2), a hypothesized subset *h* is defined as the region where a concept belongs to *h*, if and only if, it possesses feature *k* [8]. It means that *y* is considered as a newly encountered object existing in the TL ontology that should be aligned to the referent ontology of the SL according to Tversky's definition [7].

P(h, x) = P(x|h)P(h) above represents the weight assigned to the consequential subset *h* in terms of the example *x*. Therefore, as the fourth similarity algorithm, the

weight $P(h, x)$ is specifically assigned to the third algorithm based on the strong sampling scheme defined in [8] as follows:

$$P(x|h) = \begin{cases} 1/|h| & if\ x \in h \\ 0 & otherwise \end{cases} \quad (3)$$

Here, $|h|$ indicates the size of the region h [8]. In this work, the number of objects possessing the k^{th} feature in the referent ontology is considered as the size of the region h. [8] explains that the prior $P(h)$ is not constrained in their analysis so that it can accommodate arbitrary flexibility across contexts. Hence in this work, $P(h) = 1$.

3 Results and Data Analysis

Concept ID	ISCED level	Programme destination (A/B/C)	Programme orientation (G/P/V)	Theoretical cumulative duration at ISCED 5	Position in the national degree / qualification structure (intermediate, first, second, etc...)	Position in the tertiary education structure (Bachelor-Master-PhD)	Minimum entrance requirement (ISCED level or other)	Theoretical starting age	Theoretical duration of the programme	Theoretical cumulative years of education at the end of the programme	Does the programme have a work based element? (Y/N)	Programme specifically designed for adults (Y/N)	Programme specifically designed for part-time attendance (Y/N)
Danish concepts													
D1	0		G					2-5 years	4 years		No	No	No
D2	0		G					5-6 years	1 year		No	No	No
D3	1		G					6-7 years	6 years	6 years	No	No	No
D4	2	A	G				1	12-13 years	3-4 years	9-10 years	No	No	No
D7	3	C	V				2A	16-30 years	3-5 years	14 years	Yes	No	No
D19	5	B		Short	1st		3A, 3C	18-50 years	0,5-4 years	13-15 years	No	Yes	Yes
D20	5	B		Short	1st		3A, 3C	20-30 years	2-3 years	14 years	No	No	No
D21	5	A		Medium	1st	Bache-lor	3A	18-50 years	2-4 year	13-15 years	No	Yes	Yes
D22	5	A		Medium	1st	Bache-lor	3A	20-30 years	3-5 years	16 years	Yes	No	No
D23	5	A		Medium	1st	Bache-lor	3A	20-30 years	3 years	15-16 years	No	No	No
Japanese Concepts													
J35	5	B		Short	Inter-mediate		3 ABC	18	2-3	14-15	No	No	No
J36	5	B		Short Medium	Inter-mediate		5B	20	1+	15+	No	No	No
J37	5	B		Short	Inter-mediate		3	18	2-3	14-15	No	No	Yes
J38	5	B		Short	Inter-mediate		3	18	2	14	No	No	No
J40	5	B					3	18	1+	13+	No	No	No
J41	5	A		Medium	1st	Bache-lor	3	18	4	16	No	No	No
J42	5	A		Long	1st	Bache-lor	3	18	6	18	No	No	No
J44	5	A		Long	Inter-mediate		5A 1st,M	22	1+	17+	No	No	No

Table 1. Example of original datasets obtained from UIS: feature structure of selected concepts. The shadowed columns are feature values that are considered only for *a* graphs in Figures 1-2

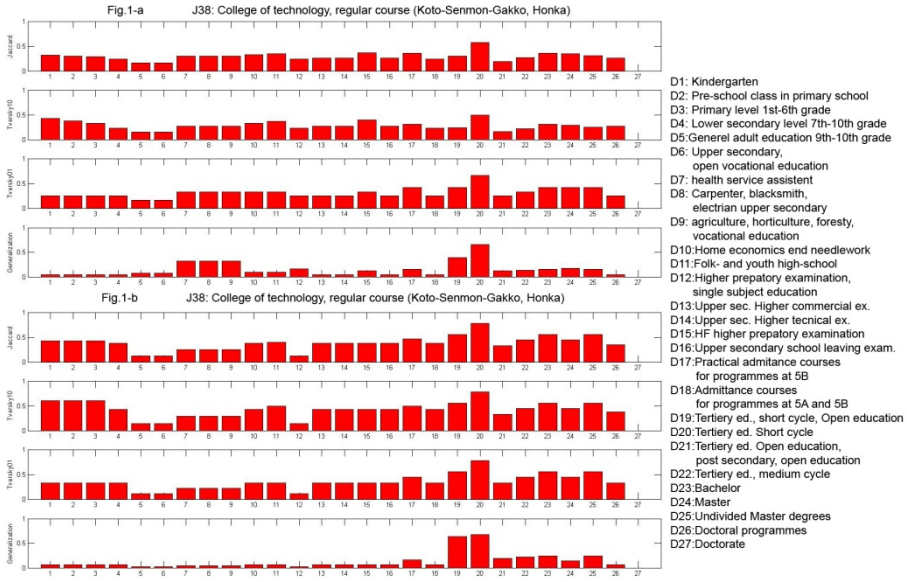

Fig. 1. Similarity scores: J38: college of technology as referent

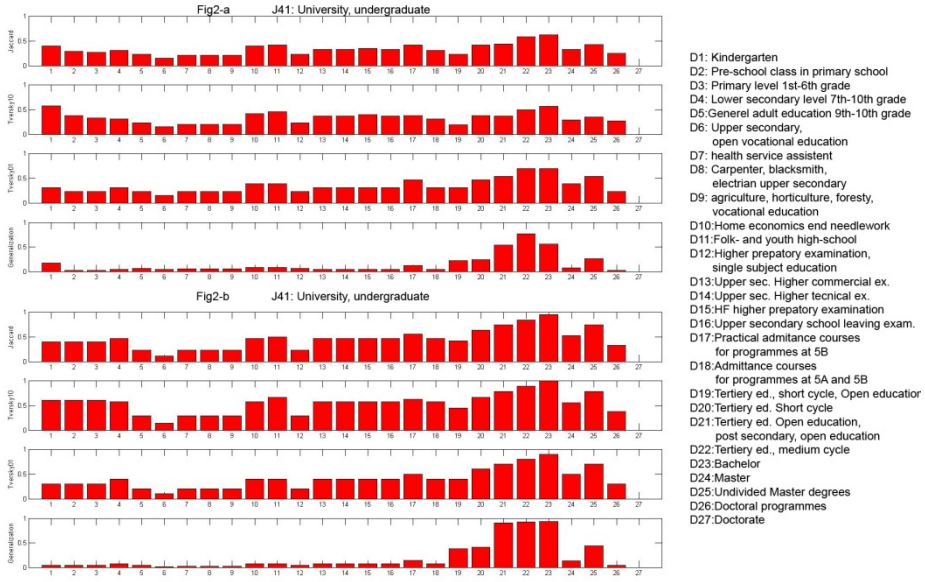

Fig. 2. Similarity scores: J41: University, undergraduate as referent

Although the datasets obtained from UIS have been developed for the purpose of statistical comparative analysis and mapping of the educational concepts among the Member States, no definite mapping pairs are proposed in a concrete form. This implies that the judgment of mapping depends on human evaluators in the respective

Member States countries. Consequently, the evaluation of results in this work focuses on a qualitative analysis, e.g. what kind of feature structures affect the results of similarity computation, instead of a quantitative analysis, e.g. recall-precision measures.

Figures 1 and 2, respectively, show similarity scores of the Japanese concepts, "J38: college of technology, regular course (高等専門学校本科 *Koto-Senmon-Gakko, Honka*)" and "J41: university, undergraduate (大学学部 *Daigaku Gakubu*)" against all accessible Danish concepts listed in the UIS dataset. The Japanese concept J38 is, from the author's own subjective point of view, an "atypical" concept compared to the more universally used concepts such as J41. While the upper part of the figures marked as *a* are the similarity scores computed based on feature dimensions/values including the numeric feature values described in Section 2, the lower part of the figures marked as *b* are computed without these feature dimensions/values.

The first thing to be noticed between the *a* and *b* graphs in general, is that the higher the number of feature values that are possessed by two concepts in question, i.e. in case of the *a* graphs, the lower the similarity scores. In particular, the first to third similarity scores in the *a* graphs show rather flat and ambiguous results. This is because the way the datasets have been created for mapping the feature values of dimensions such as "starting age" and "cumulative duration of education" simply increases the number of features. Among these, distinctive features will act as noise in the similarity computation, and hence the similarity scores decrease. In contrast to the first three similarity measures, the size principle in the fourth algorithm (BMG) effectively identifies specific concepts that are more similar than others, in all figures. For example for both J38, "D19: Tertiary, short cycle, open education" and "D20: Tertiary, short cycle education"; and for J41, "D21: Tertiary, post secondary open education", "D22: Tertiary, medium cycle education", and "D23: Bachelor" are respectively identified as the most similar concepts. On the other hand, the first to third similarity measures indicate that the aforementioned Danish concepts are only slightly more similar than the others. In addition, other Danish concepts referring to the pre-primary to lower secondary educations, i.e. D1-D4 are also considered slightly more similar than the others. Finally, the fourth similarity measure in Figure 1-*a* also identify that the Danish concepts referring to the vocational upper secondary educations, i.e. D7-9 are more similar than the others.

The results shown in Figures 1-2 indicate that the fourth similarity measure (BMG) seems to be the most effective algorithm. However, to conclude on this observation, it is necessary to investigate how the feature structures of each concept reflect the similarity computation. Table 1 shows the feature structures of selected concepts that are affected in the similarity results shown in Figures 1-2. Table 1 explains why the Danish concepts referring to the pre-primary to lower secondary educations, i.e. D1-D4 score higher with the first to third algorithms. There are two reasons for this. The first reason that apply especially for the first and second algorithms is that these algorithms consider distinctive features possessed by Danish concepts (*y*: subject to comparison), while the third and fourth algorithms consider ones possessed only by Japanese concepts (*x*: referent). Hence all feature values listed in the "programme orientation" column possessed by the Danish concepts strongly affect the similarity scores. The second reason is that the first to third algorithms equally consider all features that are

shared between two concepts in question based on additive functions. It means that for example all feature values with "no" that are matched between the two concepts are counted as "1". On the other hand, the BMG consider a feature value that is shared by many concepts as less important, which reduces similarity scores of all less relevant concepts such as pre-primary and primary education concepts. Another point is that the BMG detects that "J38: college of technology" is relatively similar to the Danish concepts referring to the vocational upper secondary educations, i.e. D7-9 in Figure 1-*a*. This is in fact true since the Japanese college of technology is a higher educational institution that is targeted for students who have graduated from lower secondary school and wish to acquire vocational skills based on 5 years education which consists of 3 years of upper secondary education and 2 years of vocationally oriented post-secondary education. The relevance between J38 and D7-9 has been effectively detected by balanced effects of feature values, i.e. feature value "14" of "cumulative duration" affects as decisive feature and other less important features reduce similarity scores of other irrelevant Danish upper secondary concepts.

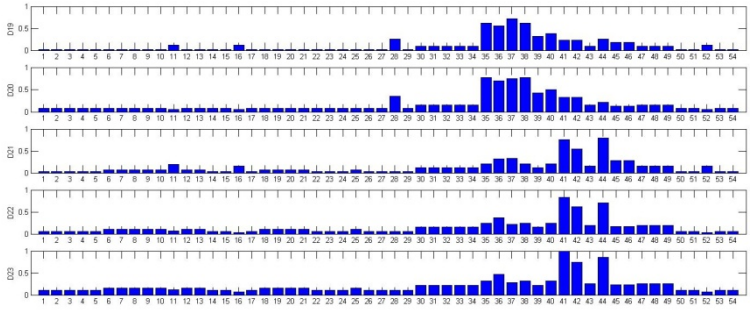

Fig. 3**.** Similarity scores: D19; D20; D21; D22; D23 as referent

Finally, equation (2) of the BMG theoretically explains that the model computes probabilities that a new object *y* falls under a hypothesized categorical region *C* provided that example *x* (prior knowledge) is observed. It means that by replacing variables *x* and *y*, it is possible to compute similarities from the Danish side, i.e. how a person who has prior knowledge of the Danish educational system selects the most similar Japanese concept as a feedback function. The results in Figure 3 show that Japanese concepts J35-37 referring to short cycle higher education provided at junior colleges and "J39: college of technology", are identified as the most similar concepts for Danish concepts "D19: Tertiary short cycle open education" and "D20: Tertiary short cycle education". In the same way, J41, J42 and J44, all of which are the Japanese bachelor degree programmes are detected as the most similar concepts for the Danish concepts "D21: Tertiary post-secondary open education", "D22: Tertiary medium cycle education" and "D23: Bachelor". These results demonstrate that, in these standardized datasets, uni-directional similarity relations from both the Japanese- and the Danish sides, are effectively computed. The feedback function of computing similarities from a Japanese or a Danish evaluator's viewpoint may be useful for detecting

asymmetric similarity relations, when mapping *independent ontologies* in a *culturally influenced domain* [1]. The theoretical argument of applying asymmetric similarity measures considering human prior knowledge is further discussed from a cognitive- and pragmatic point of view in [10].

4 Conclusions

In this work, four feature-based similarity measures are applied to the standardized datasets consisting of pre-defined feature dimensions/values developed by the UIS. The purpose of this comparison is to verify the similarity measures based on the objectively developed datasets. The results demonstrate that the BMG provides for the most effective cognitive model for identifying the most similar corresponding concepts existing for a targeted socio-cultural community.

Reference

1. Cimiano P., Montiel-Ponsoda E., Buitelaar P., Espinoza M., Gómez-Pérez A.: A Note on Ontology Localization. In: Journal of Applied Ontology Vol. 5, No. 2, IOS Press, (2010) 127-137.
2. Sperber, D., Wilson, D.: Relevance: Communication and Cognition. Blackwell, Oxford (1986)
3. Murphy,G. L.: The Big Book of Concepts. The MIT Press Cambridge, Massachusetts (2004)
4. Vossen P., Agirre E., Calzolari N., Fellbaum C., Hsieh S., Huang C.R., Isahara H., Kanzaki K., Marchetti A., Monachini M., Neri F., Raffaelli R., Rigau G., Tescon M., VanGent J.: KYOTO: A system for mining, structuring and distributing knowledge across languages and cultures. In: Proc. The 6th International Conference on Language Resources and Evaluation, Morocco, (2008) 1462-1469.
5. Declerck T., Krieger H.U., Thomas S.M., Buitelaar P., O'Riain S., Wunner T., Maguet G., McCrae J., Spohr D., Montiel-Ponsoda.E.: Ontology-based multilingual access to financial reports for sharing business knowledge across Europe. In: Rooz, J., Ivanyos, J. (Eds.) Internal Financial Control Assessment Applying Multilingual Ontology Framework, Budapest: HVG Press (2010) 67-76.
6. Espinoza, M., Montiel-Pansoda, E., Gómez-Pérez, A.: Ontology localization. In: Proc. The fifth International Conference on Knowledge Capture KCAP 09, ACM Press (2009) 33-40.
7. Tversky, A.: Features of similarity. Psychological Review, Vol., 84(4; 4), (1977) 327-352.
8. Tenenbaum, J. B., Griffiths, T. L.: Generalization, Similarity, and Bayesian Inference. Behavioral and Brain Sciences, Vol.24(4; 4) (2001) 629-640.
9. Glückstad, F.K.: Asymmetric Similarity and Cross-Cultural Communication Process. In: 9th International Conference on Terminology and Artificial Intelligence: Proceedings of the Conference. 8-10 November 2011, Paris, France. Paris : Institut National des Langues et Civilisations Orientales (2011) 59-65.
10. Glückstad F.K.: Bridging Remote Cultures: Influence of cultural prior-knowledge in cross cultural communication, In: Proc. The 26th Annual Conference of the Japanese Society for Artificial Intelligence: the Alan Turing Year Special Session on AI Research That Can Change The World (IOS-2), Yamaguchi, Japan (2012)

Acquisition, Representation, and Extension of Multilingual Labels of Financial Ontologies

Thierry Declerck, Hans-Ulrich Krieger, and Dagmar Gromann

DFKI GmbH, Language Technology Department,
Stuhlsatzenhausweg 3, D-66123 Saarbruecken, Germany
declerck@dfki.de,krieger@dfki.de
Vienna University of Economics and Business
Nordbergstrasse 15, 1090 Vienna, Austria
dgromann@wu.ac.at

Abstract. Globalization and a generally accelerated life style force companies to be flexible and ready to adapt to changes in the business environment. Integration of multilingual information as a process benefits from shared concepts of a multilingual ontology. We propose an automatic extraction of information from multilingual financial Web resources, which provide candidate terms for building ontology elements or instances of ontology concepts. Nevertheless, designations of ontology concepts need to be governed by sound terminological principles to facilitate further automation of ontology evolution as an example.

Keywords: harvesting multilingual terms, multilingual ontologies, ontology population, terminological principles

1 Introduction

Business organizations in the investment industry face a rapidly increasing need to be innovative and flexible in order to stay competitive. Thus, enterprises need to be constantly prepared to join new businesses and integrate existing systems. Semantic integration of information systems is a complex task requiring a vast variety of approaches such as conceptual modeling or requirements engineering. Ontologies can largely facilitate the integration process for multinational business partners by providing a company's information in a multilingual terminology associated to generally accepted concepts.

Ontologies are expressed in formal languages[1], describing mostly a conceptual hierarchy and associated relations[2]. Identification of concepts is done by agreed codes, which are typically not (well-formed) natural language expressions. But

[1] In the cases considered in this submission, we are dealing mainly with the languages RDF, SKOS, and OWL

[2] Depending on the framework, sometimes the words "property" or "role" are used for indicating the "relation" encoded in an ontology. We use the three words as synonyms in this submission.

modern knowledge representation languages foresee the use of annotation properties, such as `rdfs:label`, `rdfs:comment` or `skosxl:literalForm`, to include a human-readable designation of the concepts and roles in natural language.

While there is in principle no restriction to the kind of natural language expressions to be included in the labels of ontologies, there is an increasing agreement that terminological principles should be considered, for easing interpretation and translation of the content of the labels. In this context, many discussions have been pursued about the formal encoding of lexical and linguistic properties of natural language expressions used in labels[3]. Lexical and linguistic information considered for the generation of terminology compliant labels, which are thus a prerequisite of ontology engineering, might easily be lost in the final representation of the ontology, if no explicit interface between the different layers – labels and concepts – is provided, as has been argued in other works [1]. A model for such an interface for lexical information is described in [13]. This proposal for the representation of lexical information in ontologies has been designed for supporting ontology localization (see [12]) and multilingual ontology-based information extraction (see [6]).

In this submission we present a complementary approach to the direct localization/translation of ontology labels, by acquiring multilingual terminologies through the access and harvesting of multilingual Web presences of structured information providers in the field of finance, leading to both the detection of candidate terms in various multilingual sources in the financial domain that can be used as labels of ontology classes and properties but also for the possible generation of (multilingual) domain ontologies themselves. Only terms that can be transformed to validated concepts may be included in the schema of the ontology, whereas other lexical and terminological data are stored in corresponding resources, such as annotation properties.

2 Acquisition of Multilingual Terms as Building Blocks for Ontology Generation

We access different types of multilingual information related to companies – annual reports, stock exchanges, industry classification standards, etc. – from different sources. From these sources various term candidates are detected, which are to be federated. But another aspect is also worth considering: certain terms in the HTML structures of the harvested Web pages can be considered as supporting the derivation of ontology classes or properties, while other terms and contexts can be considered factual information, which can then be used for gathering individuals to populate the formerly generated schema of the ontologies.

[3] See for example the *LexInfo* Web page: http://lexinfo.net/, or the recent Linked Data in Linguistics (LDL) Workshop, http://ldl2012.lod2.eu/

2.1 Federated Financial Ontology

Crawled information was transformed into basic terminological and ontological resources, some of which were achieved semi-automatically by an extraction tool. The result was the construction of a federated ontology consisting of eleven sub-ontologies (see Figure 1).

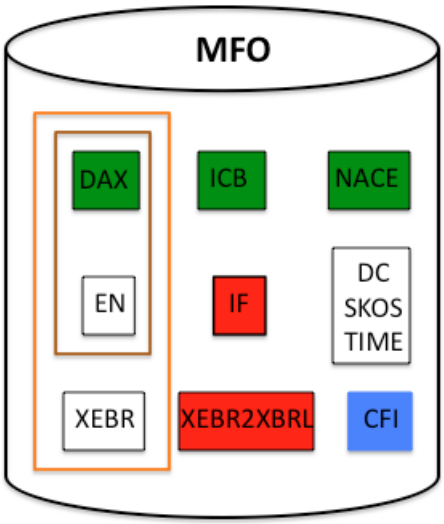

Fig. 1. The federated ontology **MFO** consists of overall 11 sub-ontologies. The color encoding refers to ontologies focusing on models of *industry sector classification* (**green**), *stock exchange* (**brown**), *reporting* (**orange**), *financial instruments* (**blue**), and *interface* (**red**). As can be seen from the picture, some of the ontologies even model several aspects of our domain; e.g. *DAX* alone deals with industry sector classification, reporting, and the description of stock exchange listed information.

Five ontologies deal with business reporting or industry classification standards as described thereafter. The federation incorporates an interface ontology (which interconnects the other ontologies), a financial instruments ontology, one ontology for temporal concepts and one for annotation properties, including SKOS elements, such as prefLabel, altLabel, and hiddenLabel.

The xEBR ontology is derived from a core taxonomy, called xEBR, which was developed for achieving comparability across national boundaries in the field of business reporting using the eXtensible Business Reporting Language (XBRL)[4]. XBRL is an XML-based language for the presentation of business information and business reports. XBRL-encoded reports describe company and financial information compliant with generally accepted accounting principles (GAAPs) and legal requirements as defined by distinct countries or legislations.

[4] http://www.xbrl.org/

The xEBR core taxonomy and the derived ontology describe conceptual links between concepts used in different GAAPs, which make use of labels in different languages.

The Deutscher Aktien IndeX (DAX) ontology is a transformation of the data crawled from the DAX company pages and the DAX sector classification, containing detailed definitions of the individual sectors, both being available in English and German. The Euronext ontology centers around the representation of companies found on the NYSE Euronext website, available in four languages: Dutch, English, French and Portuguese.

The ICB ontology is derived from the Industry Classification Benchmark (ICB)[5], which seeks to provide a global comparison of companies by industries and contains 114 subsectors represented in the ontology. The NACE ontology is based on the NACE nomenclature of the organization for industry classification, representing numerous industry sectors. Numbers describing the size of each subontology are presented in Figure 2.

	CFI	DAX	DC	EN	ICB	IF	NACE	SKOS	TIME	XEBR	XEBR to XBRL	MFO
#concepts	1	97	1	9	186	19	997	1	3	63	11	1366
#eq class axioms	--	--	--	--	--	4	--	--	--	--	7	11
#object properties	--	7	--	6	--	37	--	6	--	6	4	61
#eq obj prop axioms	--	--	--	--	--	17	--	--	--	--	2	19
#datatype properties	6	22	--	42	--	--	--	--	--	154	123	295
#eq data prop axioms	--	--	--	--	--		--	--	--	--	71	71
#individs	--	40	--	48	--	7	--	--	--	60	137	288
#annotat properties	--	1	4	--	1	--	2	3	--	--	--	11
#explicit triples (= #axioms)	28	744	11	278	1,370	81	6,983	31	10	1,164	783	11,483
size [KB]	4	111	4	41	229	12	967	8	4	180	127	1,687

Fig. 2. Statistics on the size of each sub-ontology in terms of the number of classes, properties, and axioms.

[5] http://www.icbenchmark.com/

2.2 Ontology population

A motivation for developing this federated ontology was the ability to provide a rich knowledge base supporting cross-lingual information access and presentation in the field of business reporting. The population of such a knowledge base results from the application of an ontology-based information extraction system to financial documents.

So, for example, an analyst can submit a Spanish XBRL instance document (XML documents containing only concepts and concrete values) to the system. The main concepts, i.e., those having a correspondence in the xEBR core taxonomy/ontology with their associated concrete values, are stored in a language independent way as instances of xEBR concepts and properties, while the language labels (in this case only Spanish), are accessible via the `rdfs:label` annotation property that have been adjoined to the concepts at the T-Box level of the ontology. The xEBR core ontology links concepts of other legislations to, for example, the Belgian legislation. From there, the corresponding labels can equally be linked and the result is a natural language representation in five languages for the Spanish XBRL concepts that have been extracted from a report and stored as an instance of a xEBR class or property of the xEBR ontology.

3 Linguistic Structures

Ontology labels generally consist of one designation per language, rendering the inherent terminological approach highly prescriptive. Our approach seeks to achieve a full representation of extracted terminological information in the corresponding terminological resource. If derivations and variants of existing labels of concepts as well as the financial standard-compliant version are available to the ontology engineer, a more complete view of the conceptual information can be achieved. Additionally, we seek to represent general linguistic patterns, i.e., mainly compound patterns, word patterns, collocation patterns, and syntactic patterns of existing designations to facilitate the construction process of new labels.

Compound analysis and PoS tagging represent initial steps towards a rich system of linguistic and terminological information. Lemmatization and complex morphological analysis will be described with the `lemon` model (see [13]). The information thereby obtained serves as a basis for constituency and dependency information, allowing for a comparison of labels by means of head nouns and modifiers, as well as a frequency analysis for the individual components.

On the basis of dependency analysis, head nouns of individual labels can be compared. Frequency lists on the extracted information confirmed a certain primacy of nouns in financial terminology. For example, English ICB labels contain 165 different nouns and only 13 adjectives. In comparison, German ICB labels consist of 181 distinct nouns and 19 adjectives. In terms of word patterns, nouns and adjectives belong to word classes more likely to change than conjunctions, prepositions, and so on.

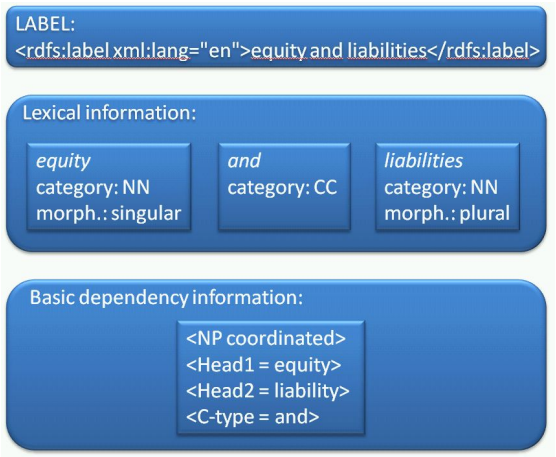

Fig. 3. Basic example of label analysis

Some lexical units are the basis for numerous different collocations, usually with strong syntactic variety. Within a specific domain these different senses can already be restricted, as for example the German *Unternehmen, die hauptsächlich im Bereich Rückversicherung tätig sind* might not refer to `reassuring`, but to `reinsurance`. Our approach to collocations is one of representing derivation patterns for each sub-ontology, referring to verbal, nominal, adjectival or adverbial derivations. For example, the DAX ontology refers to both *Industriegase* as well as *industrielle Gase*, whereas the second term is an adjectival derivation of the noun compound. However, the noun compound is used with a higher frequency in the designation of classification categories and thus, can be represented as the preferred linguistic structure of this particular resource. Nevertheless, the adjectival derivation will still be represented as a term variant in the terminological resource. Thus, derivation patterns help the ontology engineer to opt for the noun compound in this case and thereby improve the consistency of the prescriptive usage of terms in ontology labels.

The described approach to linguistic patterns is highly beneficial to other purposes as well, and having stated the similarity of the compound *Industriegase* and the simple NP consisting of a pre-modifying adjective and a head noun *industrielle Gase* supports the task of information extraction and the subsequent ontology population procedure, described in Section 2.2: both terms (variants) will be recognized as the results of the industry activity of a company.

4 Representing Multilingual Terminology in Labels

Sound terminological principles are inevitable for a consistent use of natural language terms in ontology labels, and thus, also for ontology engineering. Not only do principles foster re-usability, but they also represent a thorough guidance to-

wards term consistency and standard compliance. The following set of principles applies to the creation of a terminological resource for ontologies as well as to the automation of generating labels.

1. Some principles, such as the most important principle of concept orientation, may also apply to any type of terminological resource, however, most of them are targeted towards ontological interoperability.
2. Autonomy of ontology labels means that each designation may only exist once within one domain in the exact same wording, variation, and form. In this regard, each label might be considered autonomous unless otherwise specified as homonym in the terminological entry in a connected term base. The terminological entry in such a term base has to contain all terminological data related to one concept (ISO 1087), thereby changes of the concept can be easily propagated to dependent elements, which refers to the principle of concept orientation.
3. Compounds of simpler terms representing the same underlying reality as the simpler terms may not be included in the terminology as terms but rather as compound patterns. A compound pattern is a normalization of how to form new compound terms using the terms at hand, a highly valuable information to ontology engineers in the process of adding new concepts with new labels to an existing ontology. In line with compounds, term variants have to be examined as to their underlying concept. We will use the term "ontologically valid variants", introduced by Bodenreider et al. [3], for this purpose.
4. When it comes to multilingual terminologies, delimiting characteristics of terms might vary across languages. Nevertheless, certain principles can be generalized, such as the *avoidance of underspecification markers*. For instance, the automatic generation of the multilingual ICB ontology contains phrases such as *Specialty Finance*, defined as *financial companies engaged in financial activities not specified elsewhere*. The underspecification marker *specialty* and the corresponding definition render it impossible to classify a company as an instance of this class without taking the additional information of all other classifications elements into account.
5. The use of *standardized data categories* and *consistent terminology* contributes to the objective of automating localization. The German *Hersteller* manifests itself as *producer* and *manufacturer* among others in English. Any automation of ontology evolution largely profits from a consistent use of terminology also in the localization process.

A thorough and manual terminological analysis of the terms, which results in the construction of term bases in TBX and harmonization efforts across the different resources, provides the basis for achieving consistent ontology labels in natural language. Naturally, the financial domain provides ample numerical information, which can be used to evaluate terminology across languages. DAX information taken from the Xetra Web presence differ in terminology from the details on the company Web presence or corresponding facts on the Bundesanzeiger Web page. For instance, BASF refers to *Langfristiges Fremdkapital* on

its Website, as does the Bundesanzeiger for the same category of BASF facts. However, DAX uses *Langfristige Verbindlichkeiten*, while providing the same numerical value for the category as do BASF and Bundesanzeiger. The fact that both are localized to *Long term liabilities* in English and show the same numerical value is evidence enough to establish their equivalence.

In order to validate the strategy we initially analyzed the semantic relations of the terms before entering them as term variants into our term bases. German literature differentiates between *Verbindlichkeiten* (liabilities or debts), *Rückstellungen* (provisions), and *Rechnungsabgrenzungsposten* (Accruals and deferred income) as part of *Fremdkapital* (liabilities). Additionally, to ensure their equivalence the corresponding financial standard, i.e., IFRS, was consulted and is represented in the term base as well.

In cases where the source itself offers detailed definitions and thus a context, the consultation of numerous sources as in the example above is unnecessary. Each subsector of ICB, which represents the lowest level of the four-layered classification structure, comes with a detailed definition delimiting the category against siblings. The initial step was to analyze the classification terminology on the basis of the definitions and establish a term base. Nevertheless, we consulted profiles of companies classified therein and realized that the introduction of additional categories such as a combination of *specialty chemicals and* pharmaceuticals might be necessary. Furthermore, a frequency analysis of definitions as opposed to classification labels showed that several high-frequency terms are not used in the labels. Some of them, such as *company*, might be superfluous for the purpose of the industry classification system, however, gains importance for the ontological representation.

One approach to including term variants in a separate term base is the use of SKOS in combination with the ontology. The following example from the xEBR ontology represents one concept designation:

```
<skos:prefLabel xml:lang="en">Financial debts</skos:prefLabel>
<skos:altLabel xml:lang="en">Financial debts with a remaining
 term of more than one year</skos:altLabel>
<skos:prefLabel xml:lang="nl">Financiêle schulden</skos:prefLabel>
<skos:altLabel xml:lang="nl">Financiêle schulden op meer dan
  één jaar</skos:prefLabel>
<skos:altLabel xml:lang="fr">Dettes financières à plus
  d'un an</skos:altLabel>
<skos:prefLabel xml:lang="fr">Dettes financières</skos:prefLabel>
<skos:prefLabel xml:lang="de">Finanzverbindlichkeiten</skos:prefLabel>
<skos:altLabel xml:lang="de">Finanzverbindlichkeiten mit einer
  Restlaufzeit von mehr als einem Jahr</skos:altLabel>
```

Fig. 4. xEBR ontology labels to exemplify term variants using SKOS.

The major advantages of this approach are that the information is already encoded in RDF and can more easily be integrated with the OWL files of the ontology and no additional resources have to be used. The disadvantage is the prescriptive approach of SKOS towards terminology, as each label is always marked with a status marker `altLabel`,`prefLabel`, or `hiddenLabel`. In contrast, terms in a term base can optionally be designated with a status marker. Our current approach employs SKOS in order to harmonize terms across terminologies, as each ontology has its individual terminology. We also noted that the labels do not comply with some of the principles we have mentioned, as we think that *Financial debts with a remaining term of more than one year* should be considered a subclass of a concept bearing the label *Financial debts* due to the meaningful post nominal prepositional modification and not an alternative label.

5 Conclusion & Future Work

We have presented ongoing work dealing with the automated acquisition of multilingual candidate labels for ontology elements. We discussed issues related to the representation of such labels and how these can lead to the evolution of ontologies. As a methodology, our approach supports localization strategies by semi-automatically deriving substantial multilingual labels. The approach also supports multilingual ontology-based information extraction, since the extracted terminology and associated contexts are yielding both a list of potential natural language expressions corresponding to classes and properties, as well as values they can be associated with. Consequently, a kind of tailored multilingual domain specific lexicon and gazetteer is built. Each resource thereby obtained is modular and can easily be employed in other systems or for different purposes. We described the need to apply a manual terminological organization of the candidate terms extracted from the multilingual resources in detail, and proposed a combination of TBX and SKOS for encoding the terminology, supporting the inclusion or evolution of ontologies. Not only is the acquisition of multi-lingual labels worthwhile as the paper has shown, these labels can also be utilized to perform automatic ontology merging/alignment by using techniques developed in the *Recognizing Textual Entailment* task [5]. We are currently pursuing this road, using the industry classification from DAX, ICB, and NACE (see Section 2.1) as input data.

Acknowledgements

Part of this work has been supported by the Monnet project (Multilingual ONtologies for NETworked knowledge), co-funded by the European Commission with Grant No. 248458 and by the TrendMiner project, co-funded by the European Commission with Grant No. 287863.

References

1. Badra, F., Despres, S., Djedidi, R.: Ontology and Lexicon: The Missing Link. In: 9th International Conference on Terminology and Artificial Intelligence, Proceedings of the Workhsops, pp. 16-18. INALCO, France (2011)
2. Bizer, C., Heath, T., Berners-Lee T.: Linked data - the story so far. International Journal on Semantic Web and Information Systems (IJSWIS). 5:3, 1-22 (2009)
3. Bodenreider, O. Smith, B., Burgun, A.:The Ontology-Epistemology Divide: A Case Study in Medical Terminology. In: Varzi, A., Vieu, L. (eds.): Proceedings of FOIS 2004. International Conference on Formal Ontology and Information Systems, Turin (2004)
4. Bassey, A., Budin, G., Picht, H. Rogers, M., Schmitz, K.D., Wright, S.E.: Shaping Translation: A View from Terminology Research. Translators' Journal 50:4, 195–197 (2005)
5. Dagan, I., Glickman, O., Magnini, B.: The PASCAL Recognising Textual Entailment Challenge. In Quiñonero-Candela, J., Dagan, I., Magnini, B., d'Alché-Buc, F. (eds.): Machine Learning Challenges. Lecture Notes in Computer Science, Vol. 3944, 177-190, Springer, 2006.
6. Declerck, T., Lendvai, P. and Wunner, T.: Linguistic and Semantic Features of Textual Labels in Knowledge Representation Systems. Proccedings of the Sixth Joint ISO - ACL/SIGSEM Workshop on Interoperable Semantic Annotation 2011.
7. Federmann, C. and Hunsicker, S.: Stochastic Parse Tree Selection for an Existing RBMT System Sixth Workshop on Statistical Machine Translation (2011).
8. Federmann, C. and Hunsicker, S., Wolf, P., and Bernardi, U.: From Statistical Term Extraction to Hybrid Machine Translation. Proceedings of the 15th Annual Conference of the European Association for Machine Translation, Leuven, Belgium, EAMT (2011)
9. ISO 12620 (2009): Terminology and other language and content resources – Specification of data categories and management of a Data Category Registry for language resources, Geneva, ISO.
10. ISO 16642 (2003): Computer applications in terminology - Terminological markup framework, Geneva, ISO.
11. ISO 30042 (2008): Systems to manage terminology, knowledge, and content - TermBase eXchange (TBX), Geneva, ISO.
12. McCrae, J., Espinoza, M., Montiel-Ponsoda, E., Aguado-de-Cea, G. and Cimiano, P.: Combining statistical and semantic approaches to the translation of ontologies and taxonomies. Proceedings of the 5th Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-5) (2011).
13. McCrae, J., Spohr, D., Cimiano, P.: Linking Lexical Resources and Ontologies on the Semantic Web with Lemon. The Semantic Web: Research and Applications. Volume 6643 of Lecture Notes in Computer Science, 245-259. Springer, Berlin (2011)
14. Miles, A., Bechhofer, S.: SKOS-Simple Knowledge Organization System Reference, W3C Recommendation, 18 August (2009)
15. Montiel-Ponsoda, E., Aguado-de-Cea, G., McCrae, J.: Representing term variation in *lemon*. Extended Abstracts, 9th International Conference on Terminology and Artificial Intelligence, TIA 2011,47-50, Paris (November 2011)

Supporting collaboration in multilingual ontology specification: the conceptME approach

Manuel Silva[1,2], António Lucas Soares[2], Rute Costa[3],

[1] ISCAP-IPP - Rua Jaime Lopes Amorim, s/n, 4465-004 S. Mamede de Infesta
mdasilva@iscap.ipp.pt
[2]INESC Porto, Rua Dr. Roberto Frias, s/n 4200, Porto-Portugal
als@fe.up.pt
[3]CLUNL - Universidade Nova de Lisboa, Avenida de Berna 26-C 1069 - 61 Lisboa;
rutecosta@fcsh.unl.pt

Abstract. Despite the availability of tools, resources and techniques aimed at the construction of ontological artifacts, developing a shared conceptualization of a given reality still raises questions about the principles and methods that support the initial phases of conceptualization. These questions become more complex when the conceptualization occurs in a multilingual setting. To tackle these issues a collaborative platform – conceptME - was developed where terminological and knowledge representation processes support domain experts throughout a conceptualization framework, allowing the inclusion of multilingual data to promote knowledge sharing and enhance conceptualization.

Keywords: Multilingual ontology specification, Localization, Terminology, Collaborative networks, Knowledge Representation.

1 Introduction

The development of the diverse scientific and technical fields has its origin in the evolution and dynamics of knowledge and results from the constant interaction between individuals pursuing common objectives, knowledge that cannot be separated from its context, experience, culture and language. This interaction, especially in multinational and multicultural organizations, is increasingly taking place in collaborative and cooperative environments available online.

In these environments, language, as the means of human communication, and terminology, as a nuclear element for the specification and dissemination of specialized knowledge, assume an increasingly important mediation role in the communication taking place between the various interlocutors and in man-machine communication, emerging as the key link for the discovery and creation of knowledge and its effective conceptualization, representation, transmission and reuse.

To meet the increasing demands of the complex intra and inter-organizational processes, there was a growth in quality in the processes of interaction and sharing of resources inside organizations, on the one hand, through the implementation of inno-

vative forms of collaboration, such as collaborative networks, defined by [4] as *a network composed of a variety of entities - organizations and people - which are largely autonomous, geographically distributed, and heterogeneous in terms of their operating environment, culture, social capital and goals, where participants collaborate to (better) achieve common or compatible goals, being their interactions supported by computer network* and, on the other hand, the development of more robust information and knowledge management systems, such as ontology-based knowledge management systems.

Knowledge organization and collaboration systems are thus important instruments for the success of collaborative networks of organizations. In this context, access to and representation of knowledge implies the overcoming of difficulties inherent to the use of different natural languages, through the use of processes and methodologies that support and promote knowledge sharing and organization in multilingual settings.

2 Terminology and knowledge representation

As stated in [25], an increasing number of semantic tools and resources such as concept map editors or wiki-based platforms have been built with the goal of sharing information and knowledge in collaborative networks. Despite the availability of techniques aimed at the construction of ontological artifacts, developing a shared conceptualization of a given reality still raises questions about the principles and methods that support the collaboration process. [16] underline limitations in the development of ontologies in collaborative settings: «current knowledge about the early phases of ontology construction is insufficient to support methods and techniques for a collaborative construction of a conceptualization». Techniques may involve the (re)use of ontology design patterns (ODP), which is not without its challenges: «even users with some background on ontology modelling face difficulties when reusing ODPs for their needs» [28]. These limitations grow bigger when the setting is multilingual and the ontology has to be specified in more than one natural language.

In the light of this issue, and as [25] make clear, *tasks involving conceptualization call for interplay between terminology and knowledge representation capable of rendering intuitive and operational the notions of term and concept without blurring the theoretical distinction between the different levels of analysis triggered by them*. Practical work such as representing knowledge for ontology-building purposes tends to show them as alternate (sometimes opposing) sides rather than interdependent elements of a relation between objects, concepts and terms, as it is represented in the semiotic triangle in terminological science and research.

Under the scope of the project – CogniNET[1] – a prototype of a collaborative tool – conceptME - is being developed to implement functionalities and models that will assist experts in the process of reaching a shared conceptualization of a given domain, in the form of semi-formal ontologies, based on this interplay between terminology and knowledge representation. In this article we describe the preliminary steps of conceptME approach to conceptualization in a multilingual environment, which in-

[1] http://cogninet.tk

tends to assist experts in the discussion and modelling of the concepts of their domain in a multilingual setting.

2.1 Difficulties in multilingual ontology specification

As identified in [12], current approaches to cross-lingual information access offer only partial solutions that address the problem in a restricted way. The scarcity of formal ontologies enriched with linguistic information in more than one language has its origin in other factors such as the difficulty in choosing methodologies to support the knowledge conceptualization and representation process in an environment of construction and localization of ontologies for different languages. Although localization is a well-developed practice and its methodologies and tools have been successfully employed by the language industry in the development and adaptation of multilingual content, it has not yet been sufficiently explored as an element of support for the development of ontologies represented in more than one language.

[9] identify several problematic dimensions to be taken into account in the process of ontologies localization, namely translation problems, related to the existence (1) *of exact equivalents, (2) context-dependent equivalents and (3) of conceptualization mismatches; management problems, related to maintenance and updating of translated ontology labels throughout the ontology life cycle; and multilinguality representation problems*. In fact, part of the difficulties of any localization system lies in solving problems that we can view as traditional and which result from the translation process, such as the difficulty in finding equivalents in the target language, the existence of polysemic terms and quasi-synonyms, or problems related to terminological variation.

Other problems derive mainly from linguistic problems that arise from the association of meanings of terms in different languages to concepts represented in an ontology, as *word senses and concepts cannot be said to overlap* [10] since, as recognized by [13] *word senses are tightly related to the particular vision of a language and its culture*, whereas concepts represented in an ontology refer to *objects of the real world and are defined and organized according to expert criteria agreed on by consensus*.

As [21] acknowledges, it is generally accepted that achieving a one-to-one term-concept and concept-term relationship (*Eineindeutigkeit*) within a subject field is unattainable. [21] recalls that Wüster himself *had practical doubts about the viability of achieving this goal on a comprehensive scale, and described it as "ein frommer Wunsch"* [24]. [19] on the other hand, says we *cannot communicate and share information unless we agree on the terms we use and on their meaning*. For the author, the meaning of terms rests upon a shared and consensual representation of a domain model and it is such representation that originates an ontology.

In addition to these difficulties, localizing an ontology - understood as a specific semantic artefact used to represent the knowledge of a domain, built in a given context for a particular purpose -, raises other questions, like those related to the:

1. definition and delimitation of the domain or subdomain(s) to be conceptualised;
2. selection, adaptation and integration of existing semantic resources;
3. time constraints, usually imposed on processes of conceptualization and localization;

4. approach to integration and (re)use of already available language resources and tools.

3 Approach to Multilingual Ontology Specification

The more generic goal of ontology localization is to allow *cross-lingual semantic interoperability in large-scale information environments, which usually contain a number of heterogeneous and distributed knowledge resources* [1]. The specification of an approach by which localization may contribute to enhance the cross-lingual semantic interoperability between heterogeneous resources of a specific subject field requires taking into account and acting upon the context of the ontology construction and knowledge sharing during the ontology conceptualization phase.

It also requires that we consider the objectives and purposes the community of potential users may have for this knowledge. To do so we need to focus on apprehending the subject fields' complexity, richness and semantic diversity and, at the same time, on having a method and tool to help represent its multilinguality, what should also happen during conceptualization.

The conceptualization phase of an ontology development process is of utmost importance for the success of the ontology, as it is in this phase that a socio-semantic agreement is shaped [17]. For [22] a conceptualization process is, for an individual, a *collection of ordered cognitive activities that has as inputs information and knowledge internally or externally accessible to the individual, and as the output an internal or external conceptual representation*, and a "collaborative conceptualisation process" is *a conceptualization process that involves more than one individual producing an agreed conceptual representation*, a process which involves social activities that include the negotiation of meaning and practical management activities for the collaborative process.

For [17], *ontology engineering needs a "socio-cognitive turn"* in order to generate tools that are really effective *in coping the complex, unstructured, and highly situational contexts that characterize a great deal of information and knowledge sharing*. [17] remember [3] words when he says that *we need to go beyond the approaches that provide a high level of 'automation of the meaning'; instead, we need to address situations where human beings are highly required to stay in the process, interacting during the whole life-cycle of applications, for cognitive and cooperative reasons*. The authors place conceptualization in a phase of informal specification of the ontology (previous to any formal representation) and describe its result as a shared conceptual model.

The aim is *to support the co-construction of semantic artefacts by groups of social actors placed in organizational contexts interacting towards a set of common objectives* [16]. This co-construction and the resulting conceptual representations, which are based on the analysis of different sources, including textual, terminological, taxonomic and other, and subject to constant negotiation with the direct collaboration of domain experts, could, in our opinion, assume a multilingual dimension as early as the conceptualization phase.

3.1 conceptME conceptualization framework

Based on this view and on the analysis of the process of a shared conceptualization of domain ontologies in the context of a collaborative network, we have developed a platform – conceptME - to support the process of multilingual specification of an ontology to be implemented during the conceptualization phase. For the development of our proposal we assume that the processes of conceptualization and localization of an ontology may occur consecutively, in order to allow us to consider all available information and perspectives of the different working languages and cultures as early as the conceptualization phase. The proposed iterative and, to some extent, cyclical nature of the two processes - conceptualization and localization – intends, thus, to promote more immediate access to different perspectives about the domain's knowledge.

The conceptualization framework in the platform is structured in four phases [26]: concept elicitation, concept organization, concept sharing and concept discussion. Each of these phases is supported by a set of activities related to terminology and/or knowledge representation, being that the first phase is fully supported by terminological processes, based on texts: collection, identification and classification of resources and terminological extraction. Terminological work also supports the second phase of conceptualization, when experts engage in the organization of concepts.

The conceptualization framework depicted below underpins the advances of this research on methods and tools to support the representation of conceptual structures. This framework provides a structured and multidimensional view over the conceptualization process in what regards to its main phases, activities and artefacts, tying together the terminological and knowledge representation view.

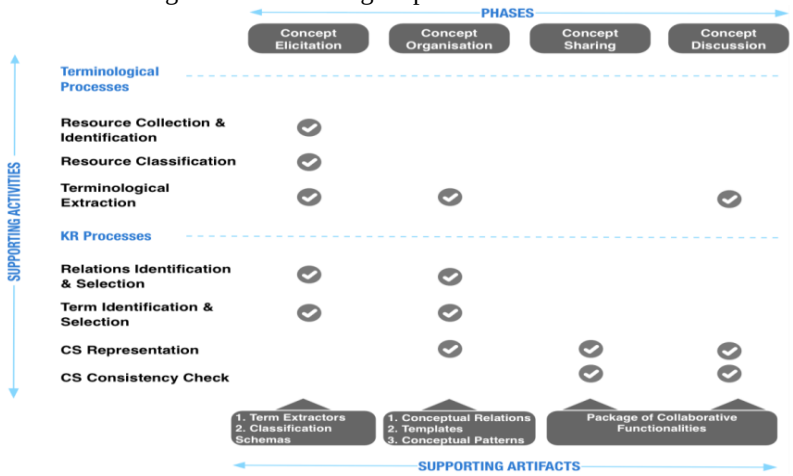

Fig. 1. Conceptualisation framework (*Sousa et al.*, 2012)

The core of conceptME platform is on supporting collaborative modelling, allowing users to create and share conceptual models, focusing on graphical knowledge representations and terminological methods, accommodated into a service's library. The platform enhances, according to [27] negotiation and discussion capabilities by

means of specific extensions, towards consensus reaching. The platform is organized as follows (see figure 1):

a) a set of functionalities to manage ongoing and previous collaborative modelling projects (generic project edition, definition and configuration of the enclosing collaborative spaces and related resources);

b) a collaborative modelling environment, which is language independent, allowing users to build their models individually or editing them collaboratively (either on their own or through available templates), while discussing around concepts;

c) a set of terminological services, based in terminological work methods and techniques, supported by the collection of domain specific textual corpus, which can be built in different languages, allowing users to associate relevant resources to their projects, performing extraction operations to retrieve candidate terms that can be used in their conceptualization process. At this level, conceptME provides: i) means for corpus organization and classification; and ii) real-time term contexts to detail existing representations;

d) a model negotiation baseline enclosing a set of features (merging individual input structures, suggestion mechanism, cross-checking corpus-based validation, autocomplete and categorization, equivalents visualization, among others) to ensure simple negotiation mechanisms, towards a common shared model. This module provides the interface and environment conditions, allowing to connect other advanced negotiation mechanisms (e.g., argumentation-based negotiation and decision-support methods), despite of their nature, domain or language.

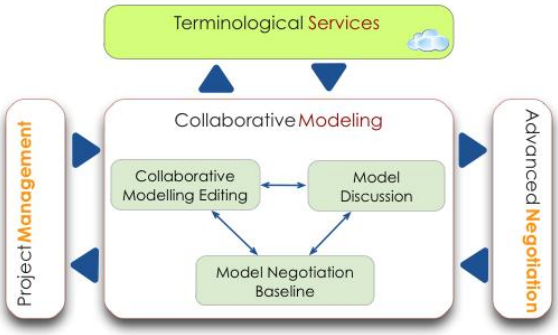

Fig. 2. - ConceptME High-level architecture (Sousa *et al.*, 2012)

3.2 Tools to support the multilingual ontology specification

Working in a collaborative space implies the availability of an environment to help promote the multilingual specification of the conceptual representation, an environment that considers both the social and organizational structure of the community and the type of existing skills. Although localization is a knowledge-based activity [23], the selection of techniques, methods and tools for the localization depend on the resources available for each particular language and for the specialized domain to be represented.

This poses a number of additional difficulties, as the available translation and localization services are almost exclusively focused on document translation and do not consider the needs of communities that operate in a multilingual network and need to deal with the presence of multiple natural languages in a same virtual collaborative space. Thus, to support the presented workflow and the subject field experts' effective participation in the localization process, we have selected a set of easily accessible Web 2.0 translation tools, lexical and terminological database, and developed a lightweight localization service support system to help the user in is search for equivalents, as depicted in the next figure.

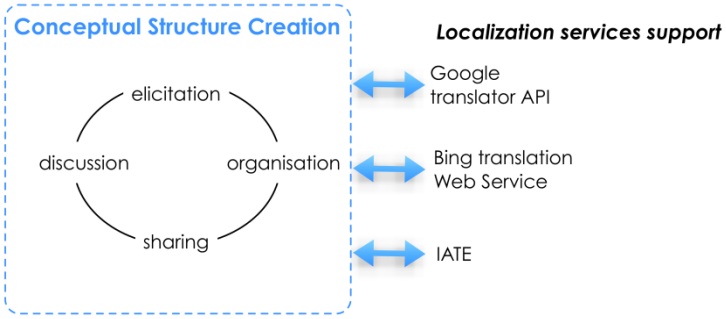

Fig. 5. Localization services support

This selection was done after an analysis of the available web translation tools which took into consideration the ease of use and access by the experts, as well as the specificities of their use in supporting localization for specialized domains. Through this service the user can either choose to localize a single term or the entire conceptual structure, and then validate or discard the results he obtained.

As we could observe, the use of these tools was, nevertheless, clearly influenced by the preexisting domain knowledge, added by the specialist to the process or that resulted from the reuse of other domain knowledge resources such as specialized multilingual dictionaries and glossaries.

4 Application scenario: development of H-Know Ontology

This approach was tested in a preliminary stage in the context of the European project H-Know - *Advanced Infrastructure for Knowledge Based Services for Restoring Buildings*. The project involved partners from five European countries and was developed in a multidisciplinary and multilingual environment involving terminologists, domain experts and knowledge engineers with the objective of building an ontology-based knowledge management platform to support the creation of cooperative and collaborative business networks to facilitate the sharing of Construction Industry knowledge in the domain of cultural heritage and old building restoration/maintenance among the network partners (SMEs and R&D Institutes). This do-

main is characterized by its cyclical and nomad activity, which involves a high number of design and production processes. The knowledge in this domain is disperse, diverse and fragmented, due to its polymorphic character and the amount of actors, rules and institutions that participate in the development of each phase of the construction process.

Management of this knowledge was based on a multilingual domain ontology for the Rehabilitation domain, the H-Know Ontology, developed with the objectives of providing an infrastructure to efficiently and effectively organize, classify and retrieve information and knowledge and to provide H-Know users with a common ground for a shared understanding of terms and concepts when engaging in the virtual collaborative network activities [6].

Implementing an approach to meet the needs of a particular process that has to be developed in a specific context has to take into consideration the users' diversity, as well as their requirements, the existing resources at the time of its implementation and the constraints that occur due to the results' integration in existing applications. In our case, and for the approach testing and implementation, it was considered that the following assumptions were gathered:

1. The collaborative network is formed, is multilingual and its objectives, mission and deadlines are established and accepted by all its members;
2. The partners will be the actors involved in the negotiation process with the aim of reaching a consensus about the representation of the domain's knowledge;
3. Each partner is seen as an expert that actively participates in the conceptualization of the domain ontology and in the localization process, according to the roles, aims and the defined calendar.

4.1 Conceptualization and multilingual specification environment

Specifying an ontology in more than one natural language is a process with its own problems, already described. When the starting point is a conceptual map there may be additional difficulties, given that the expert has to deal with both the knowledge representation specified in each conceptual map and with the localization of terms represented there.

To support the development of this task conceptME offers a conceptualization space, represented in the figure below, to support the specific communicative situation and provides a simple tool and a simple approach to facilitate access to knowledge and to represent it in a multilingual environment through the use of conceptual maps built in a shared environment, where concepts and their relations are made explicit, the equivalents displayed and where experts have the opportunity to include, together with the equivalents, other elements such as natural language definitions or share additional information considered relevant in the discussion area.

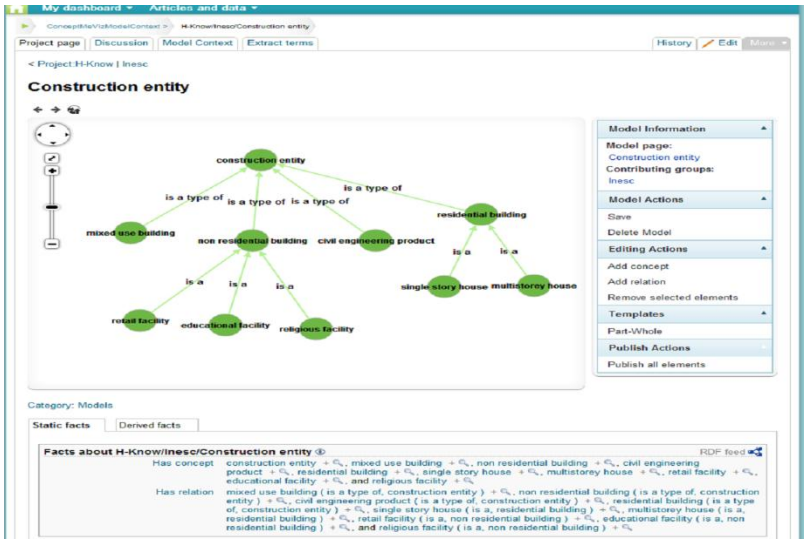

Fig. 3. Conceptual modelling space

This working environment is intended to facilitate a collaborative approach to the development of conceptual representations and to its localization and supports the inclusion of different terminological and linguistic elements, has the expert may had, along with the equivalents, terminological variants, definitions and contexts of use for his/her working language, in order to explain or support his/her choices. This environment also allows the addition and direct visualization of equivalents in the different working languages and the access to the conceptual structures in each language, as portrayed in the next figure.

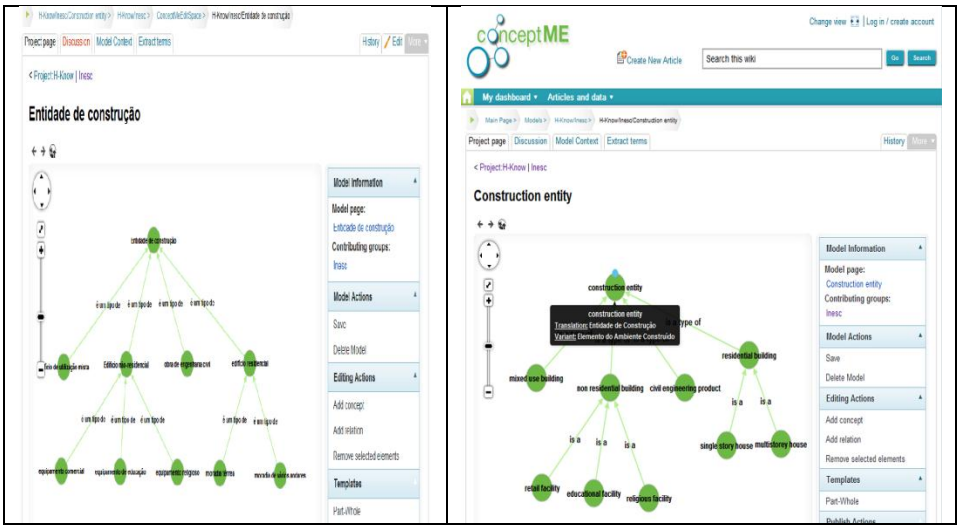

Fig. 4. Conceptual modelling space – multilingual features

The use of these elements intends to support the management of the multilingual information made available in the context of a conceptual map and to create the possibility of providing a homogeneous access to the all the partners of the network, who can thus visualize each other's work and suggestions. The use of a reduced number of elements in this space was decided after considering the time constraints that limit the process of conceptualization and localization, on the one hand, and, on the other, the need to provide a simple and functional working environment that promotes the experts' participation, for whom time is also of the essence.

5 Related work

Considering a multilingual collaborative network, in the approach we propose localization takes place after an initial conceptualization phase, developed using the English language as a starting point, and occurs in a conceptualization space, where the representation of knowledge is developed and made available to domain experts through the use of concept maps, as described in the figure presented below.
The main tasks in the conceptualization and localization for each natural language are (1) the validation of the conceptual structures; (2) translation of the terms that designate the concepts; (3) the translation of the conceptual relations and analysis of their logic validity and (4) reconceptualization, if needed. During this process, the expert must also bear in mind the need to match the represented knowledge to the purposes of the research and information management process that originated the ontology construction.

We do not use, then, a formalized ontology as a basis for localization; rather we start out from a semiformal organization of knowledge in the form of concept maps. The construction of this approach resulted also from the perception that the most commonly used approaches did not fully correspond to the prerequisites of a collaborative network where the need for localized content appears at an earlier stage, due to the short life-cycle that characterizes this type of network.

By promoting and supporting the representation of the different natural languages during conceptualization we differ from other approaches to ontology localization, as those proposed by [15], [2], [8], which focus more directly on the process of enriching formalized ontologies with linguistic elements, and we do not use either a specific ontology localization tool like LabelTranslator [7] or Ontoling [15].

Our approach to the multilingual ontology specification was chosen so as to let us consider not only the individual elements that constitute the conceptual system - concepts and relations and their equivalents in the different languages -, but also, and more importantly, the semi-formal representation as a whole, and assess, with the direct participation of the subject field experts, whether it represented knowledge as it is perceived and expressed by the community for which each expert is localizing it.

The development of this approach is based on a methodology of interlinguistic analysis that functions as a support for the conceptualization of the subject field. It is terminology-based, although it integrates elements from existing methodologies in the area of localization and translation and ontologies engineering. It follows a theoretical

framework that recognizes the conceptualization process as the basis for developing knowledge representation in more than one natural language.

6 Conclusions

The first steps given in the implementation of this approach allowed us to see that the analysis and eventual reconceptualization of the conceptual representations, reinforced by the need to simultaneously develop the localization of the represented concepts, enhanced the experts' awareness, by challenging them with the need to expose and explain their questions, doubts and uncertainties. We also observed that the clarification of doubts may lead to an attempt to conjugate different points of view between experts and between the personal highly specific knowledge and the high-level knowledge representation. This tendency for agreement happens because the expert recognizes himself as part of a collaborative network that is building a semantic representation of a specific knowledge domain which goes beyond what would be an individual representation of that same knowledge, thus valuing the ensemble of opinions and knowledge available, as well as the mediation role played by the terminologist.

This environment proved to be functional and easy to use and allowed users without great experience, who were not prepared to deal with the restrictions of formal semantics, to concentrate on the tasks of conceptualization and localization. The active participation of the experts made it possible, to a certain extent, to reduce some of the problems that hinder the swiftness and effectiveness of localizing specialized knowledge, namely conceptual problems, as experts know the domain, which contributes to reduce ambiguity and increase the semantic precision; linguistic problems, as experts are familiar with the specialized language and recognize most of the terms to localize, needing less time to find the proper equivalent; and pragmatic problems, related to the use of the term, such as its acceptance by peers, which he/she can more easily understand and anticipate.

We recognize, though, that this form of knowledge representation based on conceptual maps has a great degree of complexity which tends to increase when we use conceptual maps to develop a multilingual representation, what may hinder the understanding of the workflow and of the different tasks to be developed. Another limitation lies on the fact that this process may include a large number of the collaborative network experts which may imply, in the chain of contributions and negotiation that is generated, some loss of perception of the original meaning of a concept.

We therefore believe that this approach is adequate to the context of a multilingual collaborative network, a space where multiple partners cooperate in a common effort to represent specialized knowledge in more than one language and that it encourages interaction, knowledge sharing and consensus building.

Acknowledgements: This work is funded by the ERDF through the Programme COMPETE and by the Portuguese Government through FCT - Foundation for Science and Technology, project PTDC/EIA-EIA/103779/2008 "CogniNET".

7 References

1. Budin, Gehard: Ontology-driven translation management. In Knowledge Systems and Translation. Helle V. Dam, Jan Engberg, Heidrun Gerzymisch-Arbogast (edt). Walter de Gruyter. (2005)
2. Buitelaar, P., Declerck, T., Frank, A., Racioppa, S., Kiesel, M., Sintek, M., Engel, R., Romanelli, M., Sonntag, D., Loos, B., Micelli, V., Porzel, R., and Cimiano,P.. Linginfo: Design and applications of a model for the integration of linguistic information in ontologies. In Proceedings of the OntoLex Workshop at LREC. ELRA. (2006)
3. Cahier, J-P., Zaher, L'H., Leboeuf, J-P, Guittard, C.: Experimentation of a socially constructed "Topic Map" by the OSS community. Proceedings of the IJCAI-05 workshop on KMOM, Edinburgh. (2005)
4. Camarinha-Matos, L.: Collaborative networks in industry – Trends and foundations. In: Proc. of DET 2006 - 3rd International CIRP Conference in Digital Enterprise Technology. Springer, Heidelberg. (2006)
5. Gracia, J., Montiel-Ponsoda, E., Cimiano P., Gomez-Perez, A., Buitelaar, P., McCrae J.: Web Semantics: Science, Services and Agents on the World Wide Web. (2011)
6. H-Know - *Advanced Infrastructure for Knowledge Based Services for Restoring Buildings*. www.h-know.eu. (2011)
7. Espinoza M., A. Gomez-Perez, and E. Mena.: *LabelTranslator - A Tool to Automatically Localize an Ontology*. Neon Project. (2008)
8. Espinoza, M., Gómez-Pérez, A., and Mena, E.: Enriching an ontology with multilingual information. In Proceedings of the European Semantic Web Conference (ESWC 2008), pages 333–347. (2008)
9. Espinoza, M., Montiel-Ponsoda, E., and Gómez-Pérez, A.: Ontology localization. In Proceedings of the 5th International Conference on Knowledge Capture (KCAP). (2009)
10. Hirst, G.: Ontology and the lexicon. In S. Staab and R. Studer (eds.), Handbook on Ontologies and Information Systems, pp. 1-21. Berlin: Springer Verlag. (2004)
11. Kharatmal, Meena & G., Nagarjuna: Introducing Rigor in Concept Maps. In M. Croitoru, S. Ferre, and D. Lukose (Eds.), Lecture Notes in Artificial Intelligence: Vol. 6208. International Conference on Conceptual Structures 2010: From Information to Intelligence (p. 199-202). Berlin, Germany: Springer-Verlag. (2010)
12. Monnet Project - *Multilingual Ontologies for Networked Knowledge*. http://www.monnet-project.eu/ (2010)
13. Montiel-Ponsoda E., G. Aguado de Cea, A. Gómez-Pérez, Peters, W.: Enriching ontologies with multilingual information. Natural Language Engineering. Vol. 17 2010 17: pp 283-309. Cambridge University Press. (2010)
14. Montiel-Ponsoda, E., Gracia, J., Aguado-de-Cea, G., Gómez-Pérez, A.: Representing Translations on the Semantic Web. 2nd Workshop on the Multilingual Semantic Web. (2011)

15. Pazienza, M. Teresa, Stellato, Armando: The Protégé Ontoling Plugin - Linguistic Enrichment of Ontologies in the Semantic Web in Poster proceedings of the 4th International Semantic Web Conference (ISWC-2005) Galway, Ireland. (2005)
16. Pereira, C.; Soares, A.: Ontology development in collaborative networks as a process of social construction of meaning. On the Move to Meaningful Internet Systems: OTM 2008 Workshops, Lecture Notes in Computer Science Springer Berlin / Heidelberg. (2008)
17. Pereira, C.; Sousa, C.; Soares, A.: A socio-semantic approach to collaborative domain conceptualization. On the Move to Meaningful Internet Systems: OTM 2009 Workshops, Lecture Notes in Computer Science, 524-533. Springer Berlin / Heidelberg. (2009)
18. Rondeau, Guy – Introduction à la Terminologie. Deuxième édition. Québec: Gaëtan Morin Éditeur, 238 p. ISBN 2891051378. (1984)
19. Roche, Christophe: Terminologie et ontologie. In Langages. 39e année, n°157, 2005. pp. 48-62. Persee. (2005)
20. Dourgnon-Hanoune, A., Salaün, P., Roche, Christophe: Ontology for long-term knowledge. In Proceedings of the 19th international conference on Advances in Applied Artificial Intelligence: industrial, Engineering and Other Applications of Applied Intelligent Systems(IEA/AIE'06), Moonis Ali and Richard Dapoigny (Eds.). Springer-Verlag, Berlin, Heidelberg, 583-589. (2006)
21. Rogers, Margaret: "Lexical chains in technical translation: A case study in indeterminacy". In B. Antia (ed.): Indeterminacy in LSP and Terminology: Studies in Honour of Heribert Picht. Amsterdam/Philadelphia: John Benjamins. 15-35. (2007)
22. Sousa, Cristovão, Soares, António Lucas, Pereira, Carla, Costa, Rute: Supporting the identification of conceptual relations in semi-formal ontology development. In Proceedings of ColabTKR 2012 - Terminology and Knowledge Representation Workshop. LREC. (2012)
23. Wills, Wolfram. Knowledge and Skills in Translator Behaviour. Amsterdam & Philadelphia: John Benjamins. (1996)
24. Wüster, Eugen: "Die Allgemeine Terminologielehre – ein Grenzgebiet zwischen Sprachwissenschaft, Logik, Ontologie, Informatik und den Sachwissenschaften". In Linguistics 119. 61-106. (1985)
25. Barros, Sérgio, Costa, Rute, Soares, António Lucas, Silva, Manuel. Integrating terminological methods in a framework for collaborative development of semi-formal ontologies. Eighth international conference on Language Resources and Evaluation LREC. In Workshop In Workshop ColabTKR 2012 - Terminology and Knowledge Representation Workshop. (2012)
26. Sousa, Cristovão, Pereira, Carla, Soares, António Lucas, Costa, Rute. Supporting the identification of conceptual relations in semi-formal ontology development. Eighth international conference on Language Resources and Evaluation LREC. In Workshop ColabTKR 2012 - Terminology and Knowledge Representation Workshop. (2012)
27. Sousa, Cristovão, Pereira, Carla, Soares, António Lucas. Discussing and collaborating through concepts: the conceptME approach. KEOD 2012 - International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management. (2012)
28. Aguado de Cea, G., Gómez-Pérez, A., Montiel-Ponsoda, E., & Suárez-Figueroa, M. C.. Natural Language-Based Approach for Helping in the Reuse of Ontology Design Patterns. In A. Gangemi & J. Euzenat (Eds.), Knowledge Engineering: Practice and Patterns (Vol. 5268). Springer Berlin / Heidelberg, pp. 32-47. (2008)

Translation Politics and Terminology in Legal Texts for better community networking

Frieda Steurs[1], Hendrik J. Kockaert[1]

[1]Faculty of Language and Communication
Lessius / KU Leuven

Abstract: The globalization of activities in business, governments, and organizations, industrial markets etc. makes it clear that we live in a global, interconnected world. Translation of legal and administrative texts is a major issue both in an international context and in countries with several official languages. In this presentation, I will elaborate on TermWise, a large project dealing with legal terminology and phraseology for the Belgian public services, i.e. the translation office of the ministry of justice. Termwise aims at developing an advanced tool including expert knowledge in the algorithms that extract specialized language from textual data (legal documents). The outcome is a knowledge database including Dutch/French equivalents for legal concepts, enriched with the phraseology related to the terms under discussion.

Keywords: terminology, knowledge management, multilingual contexts, legal texts, authentication

1 Introduction

Over the last two decades, a lot has changed both in the world of business and academia. The globalization of activities in business, governments, and organizations, industrial markets etc. makes it clear that we live in a global, interconnected world. However, globalization also leads to localization and reminds us of the fact that we live in different regions, locations, nations, cultures and organizations. Globalisation even functions as a trigger to enhance localization and recognitions of locales.

Technologies and procedures may be spread all over the world, but the actual implementation, in a specific cultural and linguistic setting will vary considerably. The same tendency can be noticed in the academic world: the Bologna agreement, signed in 1999, changed the world of higher education in Europe in a more than profound way : the internationalization and the new organization of the university programmes led to interesting developments on both bachelor and master level. National legislation of different member states had to be compared and equated to come to international understanding. Due to exchange programmes and joint degrees, more international specialist communication was needed in many languages.

In this paper I will expand on terminology, knowledge management and legal texts, and present a concrete project on legal translation and knowledge management: TermWise.

2 Knowledge is the key factor in society

Communication is getting more and more complex, not only between specialists and laypeople, but even between experts in one and the same discipline. This is especially true when communicating across and beyond language and cultural borders. Today, technical and specialist communication comprises around 80% of all information exchanged across the new communication paths of a borderless and multilingual information society. Terminology can be defined as the entirety of all concepts and terms of a subject field. Therefore, one can equate terminology with specialist vocabulary. Efficient communication with regard to technical language or standardization of concepts is not possible without an exact definition of the concepts under discussion. That means that an onomasiological approach is needed; i.e. the starting point is the concept that has to be defined in an unambiguous way. Once the concept is defined within a taxonomy (a concept system), the terms can be correctly assigned and different terms in different languages can be correctly linked to the concept under discussion.

The quantity and difficulty of specialist texts have increased, along with the demands on the technical and specialist documentation (laws, norms, customer and corporate language). Experts in technical documentation must become familiar with the terminology of their field. Frequently, parts and components have different names in one and the same company. Often much time is lost before the clear terms established there find their way into the linguistic usage of technical languages, not to mention the fact that there is no way by any measure that all technical terms could be standardized. Thus, for the good of specialist communication, it is very important that the meaning of complex terms be defined as early as possible, the results be documented and made available to potential communication partners. An example: one small modification, such as changing part of a technical component, will affect all models in which this part can be found. This means that all language versions of all model descriptions must be revised. This can be very expensive and conceals the risk of errors and confusion among all stakeholders. On both a European and a national level, lawmakers place special requirements on the development of terminology, especially in the area of technical documentation. EU standards, product liability, and CE certification require companies to deliver, as an integral part of their products, documentation that meets safety requirements. Defective documentation is deemed a product defect that leads to complaints or even claims for damages.

The real wealth in a company is the knowledge that is handled and carried by the different employees. This wealth is at the same time the liability in every commercial institute; the companies will strive to explicitate the implicit knowledge, the knowledge in the heads of the individual employees. Good information handling is guided through correct communication, with clearly defined concepts and the terms related to these concepts in the different languages.

3 Multilingual challenges

The growing internationalization of the world trade and all other aspects of human life, brings an enormous challenge in multilingual communication. In the case of the European union, e.g. the number of languages grew exponentially, from 4 over 9 to 11, then 20 and now 23 official languages and the number is still growing (24 official languages in 2013, with the entrance of Croatia), with several non-official regional languages in the background.

Within the given economic context, each company and institute faces a growing technical, scientific and legal complexity. As a consequence, a lot of very sophisticated documents have to be written and translated: e.g. legal and administrative documents, technical specifications, spare parts catalogues, user manuals, procedures, reports etc.

Terminology is clearly seen as a key issue, having a crucial role in good communication. This holds both for internal AND external communication in a particular company. Not only the communication between different departments or units within a particular organization has to be clear and unambiguous; the customer must equally benefit from clear handouts, manuals and other technical communication.

The solution to this type of complex data management can be found in the creation of a central knowledge repository, where the concepts can be defined for the different subdomains under discussion, and the relevant information and terminology can be added.

The need for exact terminology in politics and public affairs is particularly obvious. Ordinances and laws must be based on clearly defined concepts and the correct nomenclature for these concepts must be used in their wording. If they are not, diplomatic complications can easily arise,. Policy statements, particularly in the international context, must also be clear regarding the terminology used.

Without the appropriate terminologies, students cannot be properly educated nor can scientists work with precision. Groups of specialists would not have the communicative means to express themselves in technical languages or to disseminate technical information and access it through information networks. There is a special need for terminology clarification in many innovative areas of medicine. A variety of synonyms develop for the same phenomena because often the same research is conducted at different locations. But because research, clinical medicine, the pharmaceutical industry, legislators, and insurance providers must work closely together, clear communication based on precisely defined terminology is so crucial in the health and medical environment.

4 Globalization, internationalization and localisation

The enormous activity on the world market is often perceived as one major ‘globalisation’ operation. There are many different definitions of globalisation, but most acknowledge the greater movement of people, goods, capital and ideas due to increased economic integration which in turn is stimulated by increased trade and investment. However, a decision by a company to “go global” will need a carefully

planned internationalization and localization script. As such, globalization triggers a lot of activity on the level of multilingual communication, translation. Globalisation encompasses all types of business aspects related to the promotion of a product on the worldwide market: internationalisation & product design, adaptation to the local markets, marketing, sales and technical support on the world market

Internationalization involves the designing of products in such a way to make them applicable to many *different markets*, in *different languages* and *cultural conventions* without changing the original design

Localization is the adaption of a product linguistically or culturally to a particular market (target locale), country or region. This can be an integrated part of the design of the product

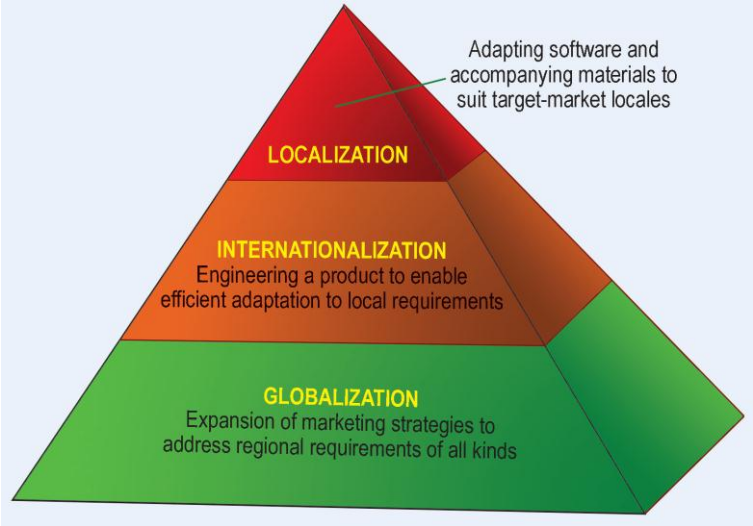

Fig. 1. Sykes, Mulitlingual (2009)

5 The language industry

Last December, the European commission published a report presenting the results of a study conducted on the size of the language industry in Europe. The language industry embraces the following domains:

- Translation, interpreting and consultancy
- Audiovisual translation
- Software localization
- Localization of products in multilingual settings (o.a.website globalization)
- Translation technology: CAT, MT, terminology databases etc.
- Language courses, e-learning

The overall results of this study for the EU showed that this industry is a major player in the field, and that it realises an average growth of more than 10% each year.

This survey only shows results for the European Union; we all know the real growing markets are outside the EU, with booming translation activities in China, India, the Middle East and the America's. As a result of this evolution, terminology management becomes more and more a core element in global communication.

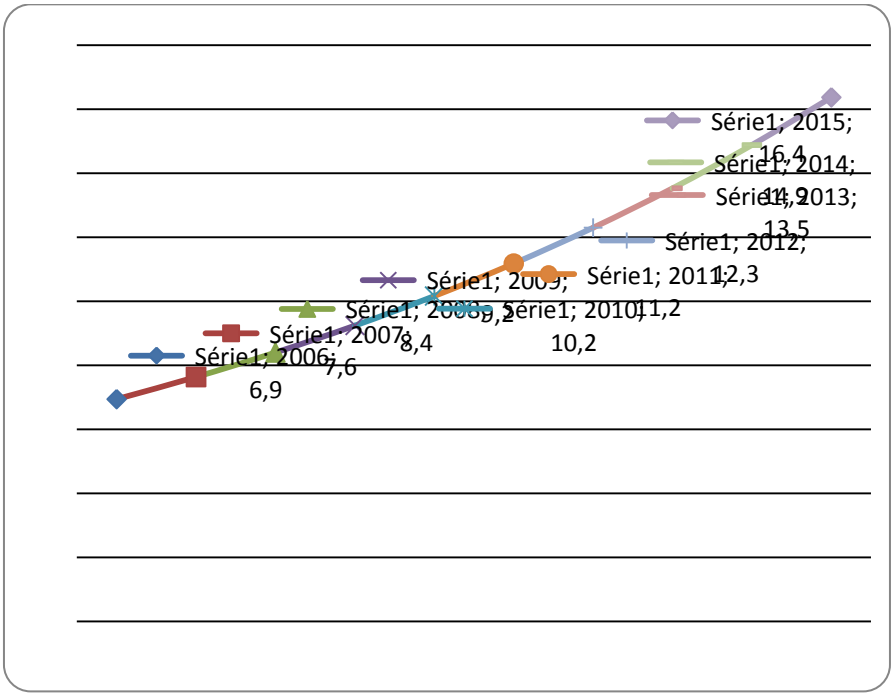

Fig. 2. (http://ec.europa.eu/dgs/translation/publications/studies/index_en.htm)

Different European countries have their own multilingual challenges: we conducted a study quite recently for the Belgian market on the average demand for translation jobs. (Barbé & Claes 2007) The following text types and topics are most heavy in demand:

- Administrative & legal documents: 25%
- Medical and biomedical material: 20%
- ICT (computing, telecom, etc.): 20%
- Manuals, user guidelines, patents : 10%
- Audiovisual translation (subtitling) : 10%
- Several other activities are also in demand: Product localisation, Websites adapation, e-sales, marketing, advertising, etc.

5.1 Terminology and general language: LSP vs LGP

Terminology (in its narrow sense of describing and inventorising terms) is very often considered to constitute a special subpart of lexicography, as it deals with a subset of language, the so called LSP 'Language for Special Purposes'. Terminology (in a broader sense, as a science) is considered to be interdisciplinary, and relates to the intersection of various fields of knowledge: logic, ontology, linguistics, information science, etc. Faber (2009) describes this as follows: terminological units can be seen as

- Linguistic entities in linguistics
- Concept entities in ontology and cognitive sciences
- Communicative units in the more restricted framework of scientific and technical distance.

This leads to a number of challenges in the field of the theory of terminology, which up to now, has not been considered to be a true theory in its own right, but variable according to the approach and the discipline involved. In the literature, we can find samples of a cognitive approach to terminology, a focus on the linguistic dimension, or a focus on the communicative dimension.

Many terminologists have been influenced by the Vienna school. The theories developed under this influence all approach terminology as the study of specialized knowledge structures in order to identify and define concepts belonging to a given domain. Once this is done, one can proceed to the inventory and analysis of the terms used to label a specialized concept, their form, their relationships with one another and their usage status among specialists of that particular domain. This onomasiological approach is the defining principle of terminology research.

Concepts are strictly delineated from each other, are clearly defined and are organised in a concept system. The term-concept relation is very strict and studies on a synchronic basis. The other principles that govern terminology research are:

- Rules for structuring knowledge based on classification systems (documentary and others)
- Rules for building concept systems by means of various representations (trees, diagrams, networks, etc.)
- Rules for defining concepts by means of selected characteristics
- Rules for term identification, collection, formation and usage
- Rules for recording terminological information based on the single-concept principle

6 TermWise

A unique project in terminology management was set up for the legal multilingual challenges of the federal department of justice in Belgium (official languages Dutch and French, pertaining to the Belgian legal system).

The TermWise Knowledge Platform is a project funded by the IOF (Industrial Research fund) of the University of Leuven (KU Leuven) ,and aims to deliver the proof-

of-concept for a software tool that offers comprehensive multilingual terminological support to language professionals like translators and copy-writers dealing with specialized language use. The core of this tool will be a rich knowledge created with computational knowledge acquisition algorithms and made accessible via a user-friendly interface. The tool will be tested and validated in the domain of Belgian legal terminology in French and Dutch.

However the algorithms will be explicitly designed to be generic and portable to other languages and domains. In a very concrete way, the tool will put the following information about the legal jargon at the user's fingertips:

- Comprehensive inventory of terminological expressions in the legal domain
- Illustration of the terms' different meanings via informative examples
- Idiomatic term usage (proper use of the terms in context)
- Differences in term use between subdomains (e.g. federal vs. regional law)
- Correct translation of French terms into Dutch and vice versa
- Meaning relations between terms (synonyms, terms for related concepts)

Giving language professionals to access to this information which is not or scarcely provided by current commercial software holds the potential for greatly improving the quality of legal translating and text writing. Delivering the proof of concept for the platform's techniques will therefore open up possibilities for further contract research for commercial companies or product development within a spin-off.

The TermWise knowledge platform is a multidisciplinary co-operation between fundamental and applied linguists, computer scientists, software engineers and end-users (translators from the federal department of justice). The project is defined in 6 workpackages:

WORK PACKAGE 1: EXTERNAL VALIDATION

TermWise is essentially a user-driven project. It was initiated in response to specific needs expressed by language professionals. Therefore, the expert users are not only integral members of the project board monitoring the project; they have also agreed to actively participate in the project at specific stages. This co-operation is essential for guaranteeing the valorization possibilities of the knowledge platform.

A first interaction takes place during the start-up phase of the project in which the terminological model will be defined. The research partners and the users will formulate what they regard as terminological expressions and which knowledge about these terminological units should be acquired. In other words, the terminological model lays out the general specifications for the rest of project. The document containing the specifications is a crucial milestone for the entire platform and has to be approved by the Project Board. At the end of the first year when a number of work packages have delivered their first results, the expert users, together with the rest of the Project Board, will check again whether the specifications have been correctly interpreted by the members of the project team.

The expert users will actively participate in the validation of the extracted terms and term translations. The most active involvement of expert users will take place in the later stages of the project when they will test-run the project’s knowledge base and interface. In practice, the translation cell of the Justice department will use the

tool to in support of their legal document translation and users will report their experiences.

These experiences and the knowledge platform's research results will be summarized in the final report written by the project manager. Based on this report, the board will then assess whether the proof of concept for the terminological tool has been delivered.

WORK PACKAGE 2: DATA PRE-PROCESSING AND SETUP

The initial phase of the knowledge platform will be dedicated to the collection and setting up of the data resources for the later stages of the project.

The project's input data will be French and Dutch, general and legal text corpora that are syntactically annotated. Large parsed Dutch and French newspaper corpora are already being compiled by the research group in the framework of other projects. The legal corpora will be provided by the translation cell of Justice Department. The project manager will carry out the automatic parsing using the work flow and computational resources established in previous projects. The Justice department has also agreed to make its current legal databases available. While not complete, these term bases and translation memories will be used for validation purposes in the first stages of the project. Since the platform's main deliverable is a rich knowledge base, the specifications for this database will be laid out and implemented from the very beginning, in compliance with the existing ISO-standards for terminological databases. All partners are future users of the database and will therefore be involved in this process

WORK PACKAGE 3: TERM EXTRACTION

Term extraction refers to the process of identifying the words and expressions that are typical for a specialized domain, in this case the legal domain in French and Dutch. The scientific goals of the WP include a better coverage of terminological units through the integration of statistical corpus analysis. This should provide better results than the commercially available state-of –the-art software.

WORK PACKAGE 4: TERM ALIGNMENT

Term alignment refers to the process of retrieving translational equivalents for terms across languages. Within the TermWise knowledge platform, we will align French and Dutch legal terms. The scientific goals of this WP include the optimization of statistical alignment algorithms for parallel corpora and their extension to comparable corpora. This research will result in publications and a doctoral thesis. In terms of valorisation, the WP will offer extra functionality as compared to existing commercial terminology management software. The latter currently allow to store previously translated chunks of texts in translation memories, but they leave the actual alignment up to manual analysis.

WORK PACKAGE 5: SEMANTIC MODELLING

A next step will analyze the meaning of terms and the semantic relations that exist between them. The scientific goals of the WP include the application of semantic vector space models to the large-scale analysis of meaning-context relationships.

In terms of valorisation, this WP aims to offer users insight into the precise meaning and correct usage of specific terms as compared to potential alternative terms for the same concept. This information is scarcely included in existing software and users are currently often left to 'googling' for informative examples to induce correct term use.

WORK PACKAGE 6: INTEGRATION AND INTERFACE DEVELOPMENT

WP6 constitutes the final phase of the project in which the R&D from the different WPs is combined, integrated and made accessible to expert users for evaluation. WP6 is of prime importance for valorisation in as far as it makes sure the knowledge platform is not just a loose collection of research results but offers a coherent answer to the needs of our target audience.

7 Conclusion

Terminology management is a crucial element in modern knowledge management, both from a conceptual point of view and a multilingual one. Projects such as Termwise, who are at the same time strongly user driven and research based, may create excellent results for modern advanced terminology management.

8 Selected Bibliography

1. Bassey, Antia(ed) Indeterminacy in Terminology and LSP. Studies in honourof Heribert Picht. John Benjamins, Amsterdam/Philadelphia. (2007)
2. Faber, Pamela, Pilar León Araúz, and Juan Antonio Prieto Velasco. Semantic Relations, Dynamicity, and Terminological Knowledge Bases In : Current Issues in Language Studies 1: 1-23. (2009)
3. Kockaert, H., Steurs, F.. Un outil de gestion terminologique pour la traduction juridique en Belgique : état de la question et perspectives. Revue Française de Linguistique Appliquée, 16, 1, 93-104. (2011)
4. Steurs, F. Thelen, M (eds). Terminology in Everyday Life. John Benjamins, Amsterdam/Philadelphia. (2010)
5. Sikes, Richard. Localization: The Global Pyramid Capstone. (D. Parrish, Ed.) MultiLingual. Localization Guide: Getting Started , pp. 3-6. (2009)
6. Van Sterkenburg, Piet (ed) A Practical guide to lexicography. John Benjamins, Amsterdam/Philadelphia. (2003)
7. Wright, Sue Ellen, Gerhard Budin (eds). Handbook of terminology management, volume I. John Benjamins, Amsterdam/Philadelphia. (1997)
8. http://ec.europa.eu/education/languages/languages-of-europe/doc4015_en.htm
9. http://ec.europa.eu/dgs/translation/publications/studies/index_en.htm

Subject Librarians working in a multi-cultural and multi-linguistic context: an ontology-based approach to semantic cataloguing and information retrieval

Deborah Grbac, Luca Losito, Andrea Sada, Paolo Sirito[1],

[1]Università Cattolica del Sacro Cuore – Biblioteca d'Ateneo
Largo Gemelli 1 - I 20123 - Milano (MI) - Italy
{deborah.grbac, luca.losito, andrea.sada, paolo.sirito}@unicatt.it

Abstract. In this paper, we report the preliminary results of a pilot project currently ongoing at UCSC Central Library, inspired to the methodologies used by translators working within European Union Institutions. User Experience (UX) Analysis techniques are being used, in order to provide subject librarians with a visual support, by means of "ontology tables" depicting conceptual linking and connections of words with concepts presented according to their semantic and linguistic meaning.

Keywords: Ontology; Academic Libraries; European Union; Semantic relationship; Terminology; Reference services; Legal Subject Headings; Information retrieval

1 Introduction: when a Librarian may become a Translator

Librarians are translators? Usually Librarians follow an education path which has got very few in common with the one followed by translators, but in some circumstances they have to deal with some similar problems and our hypothesis is that they both need instruments useful to afford close challenges.

Librarians, as also translators, have to afford various and complex domains of reference, when they have to treat with specific kind of books, as translators they do not always have a scientific background enabling them to immediately and fully appreciate the contents they have to deal with, and finally, as translators, they could find themselves in situation of shortage of time, in which they have to take a decision. This is in particular the case of librarians having to deal with specific reference domain as law, and moreover if they have to treat books in law written in foreign languages.

Not only librarians could not have a background in law, but it can also happen that they do not know the language in which the book has been written. Otherwise, into a Library is rare to find out a specialist in law domain, for different reasons: lawyers have got their own career path, determined by a customary imprinting which hardly foresees a Library as a possible outcome.

Moreover, lawyers have to forcibly determine quite early in their education if they will opt for one of the two law branches: "public law" or "private law". If you

specialise in public law, you won't have skills and won't develop culture necessary to deal in deep with private law matters, and the opposite. Second, because of the difference between roman culture law and common law it is hard for lawyers to understand different law systems and this difficulty may be enhanced by the fact that each language, as regards to law, adopts its own linguistic code (legal language). Third, if to compare different law systems expressed in different legal languages is hard, more difficult is to consider even more complex law systems as the one of international law or European law.

2 Which Instrument can be borrowed from Translators' toolbox?

Translators are really multi-tasking and they are able to deal with complex situation. Apparently, librarians are not put under the same pressure when they have to do their job, but in some cases: when they have to work in an environment containing the same elements of complexity (multi-lingual context and specific and technical domain), as also translators, they have very few instruments at their disposal.
For this reason we propose in this paper to adapt a model originally conceived for helping translators to librarians working with a particular kind of books: European law books.

3 Why European Law?

European law is a specific branch of international law, falling under the public law branch, treating with some peculiar phenomena: the integration process among a certain number of European countries in some specific areas. Not a confederation of States, nor an international organisation, the European integration process is a third way in International law, which is evolving according to the complex mechanism of the *consensus* given by participating countries to proceed in a specific way. Moreover, instead of other international organisation working with two or at least three official languages, European Union has 26 official languages (number due to grow according to new member States' accession to the Union), along with, because of its specific nature, a technical language: the "jargon" talked in European Union institutions transversal to the 26 official languages.

Finally, even if studies in European law are developing, as a topic it remains a very peculiar branch of public law, moreover students interested in European law, because of the interdisciplinary and multilingual approach given to the subject, aims at working within European institutions in Brussels and Luxembourg. If it is rare to find out a lawyer in a Library, harder is to find out a specialist in European law. The risk is that managing books treating that kind of topic could be at least imprecise, or even mistreated.

4 The Model Proposed

We propose to adapt to subject librarians, employed in large and multilingual Academic Institutions, the model used by translators working within European Union Institutions. The method: offering a visual support for people called to take rapidly decisions by means of "ontology tables" depicting conceptual linking consisting of collections and connections of words with concepts presented according to their semantic and linguistic meaning along with their official technical translation.

For the purpose of European law, the terminological data-base of reference used is the one conceived by translators of European Union institutions called IATE (Inter-Active Terminology for Europe). This data-base is access-free (http://iate.europa.eu/) and it works in 26 official languages of the European Union. Support tables will be constructed in the form of "ontology tables" and they will enable librarians, not having done studies in European law or not knowing one or more of the language of the European Union, to detect the contents of the book they have to analyse and treat them in the correct way.

A dedicated library team, skilled in semantics and ontology building, is currently working at a pilot project aimed at developing a cataloguing support tool. The methodology is based upon User Experience (UX) Analysis research and is leveraging the long standing activity of more than 12 subject librarians. In a nutshell, the ontology building team has been studying the behaviour and the semantic choices of the subject librarians. A specific tool has been developed by the Library internal IT support staff (namely: a Data Librarian and an Electronic Resources Specialist) in order to capture the inferential choices of the subject specialists.

By making use of an inductive approach, a semantic mapping tool has been released, enabling researchers to better understand the underlying choices of the specialist and – moreover – the logic behind the real time conversion from natural language to artificial one.

Obviously, as more than 35% of the documents are in English, the very same approach has been applied in multilingual context, by implementing dedicated linguistic conversion tables.

In the case proposed in this paper: cataloguing the specific domain of European Law, we refer as for the specific terminology used in Brussels to the above mentioned IATE data-base, the instrument collecting all ancient terminological data-bases used by European Union Translation Services, as it has been implemented by the European Union Translation Centre. Completed by another European Union data-base EUR-Lex (http://eur-lex.europa.eu/en/index.htm), which contains all the European Union law, treaties, international agreements, preliminary documents, further regulation (Directives and Regulations), Court of justice of the European Union and other Tribunals case-law, and finally, if necessary as further reference, the official European Union web-site: http://europa.eu/.

5 Preliminary Results

Table below shows how our model works in practice. In the first column we identify the name of authors of the documents analysed. Usually authors focalise their studies

in some specific topics and they follow them during their scholar career, this enables to detect some useful specialisations according to which organize information. For instance, in the first row we propose an example of a textbook in European Law, collecting texts and materials with a peculiar attention to the fundamentals of the topic (institutions and main legal instruments). In the second case of the first row we put the title of the book to be catalogued along with a brief description of the topic treated in order to give to the librarian a more precise context. In the third case we select some keywords according to title, sub-title, table of contents of the book. Finally according to keywords chosen we propose one or more related subjects to be used to complete the cataloguing work.

Different topics are collected in the table. We begin with the foundations of European Union Law: Institutional treaties, in particular the Treaty of Lisbon which is an amending treaty (the last one in force), amending the Treaty on European Union (Treaty of Maastricht) and the European Community Treaty (Treaty of Rome now renamed as the Treaty on the functioning of the European Union). Then as the European Union Institutional provisions include also those on the jurisdiction of the Court of Justice, we inserted the Court jurisdiction and in particular the 'previous ruling' mechanism granting an uniform European Union law interpretation.

The Treaty of Lisbon gives to European Union juridical personality and the direct responsibility for some policies. In the following table, European Union policies and internal action have been listed, in particular we made some references to: internal market, workers and capital (two freedoms of the four ones, the other two are establishment and services), area of freedom-security and justice (in particular judicial cooperation in civil and criminal matters and police cooperation), competition law (State Aid included).

European Union internal action is expressed by means of some legal acts adopted by the European Union Institutions (substantive law) and implemented by Member States in national legislation. As examples we propose a few well known Council Regulations and Directives (Brussels Regulations I and IIbis and MiFID Directive).

Finally, we treated of the external action of European Union, according to its two branches: external action and foreign policy. Then as the constitutional architecture of the European Union includes also the Charter of Fundamental Rights, we added the revised version of the Charter as proclaimed on 12 December 2007 and the consequent action pursued by European Union in human rights protection. At the end of our table we added activities of cooperation in the fields of justice and home affairs and police and judicial cooperation that, by means of the Treaty of Lisbon, have been definitely inserted among the European Union policies.

The table shown below constitutes an easy-readable conceptual map, expressed in three languages, containing links between keywords, technical concepts and related subjects that can be used by librarians to take decisions about which subjects to be chosen, enabling them to save time and to work in a specific technical domain and in more languages. Obviously, by the same instrument, also retrieval information is possible as it is already adopted at the Reference Desk of the Law Library Department at our University.

6 Concluding Remarks

The main issue we addressed in elaborating the table below was the need to deal with three different levels of translation: from technical language (Euro-crates jargon) to common language, from common language to artificial language, from the subject in the language chosen to translation into other languages.

Recurring to IATE has been useful to proceed to the final translation from one language to another and to verify the correct expressions used in the European Union law, independently from the language used. However, translation from common language to "controlled language" remains difficult, because of some peculiarities of the European Union law which are hard to be reduced to a closed list of words used by librarians. The main risk in this case is to loose some of the aspects of European Union law, which is subject to continuous development.

For instance, one of the achievements of the Treaty of Lisbon is that European Union has obtained an international (not dependent from Member States) juridical personality. This means that some attributes and powers are directly attachable to the European Union responsibility. In terms of the sentence constituting a subject the world: European Union has to take the first place and be followed by the policy or activity the European Union has a responsibility for. Unfortunately, because of the limits of the artificial language, sometimes this is not possible and the European Union has to take the final place in the sentence, meaning that that phenomena or policy are applied in the European Union as a territory.

Putting European Union at the end of the subject-sentence deprives the Union to its power of having an initiative in a given policy. Moreover, the lack of opportunity of talking directly about policies means that we have to recur to other words, not always clear, with the consequence that they may not be easy to be understood for non-lawyers. This is the reason why we insisted with keywords in our table and we tried to give examples on the main topics of the European Union law, in order to make our table comprehensive and to allow librarians to fully appreciate the linkage existing between some European Union topics and the relevant subjects.

Table 1. Preliminary Results: an Example of Semantic Cataloguing Reference Table in European Union Law

Reference Authors	Title and topic description	Keywords (in more languages)	Subjects connected
Craig Paul, Búrca de Gráinne	**'EU law. Text, Cases, and Materials'**	**European Law/Droit de l'Union européenne/Diritto dell'Unione europea**	European Union-Legal System Union européenne-Ordre juridique Unione Europea-Ordinamento istituzionale
	European Union Law: Treaties and Institutions Droit de l'Union européenne: les traités et les institutions (droit institutionnel) Diritto dell'Unione europea: i trattati e le istituzioni (diritto primario)	Treaties/traités/trattati Texts/textes/testi Institutions/intitutions/istituzioni	European Union-Treaties and International Agreements If documents are attached as texts: European Union-Treaties and International Agreements- Texts Union européenne-Traités and accords internationaux Union européenne-Traités and accords internationaux-Textes Unione Europea- Trattati e convenzioni internazionali Unione europea-Trattati e convenzioni internazionali-Testi
	European Union Judicial System and case-law Le système judiciaire de l'Union européenne et la jurisprudence L'ordinamento giudiziario dell'Unione europea e la giurisprudenza	Judicial system/Système judiciaire/Ordinamento giudiziario Case-law/jurisprudence/giurisprudenza Cases/arrêts/sentenze	European Union-Judicial System Union européenne-Système judiciaire Unione europea-Ordinamento giudiziario Court of justice of the European Union–Case-law Cour de justice de l'Union européenne-Jurisprudence Corte di giustizia dell'Unione europea-Sentenze

	European Union further regulation (Regulations, Directives and other) Le droit matériel de l'Union européenne (Règlements, directives et autre) Il diritto derivato dell'Unione europea (Regolamenti e direttive e altro)	Materials/Documents/Documenti Substantive law/Droit matériel/Diritto sostanziale	An example in European Union Competition Law Competition – European Union Regulations Concurrence – Règlements de l'Union européenne Libera concorrenza- Regolamenti dell'Unione europea
Piris, Jean-Claude	**'The Lisbon Treaty. A legal and political analysis'**	**Treaty of Lisbon/ Traité de Lisbonne/ Trattato di Lisbona** Treaty of Lisbon amending the Treaty on European Union and the Treaty establishing the European Community, signed at Lisbon, 13 December 2007 Traité de Lisbonne modifiant le traité sur l'Union européenne et le traité instituant la Communauté européenne, signé à Lisbonne le 13 décembre 2007 Trattato di Lisbona che modifica il trattato sull'Unione europea e il trattato che istituisce la Comunità europea, firmato a Lisbona il 13 dicembre 2007	Treaty of Lisbon (2007) Treaty of Maastricht (1992)-Protocols, etc.-2007 Dec. 13 Treaty of Rome (1957) -Protocols, etc.-2007 Dec. 13 Traité de Lisbonne (2007) Traité de Maastricht (1992)-Protocoles, etc.-2007 déc. 13 Traité de Rome (1957) -Protocoles, etc.-2007 déc. 13 Trattato di Lisbona (2007) Trattato di Maastricht (1992)-Protocolli, ecc.-2007 dic. 13 Trattato di Roma (1957)-Protocolli, ecc.-2007 dic. 13
Türk, Alexander H.	**'Judicial review in EU law'**	**Interpretation of European Union law/Interprétation du droit de l'Union européenne/Interpretazione del diritto dell'Unione europea** National Court requesting a preliminary ruling/Renvoi préjudicial d'une jurisdiction nationale/Rinvio preguidiziale da parte di una giurisdizione nazionale	International jurisdiction-European Union Juridisction internationale-Union européenne Giurisdizione internazionale-Unione Europea. Preliminary ruling-Procedural law-European Union Renvoi préjudiciel-Droit de la procédure-Union européenne Giudizio di rinvio-Diritto processuale-Unione europea
Gormley, Laurence W.	**'EU law of free movement of goods and customs union'**	**European Union Single Market/Marché intérieur de l'Union européenne/Mercato unico dell'Unione europea**	Customs-European Union Douanes-Union européenne Dogane-Unione europea

		European Union trade policy/ Politique commerciale de l'Union européenne/ Politica commerciale dell'Unione europea	European Union-Economic policy Union européenne-Politique économique Unione europea-Politica economica International Trade-European Union Commerce international-Union européenne Commercio internazionale-Unione europea
El-Agraa, Ali M.	**'The European Union, Economics and policies'**	**Free movement/libre circulation/libertà di circolazione**	For example: Capitals Capitals-International Movement-European Union Capitaux-Circulation internationale-Union européene Capitali-Circolazione internazionale-Unione europea For example: People Civil and political rights-European Union Droits civils et politiques-Union européenne Diritti civili e politici-Unione europea
Lovdahl Gormsen, Liza	**'A principled approach to abuse of dominance in European competition law'**	**Competition policy/Politique de la concurrence/Politica della concorrenza** Competition law/Droit de la concurrence/Diritto della concorrenza Abuse of dominant position/Abus de position dominante/Abuso della posizione dominante	Competition-European Union Concurrence-Union européenne Libera concorrenza-Unione europea
Keppenne, Jean Paul	**'Guide des aides d'État en droit communautaire. Réglementation, jurisprudence et pratique de la Commission'**	**State Aid/Aides d'Etat/Aiuti di Stato**	Business enterprises-Financing-European Union Entreprises-Financement-Union européenne Imprese-Finanziamento-Unione europea

Magnus Ulrich, Mankowski Peter	**'Brussels I Regulation'**	**Substantive law/Droit matériel/Diritto sostanziale**	Private International Law-European Union Droit international privé-Union européenne Diritto internazionale privato-Unione europea
	"Brussels I Regulation" Council Regulation (EC) No 44/2001 of 22 December 2000 on jurisdiction and the recognition and enforcement of judgments in civil and commercial matters (OJ L12 16th Januray 2001)	Regulation (EC) No 44/2001 - Jurisdiction and the enforcement of judgments in civil and commercial matters	European Union Council - Regulation (EC) No 44/2001
	Règlement dit "Bruxelles I **"** Règlement (CE) nº 44/2001 du Conseil du 22 décembre 2000 concernant la compétence judiciaire, la reconnaissance et l'exécution des décisions en matière civile et commerciale (JO L du 16 Janvier 2011)	Règlement (CE) nº 44/2001 - Compétence judiciaire et exécution des décisions en matière civile et commerciale	Conseil de l'Union européenne - Règlement (CE) nº 44/2001
	Regolamento chiamato "Bruxelles I", Regolamento (CE) n. 44/2001 del Consiglio, del 22 dicembre 2000, concernente la competenza giurisdizionale, il riconoscimento e l'esecuzione delle decisioni in materia civile e commerciale (GU L 12 del 16.1.2001)	Regolamento (CE) n. 44/2001 - Competenza giurisdizionale ed esecuzione delle decisioni in materia civile e commerciale	Consiglio dell'Unione europea - Regolamento (CE) n. 44/2001
	Conflicts of law and law applicable to contracts or else	Conflicts of law/conflits de loi/conflitti di legge	Conflicts of law-Private International Law-European Union Contracts- Private International Law- European Union
	Conflits de loi et loi applicable aux contrats ou autre	Law applicable to contratcs/loi applicable aux contrats/legge applicabile ai contratti	Conflits de lois-Droit international privé-Union européenne Contrats- Droit international privé-Union européenne
	Conflitti di legge e legge applicabile ai contratti o altro		Conflitti di legge-Diritto internazionale privato-Unione Europea Contratti-Diritto internazionale privato-Unione Europea
Magnus, Ulrich Mankowski, Peter	**'Brussels IIbis Regulations'**	Brussels IIbis Regulations/Règlement dit Bruxelles IIbis/Regolamento Bruxelles IIbis	Private International Law-European Union Droit international privé-Union européenne Diritto internazionale privato-Unione Europea

		Council Regulation (EC) No 2201/2003 of 27 November 2003 concerning jurisdiction and the recognition and enforcement of judgments in matrimonial matters and the matters of parental responsibility, repealing Regulation (EC) No 1347/2000 (OJ L 338, 23.12.2003, p. 1–29)	European Union Council - Regulation (EC) No 2201/2003
		Règlement (CE) n° 2201/2003 du Conseil du 27 novembre 2003 relatif à la compétence, la reconnaissance et l'exécution des décisions en matière matrimoniale et en matière de responsabilité parentale abrogeant le règlement (CE) n° 1347/2000 (JO L 338 du 23.12.2003, p. 1–29)	Conseil de l'Union européenne - Règlement (CE) nº 2201/2003
		Regolamento (CE) n. 2201/2003 del Consiglio, del 27 novembre 2003, relativo alla competenza, al riconoscimento e all'esecuzione delle decisioni in materia matrimoniale e in materia di responsabilità genitoriale, che abroga il regolamento (CE) n. 1347/2000 *(*GU L 338 del 23.12.2003, p. 1–29)	Consiglio dell'Unione europea - Regolamento (CE) n. 2201/2003
	Cross-border divorce and cross-border lawsuits concerning parental responsibility/ Divorce transfrontalier et contestations judiciaires concernant la responsabilité des parents/Divorzio transfrontaliero e vertenze giudiziarie riguardanti la responsabilità dei genitori	European divorce/devorce européen/divorzio europeo	Family-Private International Law- European Union Divorce-Private International Law- European Union Famille- Droit international privé-Union européenne Divorce- Droit international privé-Union européenne Famiglia-Diritto internazionale privato-Unione Europea Divorzio-Diritto internazionale privato-Unione Europea

Casey, Jean-Pierre	**'The MiFID revolution'**	**MiFID Directive/ Directive MiFID/Direttiva MiFID**	Financing-European Union Directives Financements-Directives de l'Union européenne Finanziamenti-Direttive dell'Unione Europea
		Directive 2004/39/EC of the European Parliament and of the Council of 21 April 2004 on markets in financial instruments amending Council Directives 85/611/EEC and 93/6/EEC and Directive 2000/12/EC of the European Parliament and of the Council and repealing Council Directive 93/22/EEC (OJ L 145, 30.4.2004, p. 1–44) Directive 2004/39/CE du Parlement européen et du Conseil du 21 avril 2004 concernant les marchés d'instruments financiers, modifiant les directives 85/611/CEE et 93/6/CEE du Conseil et la directive 2000/12/CE du Parlement européen et du Conseil et abrogeant la directive 93/22/CEE du Conseil (JO L 145 du 30.4.2004, p. 1–44)	European Union Council - Directive (EC) No 39/2004 Conseil de l'Union européenne - Directive (CE) nº 39/2004

		Direttiva 2004/39/CE del Parlamento europeo e del Consiglio, del 21 aprile 2004, relativa ai mercati degli strumenti finanziari, che modifica le direttive 85/611/CEE e 93/6/CEE del Consiglio e la direttiva 2000/12/CE del Parlamento europeo e del Consiglio e che abroga la direttiva 93/22/CEE del Consiglio (GU L 145 del 30.4.2004, p. 1–44)	Consiglio dell'Unione europea - Direttiva (CE) n. 39/2004
Bosse-Platière, Isabelle	**'L'article 3 du traité UE : recherche sur une exigence de cohérence de l'action extérieure de l'Union européenne'**	**International relations/Relations internationales/Relazioni internazionali** External Action/action extérieure/azione esterna Development cooperation/Coopération au développement/Cooperazione allo sviluppo	European Union-International relations Union européenne-Relations internationales Unione europea-Relazioni internazionali
Verlin Laatikainen, Katie	**'European foreign and security policy after Lisbon'**	**EFSP, European Foreign and Security Policy** **PESC, Politique étrangère et de sécurité commune** **PESC, Politica estera di sicurezza comune**	European Union-Foreign Policy Union européenne-Politique étrangère Unione europea-Politica estera
Torres Pérez, Aida	**'Conflicts of rights in the European Union. A theory of supranational adjudication'**	**Human rights/Droits de l'homme/Diritti dell'uomo**	Human rights-Protection-European Union Droits de l'homme-Protection-Union européenne Diritti dell'uomo-Tutela-Unione europea

Burgorgue-Larsen, Laurence	**'Traité établissant une Constitution pour l'Europe'**	**Fundamental rights and freedoms/Droits et libertés fondamentaux**	Charter of fundamental Rights of the European Union (2000)-Protocols, etc.,-2007 Dec. 12 Charte des droits fondamentaux de l'Union européenne (2000)-Protocoles, etc.,-2007 déc. 12 Carta dei diritti fondamentali dell'Unione Europea (2000)-Protocolli, etc.,-2007 dic. 12
Karolewski, Ireneusz Pawel	**'Citizenship and collective identity in Europe'**	**European Union Citizenship/Citoyenneté de l'Union européenne/Cittadinanza dell'Unione europea**	European Citizenship Citoyenneté européenne Cittadinanza europea
Peers, Steve	**'EU justice and home affairs law'**	**European legal area/Espace juridique européen/Spazio giuridico europeo** Cooperation in the fields of justice and home affairs/Coopération dans les domains de la justice et des affaires intérieures/Cooperazione in materia di giustizia e affari interni	Criminal justice-International cooperation-European Union Justice pénale-Coopération internationale-Union européenne Giustizia penale-Cooperazione internazionale-Unione europea
Pedrazzi, Marco	**'Individual guarentees in the European judicial area in criminal matters. Garanties individuelles dans l'Espace judiciaire européen'**	**Police and judicial cooperation/Coopération policière et judiciaire/Cooperazione giudiziaria e di polizia**	Police-International cooperation-European Union Police-Cooperation internazionale-Union européenne Polizia-Cooperazione internazionale-Unione europea

7 References

1. Alemu G., Stevens B., Ross P.: Towards a conceptual framework for user-driven semantic metadata interoperability in digital libraries: A social constructivist approach. In New Library World, vol. 113, iss.: 1/2, pp. 38-54. (2012)
2. Ballard T., Blaine A.: User search-limiting behavior in online catalogs: Comparing classic catalog use to search behavior in next-generation catalogs. In New Library World, vol. 112, iss.: 5/6, pp. 261-273. (2011)
3. Blyberg, J. : Beyond the OPAC: The Semantic Library. Michigan Library Consortium. (2007)
4. Buizza, P. : Gli OPAC: funzionalità e limiti nel mondo del web. In Bibliotime, vol. 1, year XI. (2008)
5. Choi, Y., Hsieh-Yee, I. : Retrieval effectiveness of table of contents and subject headings. In Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries, ACM Digital Library. (2007)
6. Gnoli, C. : BLOPAC semantici. In Bibliotime. number 1, year XI. (2008)
7. Markey, K. : The Online Library Catalog. Paradise Lost and Paradise Regained? In D-Lib Magazine.vol.13, number 1.2. (2007)
8. Mat Yamin, F.: An Overview of the Web Search Satisfaction. In Communications of the IBIMA. vol. 3. (2008)
9. Singer, R. : In Search Of A Really "Next Generation" Catalog. In Journal of Electronic Resources Librarianship. Vol. 20, Issue 3. (2008)
10. Shiri, A. A., Revie, C.W., Chowdhury, G. : Thesaurus-assisted search term selection and query expansion: a review of user-centred studies. In Knowledge Organization. 29 (1). (2002)
11. Veronesi, E.**,** Bertaccini, F. : Une « ontoterminologie » pour les interprètes de conférence. http://realiter.net/spip.php?article2042. (2010)