ColabTKR 2012 - Terminology and Knowledge Representation Workshop

Workshop Programme

14:00 – 14:15
Introduction by the Workshop Chairs

14:15 – 14:45

Michael Wetzel, Elena Chiocchetti, Tanja Wissik, *Putting Together Apples and Oranges: The LISE Tool Suite for Collaborative Terminology Work*

14:45 – 15:15

Sérgio Barros, Rute Costa, António Lucas Soares, Manuel Silva, *Integrating terminological methods in a framework for collaborative development of semi-formal ontologies*

15:15 – 15:45

Gabriel Bernier-Colborne, *Defining a Gold Standard for the Evaluation of Term Extractors*

15:45 – 16:15 Coffee break

16:15 – 16:45
Gian Piero Zarri, *Mapping from Lexical Resources to High-Level Data Modelling Languages*

16:45 – 17:15
Cristóvão Sousa, António Lucas Soares, Carla Pereira, Rute Costa, *Supporting the identification of conceptual relations in semi-formal ontology development*

17:15 – 17:30
Conclusions

Editors

António Lucas Soares	University of Porto and INESC Porto, Portugal
Rute Costa	New University of Lisbon, CLUNL, Portugal

Workshop Organizers/Organizing Committee

António Lucas Soares	University of Porto and INESC Porto, Portugal
Rute Costa	New University of Lisbon, CLUNL, Portugal
Carla Pereira	IPP/ESTGF and INESC Porto, Portugal
Alessandro Oltramari	Carnegie-Mellon University, USA
Christophe Roche	University of Savoie, France
Anita Nuopponen	University of Vaasa, Finland

Workshop Programme Committee

Gerhard Budin	University of Vienna
Chiara Ghidini	Bruno Kessler Foundation (FBK) - Trento, Italy
Guadalupe Aguado de Cea	Universidad Politécnica de Madrid
Hanne ErdmanThomsen	Copenhagen Business School
Mustafa Jarrar	University of Birzeit, Palestine
António Lucas Soares	University of Porto and INESC Porto, Portugal
Rute Costa	Universidade Nova de Lisboa, CLUNL, Portugal
Carla Sofia Pereira	Polytechnic Institute of Porto and INESC Porto, Portugal
Alessandro Oltramari	Carnegie-Mellon University, USA
Christophe Roche	University of Savoie, France
Anita Nuopponen	University fo Vaasa, Finland
Piek Vossen	VU University Amsterdam, Netherlands

Table of contents

Author Index

Preface

Linguistics and ontology studies have a long record of fruitful cooperation. Cross-research in areas such as computational linguistics, natural language processing, information retrieval and ontology development, maintenance and integration have produced a wealth of multidisciplinary theories, methods, models and tools (Roche, 2008) (Staab, 2008) (Costa & Silva, 2008) (Pereira et al. 2009) . More specifically, the relationship between the lexicon (lexical approaches and resources) and ontology development methods and tools, have been recently well explored in research (Huang et al., 2010). On the contrary, the relationship between terminology and ontology studies, in particular in what concerns to the initial phases of ontology development, has not received so much attention from the scientific communities involved.

Furthermore, in diverse professional areas, new challenges are appearing related with information and knowledge management in highly specialised technical domains, under tightly constrained time requirements, unfolding in collaborative networking contexts. Short-term collaborative networking between individuals, groups and organisations, is recognised by researchers and practitioners as possible solution to cope with an increasingly complex social and economic business environment. Moreover, the current demand for continuous innovation leads to an higher heterogeneity in the technical and scientific domains simultaneously involved in collaborative projects and activities (e.g involving SMEs and research centres) (Camarinha-Matos, 2006). Managing information and knowledge in this context, places new and interesting challenges to terminology and knowledge representation, particularly when these challenges are seen from an integrated terminology/knowledge representation perspective.

Terminological or ontological approaches alone are not likely to be enough in answering to the needs of precision and detail of the specialised technical domains, as much as the research efforts of articulated terminology/ontology approaches are likely to be inadequate in terms of the required resources (time and persons). Thus, these challenges call for more than the setup and configuration of common terminological or ontological resources, particularly when considering the usually accepted time-frames for developing semantic and terminological artifacts. Effective ways to collaboratively construct shared conceptualisations by the means of negotiation and representational artifacts, such as semi-formal ontologies, are then required.

The above problems and difficulties motivate challenging multi and transdisciplinary lines of research in particular where terminology and knowledge representation meet together with a double aim: to collaboratively study the phenomena from cross-perspectives and to produce practical artifacts for professional work in these two areas. This was the motivation for creating the colabTKR - Collaboration in Terminology and Knowledge Representation - workshop where terminology, information/knowledge management, ontology development, and collaboration specialists join to debate and share from problematic theoretical issues to proposals for innovative approaches. ColabTKR main subject - the interplay between terminology and knowledge representation methods and techniques in contexts of collaborative work - encompasses research in topics such as collaborative processes in terminology work, collaborative conceptualization processes and representations of knowledge, multimodal corpora for semi-formal ontology development, theory, methods and tools for conceptual negotiation, interfaces between terminology work and ontology development/maintenance.

In this workshop five papers dealing with different approaches to the collaboration within and between terminology and knowledge representation are presented, three of them describing methods and results obtained in two different projects: LISE project (http://www-lise-termservices.eu) and CogniNet (http://cogninet.tk/).

In the first case, as the authors Michael Wetzel, Elena Chiocchetti, Tanja Wissik, explain in their abstract, the LISE project aims at improving the quality of existing terminology collections and at facilitating the consolidation of administrative nomenclatures and legal terminology. To that purpose, tools and best practices are developed to enhance interoperability and cross-border collaboration, thus offering specific tools to assist the terminological workflow and also a platform to discuss and exchange data.

In the second case, a collaborative platform - ConceptME - was developed under the project cogniNET, a project addressing problems raised by information and knowledge sharing in the context of short life-cycle collaborative networks. The tool provides support to domain experts engaged in activities related to a shared conceptualization. Two presentations were held held regarding ConceptME, as part of the research developed by António Lucas Soares, Rute Costa, Carla Pereira, Sérgio Barros, Cristóvão Sousa, Manuel Silva. The first one deals with the support to the identification of conceptual relations during the development on semi-formal ontologies. The second one describes the integration of a terminological method to support experts in eliciting and organizing concepts of their domain.

In another presentation, Gabriel Benier-Colborne describes a methodology to define a gold standard (fully annotated corpus) for the automatic evaluation of term extractors that he considers relevant to evaluate protocol for term extraction systems.

Finally, Gian Piero Zarri presents a modelling and development tool – NKRL - bringing to discussion the theoretical and practical problems of transferring lexical information to ontological and knowledge-based systems.

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
António Lucas Soares
Rute Costa

Putting Together Apples and Oranges:
The LISE Tool Suite for Collaborative Terminology Work

Michael Wetzel, Elena Chiocchetti, Tanja Wissik
ESTeam AB; Institute for Specialised Communication and Multilingualism, EURAC; Zentrum für Translationswissenschaft, Universität Wien
Rungestraße 20, Berlin; Viale Druso 1, Bolzano; Gymnasiumstraße 50, Wien
E-mail: michael@esteam.se, echiocchetti@eurac.edu, tanja.wissik@univie.ac.at

Abstract

Different terminology databases contain different types of information or a diverging depth of information. To create more complete resources, it might be useful to add languages to existing collections and/or merge (part of) some terminology repositories. This being a daunting task in terms of time and staff efforts, tools allowing the semi-automatic processing of data when adding languages, cleaning termbanks from multiple entries or harmonising terminology collections would facilitate this task. The LISE project (http://www.lise-termservices.eu) aims at improving the quality of existing terminology collections and at facilitating the consolidation of administrative nomenclatures and legal terminologies. It develops tools and best practices to enhance interoperability and cross border collaboration. The main purpose is to help terminology managers in public institutions or private service providers and companies improve the coherence and completeness of their terminological resources in a more efficient way. LISE offers specific tools to assist the terminology workflow, but also a platform to discuss and exchange data. The scientific basis of the project rests in a deep insight into terminology workflow best practices, so as to understand at what point in time each specific tool might be usefully applied.

Keywords: terminology workflow, terminology tools, terminology databases, terminology harmonisation

1. Background

Every terminology database has its own peculiar objective, content, history and target users. As a consequence, the data models and entry structures might vary a great deal, even between term banks that present similar types of information (cf. Melby, 2008, Wissik, 2012). In a united Europe and globalised world it might often be sensible to add new languages to existing collections or even try and bring different repositories together, harmonising and merging the collected entries. This is true especially in the field of law and administration, which are domains of paramount importance for international collaboration. Up until now, adding new languages, merging existing terminology resources (cf. Nesculescu et al., 2011) or importing new data, as well as cleaning newly created databases from double and triple entries, meant a lot of manual work.

The LISE project (http://www.lise-termservices.eu) aims at improving the quality of existing terminology collections and facilitating the consolidation of administrative nomenclatures and legal terminologies. It develops specific tools and best practices to enhance interoperability as well as interinstitutional and cross border collaboration. The main purpose is to help terminology managers in public institutions or private service providers and companies improve the coherence and completeness of their terminological resources in a semi-automatic and hence more efficient way. To achieve this, three tools (cf. 3.1) are being further developed within the LISE project to assist and partly automatize specific steps of the terminology elaboration workflow (cf. 2.)

2. Terminology Workflow

The basic terminology workflow foresees a series of steps around the core activity of terminology elaboration (cf. KÜDES, 2002). The expression of need triggers the search for relevant documentation, i.e. the textual basis from which to extract and select terminology (cf. Ralli & Stanizzi, 2008). Terminology extraction and selection, i.e. the creation of a list of terms that will then be described in fully-fledged terminology entries, are processes that can be carried out manually or semi-automatically with dedicated terminology extraction tools. When this task is done manually, extraction and selection may partly take place at the same time, whereas automatic term extraction always requires a further step in which the relevant terminology is selected according to the specific project aims and needs. If aligned corpora or translation memories are available, it is possible to automatically retrieve also equivalents in one or more languages. Manual term extraction ensures high-quality work, but can be extremely time-consuming. A semi-automatic pre-selection with a subsequent human check may lead to good results in less time.

The selection of terms, as well as the subsequent elaboration of term entries, are usually performed by human staff with different approaches. It can be systematic, i.e. domain related on the basis of a concept tree, or text related, especially if the main aim is to support the translation process with proactive terminology work (Wright & Budin, 1997). Also ad hoc approaches are common, when the selection of terms follows some specific needs expressed. In this last case, the terms selected for elaboration might not cover an entire domain or text, but rather be 'scattered' terms following an absolutely practical and request-based approach. Before proceeding to the next step, the resulting list of terms to be treated can be formally validated (e.g. by a project manager, domain expert(s), contractor, etc.).

The core activity of the entire workflow, i.e. elaborating the terminology entries, may range from creating a collection of equivalents in two or more languages to providing a large set of information for each term in every language considered, e.g. domain, grammatical aspects, a definition, a context of use, sources, various indications on usage, restrictions of equivalence, reliability, and many more categories of information.

The number of data categories selected as well as the way each category is linked to the others in the data model may differ greatly from termbase to termbase, because it depends on the languages considered, the purpose, the target users and other factors (cf. Budin, 2010). Monolingual or multilingual entries can be elaborated from scratch, existing entries can be updated, amended and integrated, but it is also possible to import data and integrate them into an existing term collection. During import, the issues concerning often differing data models, import/export format, loss and reduplication of data are manifold and sometimes so complex, that the idea is not rarely given up soon. For example, the risk of creating double entries in the termbase is higher when importing a batch of external data (cf. Nesculescu et al., 2011).

As a consequence, automatic import of data triggers a high demand of revision work, which is the process that usually follows the elaboration of terminology entries. The completeness of data, their formal correspondence to the requirements of the termbase, the systematic coherence with other entries in the collection must be verified. A particularly important step is the elimination of duplicates and the merging of similar entries, so as to avoid misleading the end user. Finally, as part of quality control, the content might be checked by one or more experienced terminologists or subject field experts to approve the entries, with a particular focus on the selected synonyms, the content of the definition and of the possible usage notes or additional explanations.

As a last step, sometimes (selected) terms or translation equivalents are standardised, i.e. officially validated by a dedicated body (cf. TERMCAT, 2006; Chiocchetti & Stanizzi, 2009). The resulting data is often shared internally or publicly, even more so if it has been standardised. Terminology collections are made available, for example: online in specific data bases (e.g. IATE, the Interactive Terminology data base of the bodies of the European Union, cf. http://iate.europa.eu), company intranets (e.g. Volkswagen, cf. Bernardi et al., 2005), paper publications (e.g. the terminological dictionaries published by TERMCAT), etc. Standardised terminology is always disseminated in some way, for example in standards (e.g. ISO standard TC 20/SC 8 on aerospace terminology) or via official, sometimes legally binding means (e.g. the Official Journal of the Region Trentino/Alto Adige in Italy, which publishes the lists of equivalents officially validated by the local Terminology Commission; cf. p. 19 http://www.regione.taa.it/bu/2010/BO011001.pdf).

The complete workflow might be schematised like in figure 1:

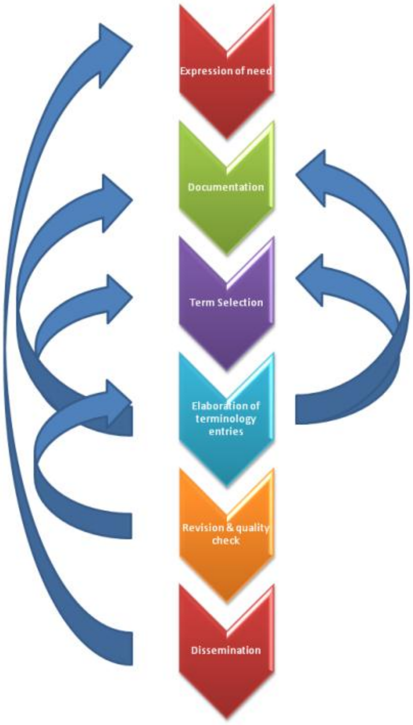


Figure 1: terminology workflow

3. The LISE terminology service

The LISE service supports maintaining large volume terminology resources together with a collaboration platform (cf. 3.2.). The goal is to achieve interoperability across terminology in the same domain and in one or more languages.

Market research carried out during the LISE project leads to a clear understanding about the lack of terminology process/workflow modelling and support technologies, making terminological resources inefficient to be applied consistently in different applications and most importantly for harmonisation. Harmonising and maintaining large termbases manually is quite often a hopeless endeavour. Based on ESTeam's experiences, when the amount of entries reaches a number of ~2,000, it is already no longer possible to maintain and fully supervise it via human activity, by scrolling through the terms or by searching and filtering. Only software with sophisticated linguistic algorithms can do this.

Three logical steps are identified to improve the quality of terminological resources with the help of the tools:

1) Remove errors and inconsistencies: spelling mistakes, wrong meta data assignments
2) Fill-in missing languages: some entries contain all required languages, some entries are not completely covered
3) Conceptualisation: data that belongs together is wrongly kept in different entries

Furthermore, terminology activities are usually not “silo” processes, but tend to leave the boundaries of a department or even an organisation. Hence, the above mentioned cleaning steps require a smooth involvement of any stakeholder that is called to contribute to the improvement of the quality of terminological data. An online collaboration portal can be a method to address this need.

The LISE Terminology Service combines a solution for all above listed requirements plus a human backed expert service that provides consulting and the customisation of technologies for appropriate data processing.

3.1. LISE Terminology Tools

The LISE project provides the applications Cleanup, Fillup and OMEO[1], that can be used at different points in time to assist and speed up some steps of the terminology workflow.

Cleanup intends to automatize the process of eliminating double and triple entries in large term collections, thus supporting the revision process. Since this is a step which results in reducing the amount of data, it is recommended to make it the first step.

Fillup can be used to automatically integrate one or more new languages on the basis of aligned translation memories. It therefore supports the process of term extraction and selection in the target language, thus speeding up the process of terminology elaboration, because the possible equivalents in the target language are already proposed by the tool.

OMEO helps reviewing the data to harmonise and streamline terminologies when merging different term collections. This results in a more complete collection of terminology, which can be harmonised semi-automatically. OMEO displays different discovered term variants for one concept, for instance variations in spelling like “eye glass” and “eyeglass”. It compares units in one language to find alternatives that share the same meaning but are written differently. Users would then *accept* or *reject* each of the variations, thus clearly determining the preferred variants.

At the Office for Harmonization of the Internal Market (OHIM) the tools have helped to clean and harmonise the trademark and classification terms in all EU languages, allowing a service offering like OHIM's EuroClass

[1] Developed by LISE partner ESTeam (http://www.esteam.se).

(http://oami.europa.eu/ows/rw/pages/QPLUS/databases/euroclass.en.do).

Generally, terminology experts and data category specialists analyse existing terminological data and import it into ESTeam language servers that are tuned for processing large volumes of data. Scanning, matching, and comparison algorithms then process the data and prepare the results for final human review in the above mentioned applications.

Seen from a more global view, it becomes more and more visible that unclean "master data" is a general problem in the data centres of enterprises and organisations – typically occurring in address book records in customer relationship management (CRM) systems, product attributes in large product information management (PIM) systems, or wrong number values in spare part systems etc. Some software vendors start addressing this and enterprise resource planning (ERP) providers have this on the horizon. Yet, no solution was available for terminological data so far.

3.2. LISE Collaboration portal

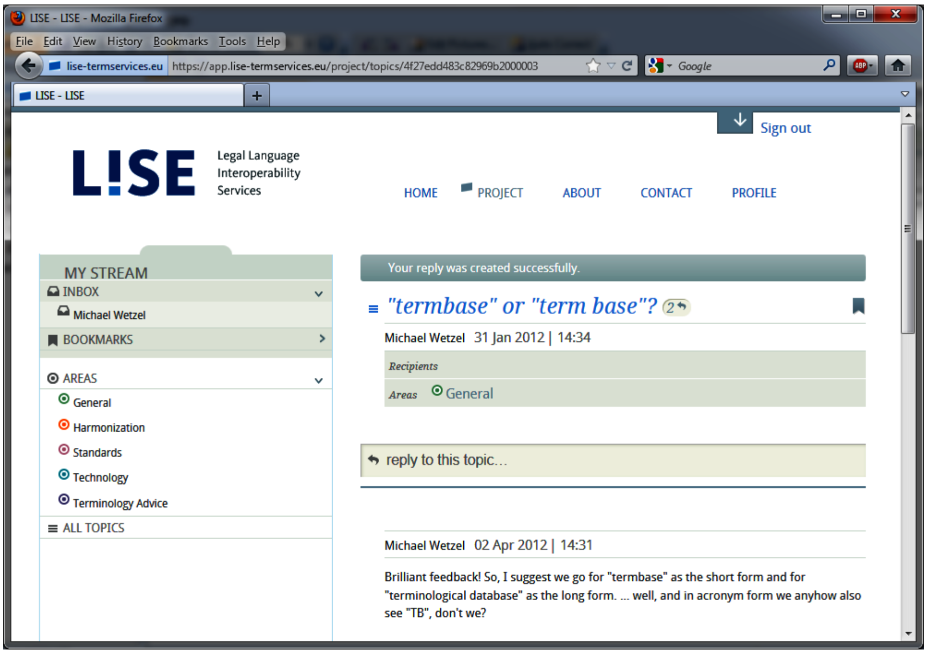
Besides the linguistic tools, the LISE service is equipped with a collaboration portal (cf. Wetzel, 2012), a common point where all stakeholders, be they from the same or from different organisations, can communicate about the terminology resources, inspect the results created by the LISE tools, discuss reviews and drive new activities. When all contributors come from different backgrounds, a common ground of knowledge or training cannot be assumed. Therefore, the collaboration portal is designed with user friendliness as the top goal in mind; it is as easy-to-use as consumer applications like Google+ or Facebook.

When logged in, the user sees all accessible posts listed in a single stream sorted by currentness. Recent activities are easily scanned at first glance. In a side column we see the work groups and subscriptions of the current user. Selecting any of these items will filter the stream to display only related posts. This makes navigating the different streams and focusing on a specific topic a very straightforward task. The fact that all related files and documents are simply attached to a post makes them directly accessible in the context of the relevant information. Several versions of the same document – for instance a set of terminology entries – can be kept together and be easily identified by date and context.

3.2.1. Functionality Outline

The LISE Collaboration Portal (see figure 2) is – state spring 2012 – in development and will be finished in the first months of 2013. Typical functionalities already available upon writing this article include:

- **Create topic**: Start a new topic with a post to inform users about the availability of a new resource, a Cleanup result file or any other discussion or question calling for input and feedback from other people.
- **Add recipients**: Adding a recipient guarantees that the recipient sees the topic very prominently in his *Inbox*.
- **Define privacy**: The creator of a topic can *lock* it, so that only specified recipients can see the topic. Both, the list of recipients as well as the privacy lock cannot be changed *a posteriori*. Thus the visibility is defined by the topic creator and can never be changed or overwritten later.


LISE - LISE - Mozilla Firefox
File Edit View History Bookmarks Tools Help
lise-termservices.eu https://app.lise-termservices.eu/project/topics/4f27edd483c82969b2000003
Google
LISE - LISE
Sign out
LISE
Legal Language Interoperability Services
HOME
PROJECT
ABOUT
CONTACT
PROFILE
MY STREAM
INBOX
Michael Wetzel
BOOKMARKS
AREAS
General
Harmonization
Standards
Technology
Terminology Advice
ALL TOPICS
reply to this topic…

Your reply was created successfully.

"termbase" or "term base"? 2

Michael Wetzel 31 Jan 2012 | 14:34

Recipients

Areas General

Michael Wetzel 02 Apr 2012 | 14:31

Brilliant feedback! So, I suggest we go for "termbase" as the short form and for "terminological database" as the long form. ... well, and in acronym form we anyhow also see "TB", don't we?

Figure 2: LISE Collaboration Portal: A topic with one reply, discussing a spelling question

- **Reply to a topic**: Obviously, users can reply to a post.
- **Areas**: Filing a topic under an *Area* keeps topics together. This guarantees easy, structural navigation when many topics are live.
- **Bookmarks**: Bookmarks make topics or even whole areas more prominent. They are listed on top of a user's own stream, hence are quickly to navigate to.
- **Add attachments**: Users can add file attachments to a topic, so that the referenced resource is directly available. This eliminates the requirement of delivering data or files via a separate channel like ftp or email. All is kept together.
- **Create links to master terminology data**: For supported external terminology management systems, a user can create a direct link to a relevant entry; a topic or a post can then directly reference and provide a pointer to the affected term entry.
- **Voting**: A special type of topic is an invitation to *vote* sent to other users with the aim of finding a majority agreement on a specific question.

3.2.1. Technical Implementation

The LISE Collaboration Portal is a modern rich web application while, on the technical side, nicely avoiding any browser plug-in, like Java or Microsoft Silverlight. This guarantees that no IT restrictions imposed by one of the contributing stakeholders may block the process and success of the project.

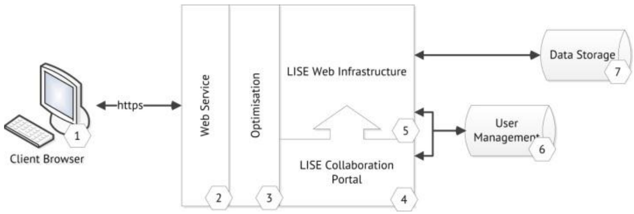


Figure 3: LISE Web Platform Architecture

As shown in figure 3 a client browser (1) establishes a secure https connection to a REST based (2) web service. Through a performance optimisation layer (3) it interacts with the collaboration portal (4), which is a major module of the LISE Web Infrastructure (5). User accounts are stored via OpenLDAP (6) technology and data is stored in a dedicated database. The technology selection was driven, inter alia, by the priority to be non-proprietary, scalable, and standards compliant.

3.3. Workflow Outline

How does this all play together, how does the LISE master workflow look like? Six major steps characterise a LISE web service driven workflow:

1) A user or an organisation applies for membership in the LISE Collaboration Portal. The account is set up. This is required only once per user or organisation. ESTeam AB administers and hosts the site and is responsible for the up-time of the service.
2) The member now delivers terminology and translation memory data for processing, together with a metadata description; the processing takes place on this basis.
3) Estimation Phase
    a. The ESTeam Terminology Service team reviews the data and estimates the effort of getting the LISE Tools customized so as to be able to process the new types of data categories.
    b. The member approves the effort and estimated delivery.
4) The data is processed with ESTeam language server technologies and prepared for above mentioned tools.
5) The processing results are made available to the member, together with the LISE Tools client software, for the member to process and post-edit the results. Terminology Expert Services may contribute here. Users discuss the result and analyse files using the collaboration portal.
6) Finally, target databases are updated. This depends on where the terminology master resources are kept.

Altogether, the service

- facilitates accessibility to high quality terminology resources in different domains and languages
- supports dissemination of best practices on how to use one's own terminology repositories
- allows to handle the diversity of coding schemes and data organisation
- improves cross-lingual and cross-domain interoperability (syntactic, semantic, pragmatic) across existing technical applications
- helps handling cultural differences across language communities and domains (administrative and legal language being the best example).

4. Conclusions

Different terminology databases contain different types of information or a diverging depth of information. To create larger resources with a more complete amount of data, it might be useful to add languages to existing resources or merge (part of) some terminology collections. This being a daunting task in terms of time and human resources, having a few tools that allow semi-automatic processing of data when adding languages, cleaning termbanks from multiple entries and harmonising terminology collections, is of great help in encouraging collaboration and data exchange or in enhancing the quality of terminology resources.

The combination of insights into terminology workflow best practices plus an inter-departmental or even inter-organisational collaborative approach to enhance and clean terminology resources is without any predecessor.

5. Acknowledgements

The LISE project has received funding from the European Community (ICT-PSP 4th call) under Grant Agreement n° 270917.

6. References

Budin, G. (2010). Socio-terminology and computational terminology – toward an integrated, corpus-based research approach. In: R. de Cillia, H. Gruber, M. Krzyzanowski, F. Menz (Eds.), Diskurs – Politik – Identität. Discourse – Politics – Identity. Tübingen: Stauffenburg, pp. 21-31.

Bernardi, U., Bocsak, A. & Porsiel, J. (2005). Are we making ourselves clear? Terminology Management and Machine Translation at Volkswagen. In EAMT 2005 Conference Proceedings. Budapest, 41-49.

Chiocchetti, E. & Stanizzi, I. (2009). Kriterien zur Normung und Harmonisierung von mehrsprachiger Rechtsterminologie. In S. Šarčević (Ed.), Legal Language in Action: Translation, Terminology, Drafting and Procedural Issues. Zagreb: Nakladni zavod Globus, 167-182.

Chiocchetti, E. & Wissik, T. (2012). Zusammenführen und Harmonisieren von rechtsterminologischen Datenbeständen: Das LISE (Legal Language Interoperability Services) Projekt stellt sich den Herausforderungen kollaborativer interinstitutioneller Terminologiearbeit. In E. Schweighofer, F. Kummer & W. Hötzendorfer (Eds.), Transformation Justischer Sprachen. Tagungsband des 15. Internationalen Rechtsinformatik Symposions (IRIS2012). Salzburg: Österreichische Computer Gesellschaft, 261-268.

KÜDES = Konferenz der Übersetzungsdienste Europäischer Staaten (Eds.) (2002). Empfehlungen für die Terminologiearbeit. 2nd ed. Bern: Schweizerische Bundeskanzlei.

Melby, A.K. (2008). TBX-Basic. Translation-oriented Terminology Made Simple. Tradumàtica, 6.

Nesculescu, S., Bel, N., Padró, M., Marimon, M., & Revilla, E. (2011). Towards the Automatic Merging of Language Resources. In IJCNLP2011 - Proceedings of the Workshop on Language Resources, Technology and Services in the Sharing Paradigm. Chiang Mai.

Ralli, N. & Stanizzi, I. (2008). Il dietro le quinte della normazione. In E. Chiocchetti & L. Voltmer (Eds.), Normazione, armonizzazione e pianificazione linguistica/ Normierung, Harmonisierung und Sprachplanung/ Normalisation, harmonisation et planification linguistique. Bolzano: EURAC, 61-74.

TERMCAT (2006). La normalització terminological en català: criters i termes 1986-2004. Barcelona: Abadia de Montserrat.

Wetzel, M. (2012). LISE Web Service: An Online Collaboration Portal to Facilitate Legal Language Harmonisation. In E. Schweighofer, F. Kummer & W. Hötzendorfer (Eds.), Transformation Justischer Sprachen. Tagungsband des 15. Internationalen Rechtsinformatik Symposions (IRIS2012). Salzburg: Österreichische Computer Gesellschaft, 259-260.

Wissik, T. (2012): International, national and regional legal terminology: challenges and perspectives for a Legal Language Interoperability Service. In L. Alekseeva (Ed.), Proceedings of the 18th European Symposium on Language for Special Purposes (LSP). Special Language and Innovation in Multilingual World. Perm: Perm State National research University, 282-297.

Wright, S.E. & Budin, G. (1997): Introduction. In S.E. Wright & G. Budin (Eds). Handbook of Terminology Management. Vol. I, Basic Aspects of Terminology Management. Amsterdam, Philadelphia: John Benjamins, 1-12.

Integrating terminological methods in a framework for collaborative development of semi-formal ontologies

Sérgio Barros$^{(1,2)}$, Rute Costa$^{(1)}$, António Lucas Soares$^{(1,3)}$, Manuel Silva$^{(2,4)}$

[1]CLUNL - Universidade Nova de Lisboa, Avenida de Berna 26-C 1069 - 61 Lisboa;
[2]INESC Porto, Rua Dr. Roberto Frias, s/n 4200, Porto-Portugal;
[3]Department of Informatics Engineering, Faculty of Engineering, University of Porto, Rua Dr. Roberto Frias, s/n 4200, Porto-Portugal;
[4]ISCAP-IPP - Rua Jaime Lopes Amorim, s/n, 4465-004 S. Mamede de Infesta

E-mail: sergio.barros@fcsh.unl.pt, rute.costa@fcsh.unl.pt, als@fe.up.pt, mdasilva@iscap.ipp.pt

Abstract (10-point Times New Roman bold, centred)

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. To tackle this issue a collaborative platform was developed where terminological and knowledge representation processes support domain experts throughout a conceptualization framework.

In this article we describe the integration of a terminological method to support experts in eliciting and organizing concepts of their domain. The method is based on a linguistic analysis of textual resources with the help of a term extraction tool and by highlighting markers of relations between concepts. An application scenario is then presented to illustrate the connection between the terminological processes and the knowledge representation processes without blurring the theoretical distinction between terms and concepts.

Keywords: terminology, collaborative network, knowledge representation

1. Introduction

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. (Pereira & Soares, 2008:613) 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 modeling face difficulties when reusing ODPs for their needs» (Aguado de Cea, G. et al., 2008:45).

In the light of this issue, 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 (e.g. Felber 1984). Considering this state of affairs, the challenge lies precisely in maintaining the premise of "terms as linguistic expressions of mental and abstract units, the concepts" throughout the conceptualization process.

In a related 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.

In this article we describe the integration of terminological methods in this tool to assist experts in the discussion and modelling of the concepts of their domain.

2. Terminological framework

Terminology is a knowledge-related discipline whose object of study is the concept. From this perspective, since a collaborative conceptualization is developed around concepts, domain experts engaged in the collaborative process and terminologists focus on the same object. Nevertheless, while the former use terms and concepts for communicative and knowledge sharing purposes the latter study them in order to facilitate communication between experts in specialized domains or to enhance interoperability between information systems.

This twofold positioning implies that terminological methods must be accommodated to a particular communicative setting depending on an application, in this case a collaborative platform, enabling the construction of semi-formal ontologies.

To develop the work carried out in Terminology, either for

[1] http://cogninet.tk/

human use or machine applications, the use of texts as a resource is a common procedure. There is, nevertheless, the question of how to approach and use the text when our theoretical perspective is conceptually-based (in the line of Wüster) and the information written in the text is of linguistic origin. It is on this double dimension, linguistic and conceptual, that the method which supports the collaborative platform conceptME is based.

The platform conceptME is a technological space that allows the user to create and share conceptual systems resulting from conceptualization processes, collective or individual, which the user accepts/wants to share with a set of partners, in order to discuss and negotiate them. In these contexts, the use of natural language is unavoidable, although it carries with it, by definition, a great number of ambiguities and imprecision, characteristics that one should avoid in any negotiation process.

3. Overview of the conceptME method

The conceptualization framework in the platform is structured in four phases: concept elicitation, concept organization, concept sharing and concept discussion (Cristóvão et al., 2012). 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.

In terminology work, text is a relevant resource since it works as a repository that gathers linguistically structured information, from which we highlight terms and linguistic markers that play a central role in the method described in this paper. Since conceptME is aimed at domain specialists, presenting them the terms and linguistic markers that specifically occur in reference texts of their professional environment equals to offering them a key to access knowledge that, in theory, they already own.

In the following sections we describe the terminological processes that support the conceptualization phases of eliciting and organizing concepts:

i. Analysis of textual and terminological data so as to display it in a structured way in the platform structure;

ii. Definition of an hypothesis (an application scenario) based on structured information, that allow experts to choose the conceptualization path that better suits their needs.

4. Text: a repository of terminological information

The status and the role of specialized texts have been studied by (Costa, 2001; Costa, 2006; Costa & Silva, 2008). Specialized texts may, simultaneously, be understood as a production and a product of a restricted communication community, either professional or scientific. The text concentrates all the linguistic elements that designate and point to extra-linguistic elements that result from the interaction between language and social life, which allows one to analyze texts both as a process and as a result (Costa, 2006:80).

Terms designate concepts which in professional contexts, specific domains or for a given purpose, form conceptual systems portraying the knowledge that individuals produce and understand, in specialized texts of specific subject fields. There are, necessarily, intersections between objects, their representation and their designations. To acknowledge this triangular relation which encapsulates beliefs, scientific ideologies and a vision of the world, authors build discourses with a mono-referential value, in given contexts and for themselves. In a specialized communicative situation, authors must limit in discourse, as much as possible, the diversity of meaning constructions so as to come closer to a discourse that will ideally have one meaning, without ambiguities. Such discourses will probably never be reached and their existence is highly difficult to prove.

Given that all discursive acts (written or oral) are reflected in texts and involve complex cognitive, linguistic and social processes, a terminological and linguistic analysis of specialized texts helps to pinpoint conceptual structures behind linguist structures. As a result, when integrated in the platform, terms and markers of lexical-semantic relationships support users in their proposals of semiformal representations, thus bridging the gap between terminology and knowledge representation.

Although knowledge has an extra-linguistic nature, it is through the discourse that in most cases one is able to reach knowledge and its representations. Words are privileged means to represent knowledge. The difficulty in theorizing about it lies in the fact that those two realities – the world and its discursive representation – create a durable and reciprocal relation.

This context calls for a closer look at the description and characteristics of the specialized text as a result, i.e. a repository, as it becomes an object of observation and analysis for those who use texts to identify terms and other terminological information necessary for conceptualization. From this perspective arises the need to manage data found in texts, which in its turn, requires the management of texts as objects of knowledge, prior to analyzing their content. In view of these requirements it is necessary to create a typology of texts.

4.1 Collecting, identifying and classifying resources

When compiling a specialized corpus, one has to rigorously select a certain number of texts in the specialized domain, which will then become the objects of analysis. Such a process leads the researcher to ponder the parameters underlying the selection, organization and systematization of the texts that will constitute his/her corpus of reference.

Previous work focused on the issue of typologies (Costa, 2006), which presupposed the classification of a series of texts organized under the same name. To that purpose texts must maintain among themselves similarity relations at the micro- and macro-structural levels through the identification of regularities which are proper to a set of texts, as opposed to regularities of another set of texts.

A typology is the result of an organization of texts based on characteristics that are common to them, which makes the classification possible. This classification allows a systematic distribution of texts in groups or types to which we attribute a label or a generic name. This grouping, which is always artificial and depends on the goals of the research and the point of view of the researcher, may take into account either linguistic or extra-linguistic factors.

A typology does not presuppose, thus, any form of hierarchy, dependency or semantic or conceptual relation between the objectives that comprise it. A typology can be built from genres or types of texts. To Maingueneu, classifying texts into types is a sociological rather than a linguistic activity, while the genre constitutes the verbal action: « Les genres de discours relèvent de divers types de discours, associés à de vastes secteurs d'activité sociale » (Maingueneau, 1998:47). For the author, constructing discourse typologies is pertinent only if you take into account the genre, founding concept of the verbal activities: « Tout texte relève d'une catégorie de discours, d'un genre de discours » (Maingueneau, 1998:45).

To talk about types of discourse means to establish parameters that are congruent with the different sectors of society, as each one of them produces discourse and texts that can be classified under a specific typology. Scientific research, for example, is a sector whose textual and discursive production constitutes a type in itself, as it constitutes the product of a specific social activity. Therefore, we think that establishing type typologies, as well as genre typologies, results from the observation of the socio-discursive conditions under which the text was produced, given the fact that it is the representative witness of a collection of texts which, in its entirety, characterizes speech.

A text corpus from a specific domain is ideally made up of texts that correspond to a typological organization with the objective of creating a certain representativeness; this representativeness is not taken in the statistical sense, but rather in the sense of texts as scientific products recognized by the members of the professional or scientific community in which and for which the text was originally written. Only with the creation of such criteria is it possible to guarantee the compliance of texts with the pre-established objectives, which are obviously the guarantee of all research work.

4.2 Towards an operable typology

Taking into account the theoretical assumptions explained above the conceptME platform integrates a typology whose goal is to allow users the organization of the texts required to extract terminological information for the purposes of a conceptualization. The typology was proposed upon the detailed analysis of texts produced in the civil construction domain, more specifically in rehabilitation.

academic text
- master dissertation
- PhD thesis
- monograph
- report

specialized publication
- journal
- dossier

legislation
- law
- decree-Law
- ordinance
- contract

technical text
- technical sheet
- technical training
- textbook
- technical report

standard
dictionary
encyclopaedia

The proposal of the categories results from the resources compiled and identified so far, that is, based on the types of documents more frequently used by the target users of cogniNET, within the rehabilitation domain. Users of the platform can increment the typology since it is an open one, in case the types already specified don't suit the users' needs. In addition, users can select a more generic type in case the more specific one is not suitable to their needs. For example, a user may not know which category suits a given text but still knows that it belongs to the 'Legislation' category. Additionally, the possibility of conceptualizing via reference linguistic resources was also considered, namely dictionaries and encyclopedias.

This typology conforms to a repository where users can organize texts of their choosing into categories, thus allowing the compilation of a customized reference corpus. Such a corpus will be dynamic and up to date at all times.

5. Terminology extraction: a different goal

The semi-automatic treatment of corpus regards the process of terminological extraction as an initial step towards the elicitation of concepts. During this phase of conceptualization domain experts can use a terminological extraction functionality which allows them to obtain from a text or group of texts a list of linguistic units that potentially designate concepts. This functionality allows the selection of one or several of these suggested candidate terms with which concepts can be organized in the following phase of conceptualization.

In the beginning of the 90s, following the rapid development of computational linguistics and the widespread availability of corpora, terminology extraction became an important research interest as a means to reduce time and effort in different tasks related to different goals. (Cabré et al. 2001:53) identify several of the goals behind terminology extraction: «building of glossaries, vocabularies and terminological dictionaries; text indexing; automatic translation; building of knowledge databases; construction of hypertext systems; construction of expert systems and corpus analysis». The task of reaching a shared conceptualization in a collaborative framework can also benefit from the potentialities of these tools. When considered in terms of such a goal, it matters to reflect on the implications that a term extraction output has for a conceptualization phase that will be carried out by individuals who have a high level of knowledge in specific domain areas, thus, capable of identifying terms and concepts without necessarily making a difference between the linguistic and the conceptual level. The challenge behind the terminological extraction is to provide to experts terminological information which serve as a starting point for their conceptualization.

5.1 Criteria for selecting a tool

After reviewing a set of existing term extraction systems, three of them were selected for an evaluation: multiwords, TermoStat and GaleXtract. The first makes use of statistical methods and the other two use a hybrid method with the incorporation of a tagger with rules of the Portuguese language.

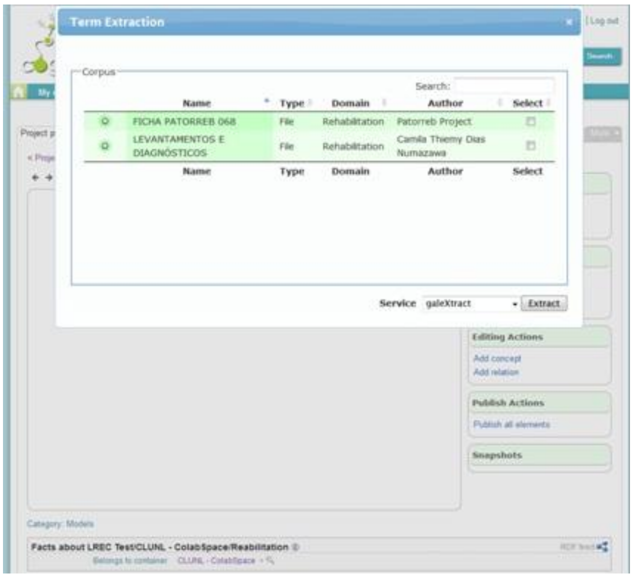
The terminology extraction methods are usually defined by linguistic and/or statistical criteria, which accounts for the linguistic dimension of terms. The possibility of extracting a set of linguistic units based on their frequency in connection with the recognition of language patterns typical of specific languages conforms to the main goal behind the evaluation of the extractors. Moreover, it bears also a connection with the requirements of an initial conceptualization activity: to obtain a list of acceptable linguistic units. In the light of these criteria hybrid methods of terminology extraction seem the most adequate for the platform: "Statistical approaches, like the linguistic ones, used alone only seldom reach truly satisfying results" (Pazienza et al. 2005:259). Furthermore, an extractor capable of accounting for several languages is preferable to a language-independent tool.

5.2 GaleXtract: description

Based on the evaluation criteria described above GaleXtract[2] was selected as the term extraction tool since it is based on a hybrid method, and it handles several languages: Galician, Spanish, English, French, Portuguese. Its extraction allows the use of either Freeling or Treetagger for the tagging phase. Furthermore, five statistical measures can be employed although only one is available in the collaborative platform[3].

6. Integration of GaleXtract in the platform

The terminological extraction process consists in automatically extracting term candidates from a single text or group of texts and then select one or several to work with.

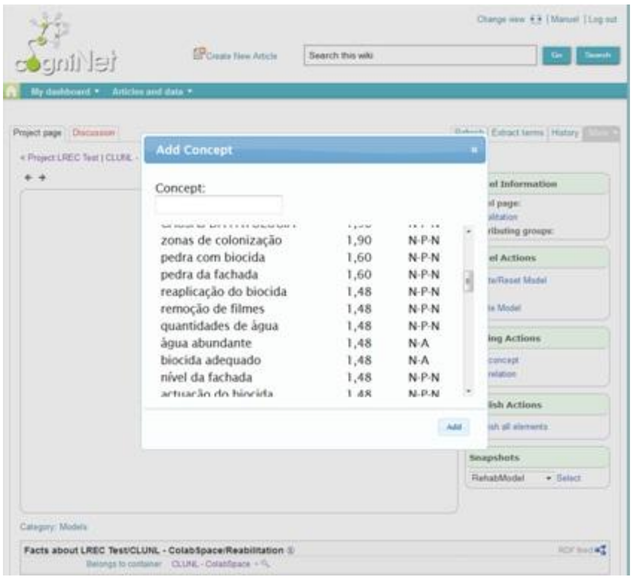


	Name	Type	Domain	Author	Select
○	FICHA PATORREB 068	File	Rehabilitation	Patorreb Project	□
○	LEVANTAMENTOS E DIAGNÓSTICOS	File	Rehabilitation	Camila Thiemy Dias Numazawa	□
	Name	Type	Domain	Author	Select

Facts about LREC Test/CLUNL - ColabSpace/Reabilitation

Figure 1: Term extraction from resource(s)

The output list that is presented to the expert can be sorted alphabetically or by ordering the results from the highest to the lowest statistically measured linguistic unit.



zonas de colonização	1,90	N-P-N
pedra com biocida	1,60	N-P-N
pedra da fachada	1,60	N-P-N
reaplicação do biocida	1,48	N-P-N
remoção de filmes	1,48	N-P-N
quantidades de água	1,48	N-P-N
água abundante	1,48	N-A
biocida adequado	1,48	N-A
nível da fachada	1,48	N-P-N
actuação do biocida	1,48	N-P-N

Facts about LREC Test/CLUNL - ColabSpace/Reabilitation

Figure 2: Term extraction result

[2] Developed under the Gari-Coter project: http://gramatica.usc.es/proxectos/Gari-Coter/?lang=gl.

[3] Although the measures of coocurrences, loglike, chi-square, mutual information and scp generated similar results chi-square was the measure chosen.

Nevertheless, tools are only a means to save time and effort in terminology work: «Terminology extraction implies, almost invariably, that whatever is initially collated is a collation of candidate terms» (Ahmad 1998:141).

From a terminologist's stance, the initial selection of acceptable term candidates follows linguistic criteria, by selecting combinations of words that match patterns that are typical of the Portuguese language: noun, noun + adjective, noun + preposition + noun, i.e. *colonização biológica* (*biological colonization*). Or by selecting combinations of words whose meaning is not the result of the sum of its parts, i.e. *filmes negros* (*dark spots*).

~~*intensidade nas zonas*~~
filmes negros
colonização biológica
~~*PALAVRAS-CHAVE Parede&Exterior*~~
~~*DESCRIÇÃO DA PATOLOGIA*~~
~~*CAUSAS DA PATOLOGIA*~~
rocha total
microscopia electrónica
~~*certa regularidade*~~
~~*sua aderência*~~
raios X
~~*Barroso de Aguiar*~~

Table 1: Sample of extraction.

One obstacle resulting from the semi-automatic method is that the output may not conform to the needs of experts, such as combinations of linguistic units that don't designate a concept or a conceptual unit, i.e. *certain regularity* (*certa regularidade*).

Although term extraction systems are useful to obtain lists of terms, a crucial methodological step consists in resorting back to their natural habitat, texts. Furthermore, since the terminological extraction process establishes a connection with the conceptualization phase where experts organize concepts something else is required in order to guide them towards relations between concepts.

6.1 Term candidates and lexical-semantic relations

Specialized texts are undoubtedly a vehicle of knowledge. In terminology, terms play a fundamental role as nuclear elements of lexical and semantic relations that language professionals or experts are able to recognize in texts. Such relations, held between the meanings of words, form the basis for the construction of semantic networks and allow the representation of the knowledge available in a text or set of texts.

Within the conceptualization framework designed for the conceptME platform the notion of knowledge representation covers several activities, namely the identification and selection of relations, the identification and selection of terms, the representation and consistency check of conceptual structures.

Recalling the motivations behind this research, the focus of integrating terminological methods in the platform is to establish a suitable and operable connection between the terminological processes and the knowledge representation processes as a means to support experts in the organization, sharing and discussion of concepts.

Considering the principles behind the terminological processes and how such principles relate to the knowledge representation processes involved in the platform, identifying potential terms during the concept elicitation phase must be complemented with a technique/method that allows one to understand not only how a given term is used but also the relation that it can have with other terms/concepts.

Since concepts can be expressed through linguistic forms, specialized texts are valuable sources of information for terminologists carrying out tasks related to concept analysis, like semi-automatic extraction of terminology or of relations between concepts.

Domain experts will also use specialized texts – previously selected by terminologists or by themselves – as a source of knowledge for their conceptualization tasks. Therefore, a natural step in our approach is to consider contexts as a source of information about concepts and about relations between concepts.

Following the work of (Hearst, 1992) several researchers developed the idea of extracting from texts linguistic patterns that express information about concepts, as contexts from a corpus of urban rehabilitation exemplify. For example, the structure is a typically expresses a relation between a subordinate concept and a superordinate concept:

> «A pre-dosed industrial mortar **is a** mortar whose components are dosed in the factory and supplied to the construction site, where they will be mixed according to instructions and conditions of the manufacturer»

The structure *is composed of* points to a partitive relation:

> «The floating floor **is composed of** laminated wooden boards arranged in opposite layers, so as to reduce the movement of the timber.»

The structure *X is caused by Y* expresses a relation between an effect and a cause:

> «The moisture is usually **caused by** the inadequate protection of the outer wall with respect to the atmospheric conditions to which it is subjected.»

Applications of this type include the writing of definitions (Pearson, 1998), concept analysis (Meyer, 2001), semi-automatic ontology building (Gillam, Tariq, & Ahmad, 2005) or the reuse of ontology design patterns (Aguado de Cea, Gómez-Pérez, Montiel-Ponsoda, & Suárez-Figueroa, 2008).

Based on the hypothesis that contexts such as these provide useful input to those who engage in a conceptualization process an application scenario related to the domain of civil construction, specifically rehabilitation, exemplifies how this terminological data can be applied.

7. Scenarios: an application

Considering the theoretical principles described above plus the criteria behind the terminological approach to the elicitation of concepts and the support to the concept organization phase, the integration of a terminological method in the platform is illustrated below.

A scenario implies starting a conceptualization with input, which consists of term candidates manually selected from the term extraction process, complemented with contexts with information about concepts, evidenced by the presence of linguistic markers.

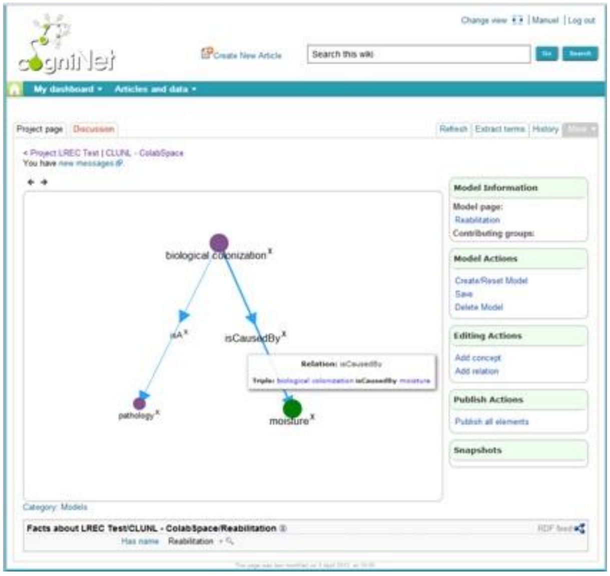
The first part of the application scenario draws on the first phase of the conceptualization framework, whose goal is to elicit concepts from textual resources**.**

patologia
colonização biológica
liquens
musgos
filmes negros
humidade
microscopia electrónica de varrimento
MEV
difracção de raios X
DRX
rocha total
biocida
nebulização

Table 2: Terminological input.

When compared with the raw output of the extraction tool these linguistic units illustrate the advantage of having a list of candidate terms that will serve to select designations of concepts to be organized in the following conceptualization phase.

The table above presents candidate terms which can trigger a conceptualization, i.e. a generic relation between the concept of biological colonization (*colonização biológica*) and that of pathology (*patologia*), a causal relation between the concept of humidity (*humidade*) and that of biological colonization. Relations such as these can be established by domain experts using a catalogue of concept relations that is available in the platform as a resource.



Relation: isCausedBy

Triple: biological colonization isCausedBy moisture

Facts about LREC Test/CLUNL - ColabSpace/Reabilitation

Has name Reabilitation

Figure 3: Possible conceptual relation(s)

As a means to support the concept organization phase, more specifically the use of the catalogue of concept relations, experts can consult contexts where those terms occur, thus obtaining further information about the respective concepts.

The objective behind the contexts, previously selected and filtered by a terminologist, is to call the attention of experts to the presence of markers of concept relations within contexts, which helps them to decide which type of conceptual relations exist between certain concepts.

Below we present a context with a linguistic marker of cause-effect relation between the concepts of biological colonization and moisture:

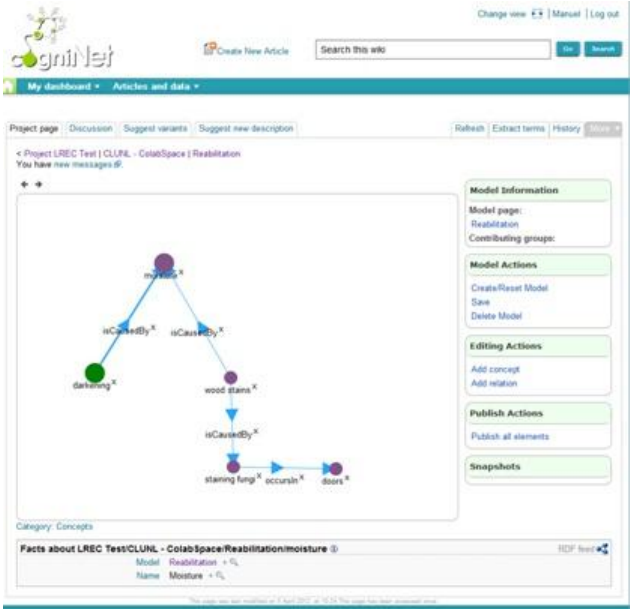
> «The biological colonization of the surface of the stone facade was mainly **due to** the presence of moisture, and there has been a greater intensity in areas where run-off are larger and darker on the front (north).»

Despite the potentialities of linguistic markers, research in the field of terminology has shown that their reliability is limited by factors such as their degree of dependency to the corpus (Meyer 2001, Condamines, 2002), their portability across different domains (Marshman, L'Homme, & Surtees, 2008) or the presence of uncertainty markers (Marshman, 2008).

An interesting example of such limitations is provided by the following context:

> «Darkening and wood stains **caused by** the presence of moisture and staining fungi most often located at the bottom of the door, **due to** lack of inclination of the sill with the accumulation of a water layer which penetrates inside the wood.»

The context above should give rise to causal relations such as the one between the concepts of moisture and that of wood stains or between the concepts of staining fungi and that of wood stains.



Facts about LREC Test/CLUNL - ColabSpace/Reabilitation/moisture

Model Reabilitation

Name Moisture

Figure 4: Possible conceptual relations

In addition to the relations modeled above this context is particularly interesting for a distinction between the markers *caused by* and *due to*, both causal but in principle expressing different types of causality that only experts can recognize. The marker *caused by* refers in principle to a causal agent of darkening and wood stains (*moisture*, *staining fungi*) and the marker *due to* possibly refers to its explanatory cause (*lack of inclination of the sill*).

Several authors studied the nature and number of concept relations (Feliu, 2004; Nuopponen, 2005, 2011; Sager, 1990). For example, (Nuopponen, 2011) has devised a model for cause-effect relations where she distinguishes various types of causes and of effects. Around the core concept of effect the author underlines different relations, i.e. a patient relation, a symptom relation, a consequence relation, a counteraction relation and a cause-effect relation. She also sees three types of effects (resulting product, resulting state, resulting event) and different possible causes (causal agent, producing cause, explanatory cause) (cf. Nuopponen 2011:12).

The author's perspective is: «Causal relation is often seen as a relation between the concepts of cause and effect (causal sequence), but this is only the basis for a complex concept system that is often involved» (Nuopponen, 2011:12). Some authors suggest that it is not very practical to have a very detailed account of concept relations (Madsen, Pedersen, & Thomsen, 2001:7). However, if the purpose is to negotiate meaning and clarify concepts then it may be a good idea to have a breakdown of the most general conceptual relations into more detailed ones such as the ones that (Nuopponen, 2011) proposes in her causality model.

From a terminologist's perspective it would be interesting to see whether users recognize and discuss the meaning of different markers of causal relations such as the ones that occur in the context above.

s of causal relations such as the ones that occur in the context above.

8. Concluding remarks

This article described the integration of a terminological method in a collaborative framework to assist the domain expert throughout the initial phases of a conceptualization process. More specifically, we focused on the integration of a tool in the platform to extract term candidates and on supporting the use of a catalogue of conceptual relations that will be available in the platform. The organization, share and discussion of concepts is supported by natural language, more specifically texts that provide the terms to designate concepts or the linguistic mechanisms to establish relations between concepts. Nevertheless, those texts contain ambiguities and uncertainties that experts may not recognize.

The hypothesis behind this method is that eliciting concepts from textual resources and identifying concept relations for conceptualization purposes can benefit from an approach that maintains a distinction between terms and concepts throughout the conceptualization process.

To obtain insights on the usability of the terminological method in the platform several scenarios with terminological data were prepared as application examples. Scenarios such as these are important not only to obtain an empirical insight on the connection between terminology and knowledge representation but also on the usefulness of contexts with markers of conceptual relations as a means to complement the use of the catalogue.

9. 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".

10. References

Aguado de Cea, G., Gómez-Pérez, A., Montiel-Ponsoda, E., & Suárez-Figueroa, M. C. (2008). 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.

Ahmad, K. (1998). Specialist Texts and their Quirks. In *TAMA '98: Proceedings of the 4th TermNet Symposium*. Vienna, Austria: TermNet, pp. 141-157.

Cabré, M. T., Estopà, R., & Vivaldi, J. (2001). Automatic term detection. A review of current systems. In D. Bourigault, C. Jacquemin, & M.-C. L'Homme (Eds.), *Recent Advances in Computational Terminology*. Amsterdam; Philadelphia: John Benjamins Publishing Company, pp. 53-87.

Condamines, A. (2002). Corpus Analysis and Conceptual Relation Patterns. *Terminology*, 8(1), pp. 141-162.

Costa, R. (2001). Pressupostos teóricos e metodológicos para a extracção automática de unidades terminológicas multilexémicas. PhD Thesis. Lisboa: Universidade de Lisboa.

Costa, R. (2006). Corpus de spécialité : une question de types ou de genres. In H. Béjoint & F. Maniez (Eds.), *De la mesure dans les terme. Hommage à Philippe Thoiron*. Lyon : PUL, pp. 313-324.

Costa, R. & Silva, R. (2008). De la typologie à l'ontologie de texte. In *Terminologie & Ontologies : Théories et Applications. Actes de la deuxième conférence - Toth Annecy – 2008*. Annecy : Institut Porphyre. Savoie et Connaissance, pp. 3- 16.

Felber, H. (1984). M*anuel de terminologie*. Paris: United Nations Educational, Scientific and Cultural Organization - International Information Centre for Terminology.

Feliu, J. (2004). Relacions conceptuals i terminologia: anàlisi i proposta de detecció semiautomática. PhD Thesis.

Gillam, L., Tariq, M., & Ahmad, K. (2005). Terminology and the construction of ontology. *Terminology*, 11(1), pp. 55-81.

Hearst, M. A. (1992). Automatic Acquisition of Hyponyms from Large Text Corpora. In *Proceedings of the Fourteenth International Conference on Computational Linguistics*. Nantes : France, pp. 539-545.

ISO 704:2000 - Terminology work - Principles and methods.

Madsen, B. N., Pedersen, B. S., & Thomsen, H. E. (2001). Defining Semantic Relations for OntoQuery. In P. A. Jensen & P. Skadhauge (Eds.), *Proceedings of the First International OntoQuery Workshop, Ontology-based interpretation of NP's*. Kolding: Department of Business Communication and Information Science, pp. 57-88).

Maingueneau, D. (1998). *Analyser les texts de communication*. Paris: Dunod.

Marshman, E. (2008). Expressions of uncertainty in candidate knowledge-rich contexts: A comparison in English and French specialized texts. *Terminology*, 14(1), pp. 124-151.

Marshman, E., L'Homme, M.-C., & Surtees, V. (2008). Portability of cause–effect relation markers across specialised domains and text genres: a comparative evaluation. *Corpora*, 3(2), pp. 141-172.

Meyer, I. (2001). Extracting knowledge-rich contexts for terminography - A conceptual and methodological framework. In D. Bourigault, C. Jacquemin, & M.-C. L'Homme (Eds.), *Recent Advances in Computational Terminology*. Amsterdam; Philadelphia : John Benjamins Publishing Company, pp. 279-302.

Nuopponen, A. (2005). Concept relations – An update of a concept relation classification. In B. N. Madsen & H. E. Thomsen (Eds.), *Proceedings of TKE 2005 – 7th International Conference on Terminology and Knowledge Engineering*. Copenhagen, pp. 127-138.

Nuopponen, A. (2011). Methods of concept analysis – tools for systematic concept analysis. Part 3 of 3. *LSP Journal*, 2(1), pp. 4-15.

Pazienza, M. T., Pennacchiotti, M., & Zanzotto, F. M. (2005). Terminology Extraction: an analysis of linguistic and statistical approaches. In S. Sirmakessis (Ed.), *Knowledge Mining*, Vol. 185, pp. 255-279.

Pearson, J. (1998). *Terms in Context*. John Benjamins Publishing Company.

Pereira, C., & Soares, A. L. (2008). Ontology Development in Collaborative Networks as a Process of Social Construction of Meaning. In R. Meersman, Z. Tari, & P. Herrero (Eds.), *On the Move to Meaningful Internet Systems: OTM 2008 Workshops*, Vol. 5333. Springer Berlin / Heidelberg, pp. 605-614.

Sager, J. C. (1990). *A practical course in terminology processing*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Defining a Gold Standard for the Evaluation of Term Extractors

Gabriel Bernier-Colborne
Observatoire de linguistique Sens-Texte
Université de Montréal
E-mail: gabriel.bernier-colborne@umontreal.ca

Abstract

We describe a methodology for constructing a gold standard for the automatic evaluation of term extractors, an important step toward establishing a much-needed evaluation protocol for term extraction systems. The gold standard proposed is a fully annotated corpus, constructed in accordance with a specific terminological setting (i.e. the compilation of a specialized dictionary of automotive mechanics), and accounting for the wide variety of realizations of terms in context. A list of all the terminological units in the corpus is extracted, and may be compared to the output of a term extractor, using a set of metrics to assess its performance. Subsets of terminological units may also be extracted, due to the use of XML for annotation purposes, providing a level of customization. Particular attention is paid to the criteria used to select terminological units in the corpus, and the protocol established to account for terminological variation within the corpus.

Keywords: term extraction, evaluation, annotated corpora, gold standard

1. A Gold Standard for Term Extraction

Terminological resources are compiled using various technologies, but few of these technologies have been evaluated from the point of view of their contribution in a specific terminological setting. Since methodologies for compiling these resources are increasingly corpus-based, one of the main tools is the term extractor. Term extractors are used for compiling specialized dictionaries (L'Homme, 2008), ontologies (Biébow & Szulman, 1999) and back-of-the-book indexes (Nazarenko & Aït El Mekki, 2005). This paper describes a proposal for the definition of a gold standard for automatically evaluating term extractors, an important step toward establishing a much-needed evaluation protocol for term extraction systems.

Term extractors are tools designed to retrieve specialized terms from running text, which play a role in a variety of applications, such as terminology, thesaurus building, document indexing, technological watch and ontology development. Like all language technologies, the design and improvement of term extractors requires that developers evaluate these systems.

When new extraction techniques are introduced, attempts are usually made to measure their performance, but how this evaluation is undertaken varies greatly and is often not described in much detail. In some cases, extractors are evaluated by manually scanning their output. In others, extractor output is compared to some sort of term list, but this reference is seldom given much attention in the literature.

This lack of a standardized evaluation protocol has motivated some researchers in the field (L'Homme et al., 1996; Sauron, 2002; Vivaldi & Rodríguez, 2007; Nazarenko & Zargayouna, 2009) to make proposals for such a protocol. Large-scale evaluation efforts have been undertaken in the form of campaigns or workshops, such as ARC A3 and CESART (Timimi & Mustafa el Hadi, 2008) as well as NTCIR-TMREC (Kageura et al., 2000), but these efforts pale in comparison with those made in other branches of NLP (Nazarenko et al., 2009).

If a standardized automatic evaluation protocol is to be established, a gold standard must be defined. To this end, we propose a fully annotated corpus, accounting for the wide variety of realizations of terms in context. This standard is the reflection of a specific terminological setting, namely compiling a specialized dictionary; other applications would necessarily derive a different set of terms.

This gold standard can also be customized, due to the use of XML for annotation purposes. Combined with a set of metrics, such a standard will enable an automatic evaluation of the performance of term extractors, which would be helpful in assessing the performance of a particular system given a specific setting, or comparing different techniques. It would also allow developers to fine-tune their systems by measuring how a given component affects the overall output or how a change in design affects performance (Popescu-Belis, 2007: 77).

2. Specific Problems in Term Extraction

Term extraction raises challenges that are not found in other NLP technologies:

- The notion of "term" is linked to a specific application, as users of a term extractor have different needs in accordance with their professional activity (Estopà Bagot, 2001), be it knowledge organization, indexing or specialized lexicography. Thus, the relevance of terms depends on the task at hand.

- The use of terms in context involves various phenomena that modify their normal structure and can make the identification of term boundaries difficult. These include coordination of complex terms (e.g. *room temperature vulcanizing and anaerobic sealants*), embedding of terms or other elements within compound terms (e.g. *inline (placed in the fuel line) filter*) and anaphoric references (e.g. using *tank* when *fuel tank* has been used previously).

- Concepts may be denoted by more than one term, and terms are subject to various kinds of variation, such as regional variations (e.g. *gearbox/transmission*), spelling variations (e.g. *disc brake/disk brake*), syntactic variations (e.g. *piston head/head of the piston*) and acronyms, which adds a level of complexity to term selection. These variants must be encoded in terminological resources to allow language technologies to recognize them.

Each of these factors also has an impact on the evaluation of term extractors. First, since the ideal set of terms provided by an extractor varies according to the task involved, the reference used to evaluate an extractor must be compiled in accordance with a specific application. With this in mind, we chose the compilation of a specialized dictionary as the application guiding the selection of terms; thus, the gold standard is meant to reflect the work of a terminologist.

The next section will focus on the factors that make term selection and term boundary identification difficult, and how they were dealt with during annotation of the corpus.

3. Annotating the Corpus

The corpus that was annotated and is used to establish the gold standard set of terms -- which will also be used as the test corpus for evaluating term extractors -- consists of three manuals on automotive mechanics, containing some 224,159 tokens. A set of guidelines for selecting terms was established, which includes some of the term selection criteria described by L'Homme (2004: 64--66).

First, units must convey a meaning that is related to the chosen subject field. Units that are morphologically related to previously selected terms (e.g. *cooling pump* and *coolant pump*) are also valid, as long as they are also semantically related. Units that share a paradigmatic bond (e.g. synonymy, meronymy, etc.) with valid terms are also likely candidates. The criteria set out by L'Homme concerning predicative units were not used, as it was decided that only nouns and noun phrases were eligible, since most of the concepts that should be included in a dictionary of this field are denoted by nouns.

Moreover, only units of maximum length are selected, such that terms embedded within terms are not tagged as such. Regional variations, spelling variations and acronyms are included, and the type of variation is specified in the term bank (see Section 4).

More general, or thematic, guidelines were also followed, based on the idea that the application guiding the selection of terms was the production of a specialized dictionary that focuses mainly on the structure of an automobile. Accordingly, terms denoting parts, types of cars and products that a car needs to work are included, whereas terms denoting damages or units of measurement are excluded.

All terms considered relevant according to these guidelines are tagged within the corpus in XML format. These tags serve not only to segment the text into terms and non-terms, but to identify them using a unique identification number and describe certain features of the terms (simple or compound, types of variation), as shown if Figure 1. The selection and tagging process was entirely manual -- term extractors were not used to pre-process the text.

```
One <term id="1" type="s">transaxle</term>
design is the <term id="2" type="c">continuously
variable transmission</term> (<term id="3"
type="a">CVT</term>).
```

Figure 1: Tagged text (with simplified tags).

Rules were established for cases where term segmentation is not so straightforward, as compound terms may be truncated for various reasons.

Coordinated compound terms (e.g. *intake and exhaust valves*) are tagged separately. In compound terms that are disjoined by punctuation marks, embedded terms or paraphrases (e.g. all wheel drive (AWD) systems), any linear sequence that corresponds to a term or part of term is tagged as such, and extraneous elements are excluded whenever possible, as shown in Figure 2. Anaphora can also result in compound term truncation, and these forms are tagged as well.

In all cases, shortened forms are linked to the base term, as described below, and their tags contain an attribute indicating that they are variations of some base term.

```
Many manufacturers have introduced <term
str=”coord”>full time FWD drive</term> or
<term str="coord">all wheel drive</term>
(<term str="disj">AWD) systems</term>
```

Figure 2: Disjoined compound terms (simplified tags).

4. Building the Term Base

The tagged terms are then entered, in their lemmatized form, in a separate term base, in which each term has a record, as shown in Figure 3.

Each sense of a polysemous term receives its own record and identification number. This number not only establishes a distinction between senses, but also allows for easy retrieval of term occurrences from the corpus. Records include a definition, generally adapted from a term base or specialized dictionary, which allows the annotator, as well as any future users of the term base, to obtain the term's meaning, and distinguish between polysemous terms.

Also included in the record is information concerning synonymy and term variation. Synonyms and variations are linked together, all forms pointing to one base term, which is chosen by looking for the headword most often used in dictionaries and term bases, and for the term that has the highest frequency in the corpus. Compound terms that are truncated for the reasons described above (coordination, embedding, anaphora) are given their own record, including a link to the base term. If the base term did not occur in the corpus, the term is "reconstructed" and given a record in the term base.

ID	307
Lemma	EGR valve
Variation type	Acronym
Base term	exhaust gas recirculation valve
Definition	A valve that regulates the flow of exhaust gas into the intake manifold.

Figure 3: Part of the record for the term *EGR valve*.

Inspection of the term base reveals some interesting properties. The term base contains 5489 records, more than half of which, interestingly, are not base terms: 1257 are synonyms of a base term, 1447 are truncated forms of compound terms, and 55 are acronyms. 174 terms were reconstructed from truncated compound terms. Furthermore, of the 23 terms that have a frequency greater than 100, none is compound, if we exclude the base terms that two variations derive from. The corpus contains 28,658 term occurrences, yielding an average term frequency of 5.22, and contains 2,656 hapax legomena -- regarding these figures, it is important to remember that the different meanings of polysemous terms are considered separately. Although this is outside the scope of this paper, a clearer picture of the distribution of terms in the corpus might be provided if frequencies were calculated not only on individual terms (or senses), but also on sets comprising a term and its variations.

The term base can be used as is and compared to the output of a term extractor using a set of metrics. These could be traditional metrics, such as precision and recall, or other metrics that have been proposed for term extraction evaluation (Nazarenko et al., 2009). It is also possible to extract subsets of the term list, for example by excluding uniterms or specific types of terminological variations. This can be easily accomplished using XSLT, and produces a customized term list for the purposes of evaluation.

5. Conclusion

In this paper, we have described a methodology for constructing a gold standard for the automatic evaluation of term extractors. Particular attention has been paid to term selection criteria and term segmentation, as well as the processing of terminological variations.

The gold standard was built by annotating a corpus on automotive mechanics in accordance with a specific application, namely compiling a specialized dictionary. Extensions of this work might include annotating terms in a corpus in accordance with more than one application (ontology development, document indexing, etc.), which would allow evaluators to measure the relevance of extractor output to different applications.

Using the gold standard to evaluate a term extractor is fairly straightforward. The tags are removed from the corpus, which then serves as the test corpus: it is fed to a term extractor, and the output is compared to the standard using an appropriate set of metrics. This enables the performance of a term extractor to be assessed automatically, an important step toward establishing a standardized automatic evaluation protocol for term extractors.

6. Acknowledgements

The work presented in this paper is supported by the Quebec funding agency *Fonds de recherche Société et culture* and the Social Sciences and Humanities Research Council of Canada.

The author also wishes to thank Patrick Drouin and Marie-Claude L'Homme for their helpful suggestions regarding this paper.

7. References

Biébow, B., Szulman, S. (1999). Terminae: A Linguistics-Based Tool for the Building of a Domain Ontology. In *Proceedings of the 11th European Workshop, Knowledge Acquisition, Modelling and Management (EKAW 99), LNAI 1621*. Berlin: Springer Verlag, pp. 49--66.

Estopà Bagot, R. (2001). Les unités de signification spécialisées : élargissant l'objet du travail en terminologie. *Terminology*, 7(2), pp. 217--237.

Kageura, K. et al. (2000). Recent advances in automatic term recognition: Experiences from the NTCIR workshop on information retrieval and term recognition. *Terminology,* 6(2), pp. 151--173.

L'Homme, M.-C., Benali, L., Bertrand, C., Lauduique, P. (1996). Definition of an Evaluation Grid for Term-Extraction Software. *Terminology*, 3(2), pp. 291--312.

L'Homme, M.-C. (2004). *La terminologie: principes et techniques*. Montréal: Presses de l'Université de Montréal.

L'Homme, M.-C. (2008). Le DiCoInfo. Méthodologie pour une nouvelle génération de dictionnaires spécialisés. *Traduire*, 217, pp. 78--103.

Nazarenko, A., Aït El Mekki, T. (2005). Building back of the book indexes. *Terminology*, 11(1), pp. 199--224.

Nazarenko, A., Zargayouna, H. (2009). Evaluating Term Extraction. In *Proceedings of RANLP 2009*, pp. 299--304.

Nazarenko, A., Zargayouna, H., Hamon, O., Van Puymbrouck, J. (2009). Évaluation des outils terminologiques : enjeux, difficultés et propositions. *TAL*, 50(1), pp. 257--281.

Popescu-Belis, A. (2007). Le rôle des métriques d'évaluation dans le processus de recherche en TAL. *TAL*, 48(1), pp. 67--91.

Sauron, V. (2002). Tearing out the Terms: Evaluating Terms Extractors. In *Proceedings of the Aslib conference Translating and the Computer 24*.

Timimi, I. Mustafa el Hadi, W. (2008). CESART : une campagne d'évaluation de systèmes d'acquisition de ressources terminologiques. In S. Chaudiron & K. Choukri (Eds.), *L'évaluation des technologies en traitement de la langue : les campagnes Technolangue*. Paris: Hermes science, pp. 71--91.

Vivaldi, J., Rodríguez, H. (2007). Evaluation of Terms and Term Extraction Systems: A Practical Approach. *Terminology*, 13(2), pp. 225--248.

Mapping from Lexical Resources to High-Level Data Modelling Languages

Gian Piero Zarri
Sorbonne University, LaLIC/STIH Laboratory
Maison de la Recherche, 28 rue Serpente, 75006 Paris, France.
E-mail: zarri@noos.fr, gian_piero.zarri@paris-sorbonne.fr

Abstract

This paper deals with some theoretical and practical problems involved in the transfer from an ‘external’, lexical level to a ‘deep’, conceptual one. The main thesis defended in the paper is linked with the remark that the ‘lexical information’ (in the most general meaning of these words) used to feed the ontological and knowledge-based systems after the passage through some sort of knowledge representation system is *not homogeneous* from a syntactic and semantic point of view. The recourse to a unique conceptual representation model (the well-known ‘uniqueness syndrome’) is then *methodologically erroneous*. In NKRL (Narrative Knowledge Representation Language), for example, several representation models are used. The usual “binary” model is utilized for the ‘standard’ NKRL ontology, HClass (ontology of classes). An “*n*-ary” model, based on the notions of “conceptual predicate” and “functional roles” is used for representing the nodes of HTemp (ontology of templates, i.e., the NKRL “ontology of events”). Recursive lists of (reified) symbolic labels are used for modelling the “connectivity phenomena” and for representing correctly full narratives, complex events, multifaceted eChronicles etc.; special representations are employed for representing the temporal phenomena, and so on.

Keywords: knowledge representation, ontologies, inferences

1. Introduction

Lexical information (in the most general meaning of these words) used to feed the current ontological and knowledge-based systems after the passage through some sort of knowledge representation system is *not homogeneous* from a syntactic and semantic point of view – it can, e.g., simply denote ordinary, unrelated ‘static’ objects and entities or describe the interconnections of structured ‘dynamic’ events. Accordingly, also the practical modalities of use of this ‘transformed’ lexical information within a target system can be totally different. We argue that – in contrast with the normal practice in the knowledge-based and ontological domain, where a same representation model is *repetitively used* to designate entities and phenomena conceptually very different (the well-known ‘*uniqueness syndrome*’) – these differences must be *correctly represented* in the modelling representation language(s) that define (s) the link between the original lexical information and the final conceptual knowledge, and eventually managed according to appropriate, different modalities. This calls, in short, for the use of *diverging data models* – binary, *n*-ary, labelled recursive lists, for the different categories of the original lexical resources.

In the following, we will show how the problem of the intrinsic dissimilarity of the original lexical sources is dealt with in the context of NKRL (Narrative Knowledge Representation Language). More precisely, Section 2 will concern the modelling tools used in NKRL to deal with the so-called “plain/static” knowledge, and Section 3 will describe the (totally different) tools used for the “structured/dynamic” knowledge. Section 4 will deal with the NKRL representation of the “connectivity phenomena”, i.e., those mutual relationships between structured/dynamic units (“elementary events”) within larger conceptual entities (“complex events”, “full narratives”, “multifaceted eChronicles” etc.) that are signalled by surface lexical/syntactic structures like goal, cause, coordination, subordination ,indirect speech etc. Section 5 will supply some basic information on the query/inference system of NKRL, and Section 6 will consist in a short “Conclusion”. NKRL is both a modelling and a development tool – built-up mainly thanks to the contribution of the European Commission through several EC-financed projects – for the representation and management, in a normalised way, of structured multimedia ‘surface’ information, see Zarri (2009; 2011a; 2011b).

2. “Plain/static” lexical knowledge

“Plain” surface lexical information denotes some *stable, self-contained and basic notions* that can be considered, *at least in the short term*, as ‘*a-temporal*’ (static) and ‘*universal*’. This means that the formal descriptions/definitions of these notions (obtained making use of a particular knowledge representation language and used to set up the target knowledge bases) *are not subject to change, at least within the framework of a given application* – even if they can evolve *in the long term* as a consequence, e.g., of the progress of our knowledge. These static notions (that match to separate “*concepts*” at deep level) can be very *general* (corresponding then to surface lexical terms as “human being”, “company”, “colour” or “artefact”) – and proper, then, to several application domains – or linked to a specific application domain (like “control room operator”, “level of temperature”, “valve” or “heat exchanger”). Their self-contained and stable character – where, e.g., the temporal phenomena can be ignored – justifies a conceptual representation/definition *according to some simple model including only, basically, a description of* *the corresponding main (static) properties.* This model can then correspond to the *usual binary model*, where properties are simply expressed as a *binary* (i.e., accepting only two arguments) *relationship* linking two individuals or an individual and a value. And this independently from the fact that these binary relationships are organised into, e.g., frame format as in the original Protégé software (Noy *et al.*, 2000) or take the form of a set of "property" statements used to define a "class" (a "concept") in a W3C language like OWL or OWL 2 (Bechhofer *et al.*, 2004; Hitzler *et al.,* 2009).

Accordingly, the NKRL *conceptual representation* of the "plain/static" surface lexical information is obtained making use of the usual binary model to produce a (quite standard) "*ontology of concepts*" called *HClass* (hierarchy of classes). Note that this "binary" choice is not fundamentally different from the modelling choices adopted in the context of the EC Monnet project (http://www.monnet-project.eu/Monnet/Monnet/English?init=true) to set up the LEMON proposal (http://www.monnet-project.eu/Monnet/resource/Monnet-Website/0000%20-%20Library/0700%20-%20Downloads/lemon-cookbook.pdf) as a standard, RDF/S-based (Brickley and Guha, 2004) tool for sharing lexical information on the Semantic Web. A tiny fragment of HClass (chiefly, of its "non sortal branch") – which shows only, for clarity's sake, the subsumption relationships – is shown in Fig. 1.

We will limit us to note here that the main architectural principle underpinning the HClass' "upper level" concerns the partition between sortal_concept and non_sortal_concept. This corresponds to the differentiation between "*(sortal) notions that can be instantiated directly into enumerable specimens (individuals)*", like chair_ (a physical object) and "*(non-sortal) notions that cannot be instantiated directly into specimens*", like gold_ (a "substance"), white_ (a "colour") or student_ (a "property", more exactly, a "*semantic role*"). Therefore, the specializations of sortal_concept, like chair_, city_ or european_city, can have *direct instances* (CHAIR_27, PARIS_: in NKRL, the instances of concepts, i.e., the "individuals", are denoted conventionally in upper case Roman characters), whereas the non_sortal_concept like gold_, white_ or student_ *can admit further specializations*, see red_gold, whitish_ or university_student, but *do not have direct instances*. A discussion about some (*partial*) correspondences between "sortal concepts" and "count nouns" and "non-sortal concepts" and "uncount (or mass) nouns" can be found in (Zarri, 2009: 125-126). Note, however, that "count/uncount nouns" are (well-known) *linguistic/lexical surface notions*: "sortal/non-sortal concepts" are *deep-level conceptual entities* that can also correspond, in case, to *surface verbs* like "buy" or *adjectives* like "big". Note also that sortal_concept, see Fig. 1, are classified into entity_ and situation_: these last concepts share some similarities with well-known DOLCE's notions like Endurant and Perdurant (Gangemi *et al.*, 2002). With respect then to the NKRL's analysis of controversial notions like "substances" and "colours", see (Zarri, 2009: 132-135); see in general (Zarri, 2009: 43-55, 103-137) for a complete description of the HClass architecture, the arrangement of the nodes according to a frame format, the axioms, some comparisons with alternative models, etc.

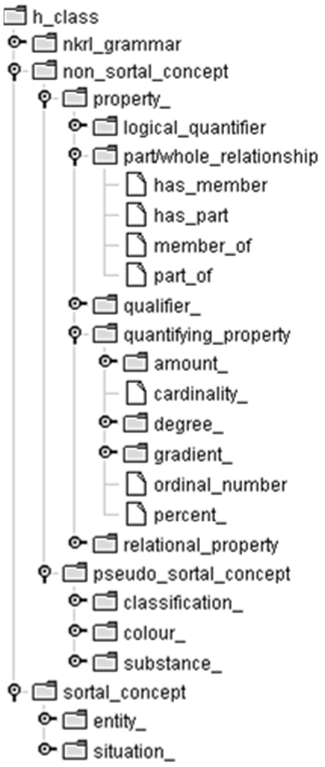


Figure 1. Fragment of HClass, the NKRL standard "ontology of concepts".

3. "Structured/dynamic" lexical knowledge

Structured/dynamic lexical information denotes, on the contrary, the "*elementary events*" – to be collected then within "narratives", "complex events", "eChronicles", "digital storytelling scenarios", "history management systems" etc., i.e., in practice, within *structured streams* of elementary events – *that describe the active or passive, spatio-temporal constrained 'behaviour' of specific subsets of the plain/static lexical entities introduced above*. In a "structured/dynamic" context, the plain/static lexical entities correspond then to the "characters", "actors", "personages" etc. that are involved in the different elementary events: they try to attain a specific result, experience particular situations, manipulate some (concrete or abstract) materials, send or receive messages, etc. In short, *they have a specific "role" in a particular elementary event and/or in a global narrative/complex event etc.* Examples of this sort of *dynamic information* are *elementary events denoted by structured and coherent surface lexical units* like "The Control Room operator presses a button in the context of a start-up sequence", "The oil extractor moves from the state 'idle' to the state 'running'", "Lucy was looking for a taxi" or "Peter lives presently in Paris", etc.

The necessity of making use i) of "*conceptual predicates*" (corresponding to surface verbs like "press", "move", "look for", "live" in the examples before) for specifying the basic type of state, action, situation etc. described in each event, and ii) of the notion of "*functional role*" (Zarri 2011c) to denote the logical and semantic function of the static entities involved in the different events – in "The Control Room operator presses a button...", the "individual" (instance of a concept) CONTROL_ROOM_OPERATOR_1 is the SUBJ(ect) of the action of "pressing" and the individual BUTTON_1 the OBJ(ect) – *makes it difficult to use the common binary approach to represent correctly and effectively this sort of dynamic information*. As it is now widely recognized see, among many other inquiries, (Mizoguchi *et al.*, 2007; Salguero *et al.*, 2009; Liu *et al.*, 2010), the standard (binary) ontologies and the W3C solutions (RDF/S, OWL, OWL 2 etc.) may be in fact sufficient to represent correctly the "plain/static" knowledge but, because of their lack of "expressiveness", *are conceptually inadequate – or at least, very inefficient from a practical point of view – to represent the structural/dynamic knowledge (and the temporal information)*. For this type of information it is then necessary to have recourse to the well-known "*n*-ary" schema denoted by Eq. 1 below:

$$(L_i\ (P_j\ (R_1\ a_1)\ (R_2\ a_2)\ \dots\ (R_n\ a_n)))\ ,\qquad (1)$$

where L_i is the symbolic label identifying the particular *n*-ary structure (the particular *elementary event*, e.g., that corresponding to the surface lexical unit "The Control Room operator presses a button …"), P_j is the conceptual predicate, R_k is the generic functional role and a_k the corresponding argument (e.g., the individuals CONTROL_ROOM-OPERATOR_1, BUTTON_1 etc.). Note that each of the (R_i a_i) cells of Eq. 1, *taken individually*, represents a sort of *binary relationship*. The main point here is, however, that the whole conceptual structure represented by Eq. 1 can be fragmented for practical purposes (e.g., storing within a database), *but must be considered globally whenever significant querying/inference operations must be envisaged about its global 'meaning'*.

According to the NKRL's jargon, the *n*-ary structures represented in Eq. 1 format are called "*templates*" and the corresponding hierarchy – denoting, then, an ontology of "*elementary events*" – is called *HTemp* (hierarchy of templates). Templates can be conceived as the formal representation of generic classes of elementary events like "move a physical object", "be present in a place", "produce a service", "send/receive a message", etc. When a specific, 'surface' elementary event pertaining to one of the general classes included in HTemp must be encoded, the corresponding template is *instantiated*, giving then rise to what, in NKRL's terms, is called a "predicative occurrence". To represent a complete "gas/oil" elementary event expressed in natural language as: "On October 16th, 2008, the Control Room operator pushes SEQ1_BUTTON in the context of a particular sequence of operations (SEQ1) associated with the start-up of the turbine", we must select firstly, in the HTemp hierarchy, the template corresponding to "perform a task or an activity", represented in the upper part of Table 1. This template pertains to the Produce: branch (see its "father" code) of HTemp; note that the elements of a template (as SOURCE etc. in Table 1) included in square brackets are 'optional', i.e., they can be present or not in the instances (predicative occurrences) of the template.

Table 1. Deriving a predicative occurrence from a template.

```
name:   Produce:PerformTask/Activity
father:   Produce:
position:   6.3
natural language description: "Execution of Intellectual or Industrial
Procedures, of Economic Interest Activities, etc."

PRODUCE          SUBJ              var1: [var2]
                 OBJ               var3
                 [SOURCE           var4: [var5]]
                 [BENF             var6: [var7]]
                 [MODAL            var8]
                 [TOPIC            var9]
                 [CONTEXT          var10]
                 { [modulators], ≠abs }

var1         =   human_being_or_social_body
var3         =   activity_, process_, temporal_development
var4         =   human_being_or_social_body   (1)
var6         =   human_being_or_social_body
var8         =   activity_, artefact_, process_, temporal_sequence
var9         =   pseudo_sortal_concept, sortal_concept
var10        =   situation_, symbolic_label
var2, var5, var7  =  location_

virt2.c32)   PRODUCE   SUBJ       INDIVIDUAL_PERSON_102:
                                    (GP1Z_MAIN_CONTROL_ROOM)
                       OBJ        button_pushing
                       TOPIC      SEQ1_BUTTON
                       CONTEXT (SPECIF SEQ1_GREASING_PUMP
                                    (SPECIF member_of
                                    F17_STARTUP_SEQUENCE))
                       date-1:    2008-10-16-08:26
                       date-2:
```

Fig. 2 reproduces a fragment of the HTemp hierarchy that displays, in particular, the conceptual labels of some offsprings of the Produce: and Move: sub-hierarchies – once again, only the subsumption relationships are shown in this figure, see the upper part of Table 1 for an example of actual HTemp node. As it appears from Fig. 2, HTemp consists of *seven branches*, where each one of them includes *only* the templates structured – following the general syntax of Eq. 1 – around one of the seven conceptual predicates (P_j) admitted by the NKRL language, see (Zarri, 2009: 56-59) for a discussion about this architectural choice. HTemp includes presently (April 2012) more than 150 templates, very easy to specialize and customize. A detailed discussion of many of them is given in (Zarri, 2009: 149-177).

From Table 1 we can see that, in a template t_i, the arguments of the predicate (the a_i terms of Eq.1) are represented in practice *by variables with associated* *constraints*. The constraints are expressed as *HClass concepts or combinations of concepts*: the two NKRL ontologies, HClass and HTemp, *interact then strictly*. When creating a predicative occurrence like virt2.c32, see the lower part of Table 1, to represent a particular elementary event, the role fillers of this occurrence *must conform to the constraints of its father-template*. In the predicative occurrence virt2.c32, e.g., INDIVIDUAL_PERSON_102 is an *"individual"*, *instance* of the HClass concept individual_person; this last is a specialization of human_being, specialization in turn of human_being_or_social_body, see the constraint on the argument *var1* associated with the SUBJ(ect) role in the template of Table 1. The "location" GP1Z_MAIN_CONTROL_ROOM is also an individual, instance of the concept control_room, a specialization of the HClass concept location_ – see the constraint on the variable *var2* in Table 1 – through a specialization chain of concepts that includes, among other things, office/room_, building/area_component and extended_location. Note that, in the templates and predicative occurrences, the "determiners/attributes" of the location type are denoted by a vector and are associated with the corresponding arguments of the predicate through the "colon" operator, "**:**", see the SUBJ filler of the occurrence virt2.c32 in Table 1. button_pushing is a specialization of activity_ through device_use and other HClass terms; etc.

Two special operators, date-1 and date-2 – that can be assimilated to specific functional roles – are used to introduce the *temporal information* associated with an elementary event: a detailed description of the formal system used in NKRL for the representation and management of temporal information and its use for indexing purposes can be found, e.g., in (Zarri, 1998). The meaning of "(SPECIF SEQ1_GREASING_PUMP (SPECIF member_of F17_STARTUP_SEQUENCE))" in virt2.c32 is: the general framework (functional role CONTEXT) of the action of pushing the button is a particular process_phase (i.e., SEQ1_GREASING_PUMP) that is a constituent (member_of, see also Fig. 1) of the specific industrial_temporal_sequence represented by the individual F17_STARTUP_SEQUENCE. The "attributive operator", SPECIF(ication), is one of the *four operators that make up the AECS sub-language*, used for the set up of *structured arguments (expansions)*. Apart from SPECIF(ication) = S, AECS includes also the disjunctive operator ALTERN(ative) = A, the distributive operator ENUM(eration) = E and the collective operator COORD(ination) = C. The interweaving of the four operators within an expansion is controlled by the so-called "*priority rule*" (Zarri, 2009: 68-70).

In the context of the conceptual representation, according to the NKRL model, of the "structured/dynamic" lexical knowledge, it can be of some interest to add some comments about those "*functional roles*" R_k that are of paramount importance for asserting the *n*-ary character of Eq. 1 above. Lists of (functional-like) roles created in a Computational Linguistics context are described in, e.g., Bruce (1975), Spärck Jones and Boguraev (1987), Sowa (2000), etc. In a report on "Lexical Semantic Encoding", see http://www.ilc.cnr.it/EAGLES96/EAGLESLE.PDF, the EAGLES researchers supply "…a list of the most popular roles and the properties usually associated with them" that is widely reproduced in the literature as a sort of 'consensus list' about "thematic roles". This list includes 7 items: Agent, Patient, Experiencer, Theme, Location, Source and Goal. A Beneficiary role is added in Palmer *et al.* (2010: 4).

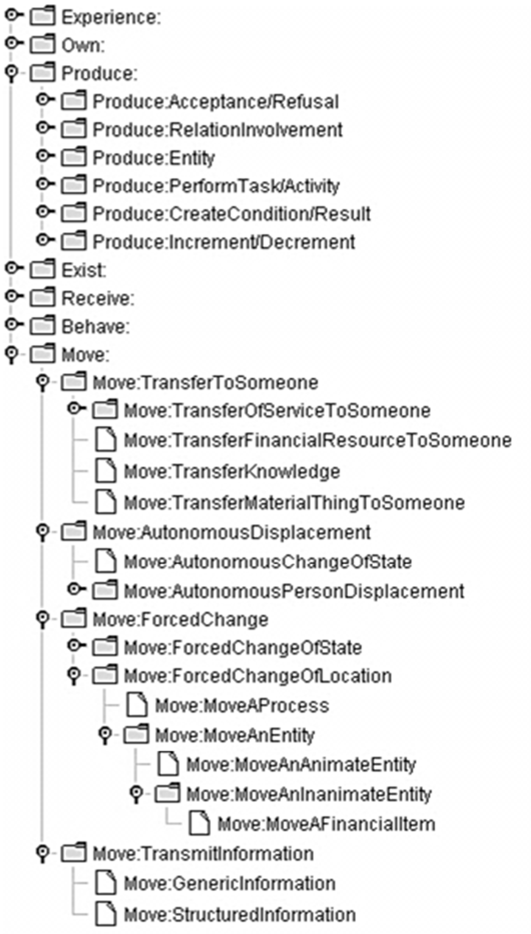

Experience:
Own:
Produce:
Produce:Acceptance/Refusal
Produce:RelationInvolvement
Produce:Entity
Produce:PerformTask/Activity
Produce:CreateCondition/Result
Produce:Increment/Decrement
Exist:
Receive:
Behave:
Move:
Move:TransferToSomeone
Move:TransferOfServiceToSomeone
Move:TransferFinancialResourceToSomeone
Move:TransferKnowledge
Move:TransferMaterialThingToSomeone
Move:AutonomousDisplacement
Move:AutonomousChangeOfState
Move:AutonomousPersonDisplacement
Move:ForcedChange
Move:ForcedChangeOfState
Move:ForcedChangeOfLocation
Move:MoveAProcess
Move:MoveAnEntity
Move:MoveAnAnimateEntity
Move:MoveAnInanimateEntity
Move:MoveAFinancialItem
Move:TransmitInformation
Move:GenericInformation
Move:StructuredInformation

Figure 2. PRODUCE, MOVE etc. branches of the HTemp hierarchy.

When comparing the seven NKRL functional roles with the above solutions a fundamental principle to be kept in mind is that NKRL functional roles *are strictly relative to an elementary event framework*. This means that their duty consists solely in denoting, in the best possible way, the *functional relationships* of the a_i arguments with respect to the predicate P_j within the context of Eq. 1. This principle allows us to discard all the 'roles' that, in the above solutions, can be associated with notions in the CAUSE (e.g., Force and Reason in Spark Jones/Boguraev) or GOAL (e.g., Goal in Spark Jones/Boguraev and EAGLES/Palmer and, at least partially, Completion, Destination and Result in Sowa) style. These last 'roles' do not concern, in fact, the *internal structure* of an elementary event but, on the contrary, the *mutual relations between two (or more) of* *these events*. Let us consider, e.g., examples like "The girl died from an accident" and "John went to town in order to buy a shirt", introduced by Spark Jones and Boguraev as illustrations of the use of their Force and Goal 'roles'. For each of them, we have to deal in reality with some *logical/temporal relationships of the CAUSE/GOAL type between two (or more) elementary events*, identified by recognizing the presence of surface predicative forms like "die" and "accident" or "go" and "buy". The above 'roles' refer then, in reality, to those "*connectivity phenomena*", already mentioned in Section 1, *which allows us to associate together several elementary events* and that are dealt with in the next Section. The seven NKRL functional roles are described informally in Table 2.

Table 2. NKRL's functional roles.

Role	Acronym	Mnemonic Description
Subject	SUBJ	The *main actor* (the "agent") of the elementary event, independently from the grammatical/syntactic form of the corresponding expression in natural language, see "Caesar has been stabbed by Brutus (the SUBJ)". The "filler" (argument of the predicate) of this role is often, but not necessarily, an animate entity or a group of animate entities (e.g., a social body).
Object	OBJ	The entity, *animate* (e.g., Caesar, the "patient", OBJ in the previous example) *or inanimate* (e.g., the book that is moved from John to Mary), which is *acted upon* in the context of the elementary event.
Source	SOURCE	The animate entity (group of entities), *if any*, who is *responsible for* the behaviour, situation, state etc. of the SUBJ of the elementary event.
Beneficiary	BENF	The animate entity, ("Mary" in the "book" example), or the group of entities, which constitutes the "*addressee*" (the "recipient" etc.) of the OBJ mentioned in the event (or, more generally, the addressee of the global behaviour of the SUBJ of the event).
Modality	MODAL	The (often inanimate) *entity* (e.g., the knife) or the *process* (e.g., "stabbing", if the elementary event to be represented was "Brutus killed Caesar by stabbing him") that is *instrumental* in producing the situation described in the elementary event.
Topic	TOPIC	The *theme* ("à propos of…") of the fact(s) or situation(s) represented in the elementary event (e.g., "Mary's birthday", in the context of the "book" example and in the absence of further, more complete details).
Context	CONTEXT	The *general context* ("in the context of…") – often represented by other events/streams of events – of the fact(s) or situation(s) represented in the elementary event under examination, e.g., "Roman Senate's fears about Caesar's ambitions", "John's love for Mary", etc.

5. Connectivity phenomena

To represent completely the *structured/dynamic lexical knowledge*, it is also necessary to have a way of representing the *coherence links* that bring together into a *unique, global entity* (complex event, narrative, multifaceted eChronicles etc.) the different, *constitutive elementary events*. These links are normally expressed in natural language through *lexical/syntactic constructions like causality, goal, indirect speech, co-ordination and subordination*, etc. The term "*connectivity phenomena*" is used here to denote this sort of clues, i.e., to denote what, in the stream of elementary events representing together a structured, dynamic situation i) leads to a *'global meaning' that goes beyond the simple addition of the 'meanings' conveyed by the single elementary events*; ii) defines *the influence of the context* on the meaning of these events. To represent, at the conceptual level, the connectivity phenomena, NKRL makes use of a specific modelling tool, *i.e., second order structures created through reification of conceptual labels in the L_i (see Eq. 1) style*. These structures are formalized as recursive lists differentiated making use of specific *binding operators* as GOAL, CAUSE, COORD(ination), ALTERN(ative), etc.

To supply then an at least intuitive idea of how a *complete* narrative/complex event is represented in NKRL and returning to the Table 1 example, let us suppose we would now state that: "… the production activities leader pushes the SEQ1_BUTTON in the context of … *in order to* start the auxiliary lubrication pump", where the specific elementary event corresponding to the action of pushing is still represented by the predicative occurrence virt2.c32 in Table 1. To encode correctly the new information, see Table 3, we must introduce first an *additional predicative occurrence* labelled, e.g., as virt2.c33 and meaning that: "[the aim of the previous action is to …] move the auxiliary lubrication pump from an 'idle' to a 'running' state". We will eventually add a *binding occurrence* virt2.c30 – labelled using the GOAL binding operator and involving only two arguments – to link together the conceptual labels virt2.c32 (the planning activity) and virt2.c33 (the intended result). The global meaning of virt2.c30 is then: "the activity described in virt2.c32 is focalised towards (GOAL) the realization of virt2.c33". Note also that, in agreement with the semantics of the GOAL operator, see (Zarri, 2009: 71), virt2.c33, the 'result', is *characterized by the presence of an uncertainty code*, "*****", to indicate that, at the moment of 'pushing', the real instantiation of a situation corresponding to 'pump running' cannot be categorically stated.

6. Remarks on the inference procedures

The human and computational effort of transforming from the 'surface' lexical level to the 'deep' conceptual level can only be justified in the context of some practical application implying the concrete utilisation, under the form of *querying/inference operations*, of the obtained formal code. The querying/inference features represent indeed an essential aspect of the NKRL effort: they are, however, out of scope in the framework of the present paper and we refer then the reader, e.g., to (Zarri, 2009: 183-243) for detailed information on this topic.

Table 3. Binding and predicative occurrences.

```
virt2.c32)  PRODUCE   SUBJ      INDIVIDUAL_PERSON_102:
                                  (GP1Z_MAIN_CONTROL_ROOM)
                      OBJ       button_pushing
                      TOPIC     SEQ1_BUTTON
                      CONTEXT (SPECIF SEQ1_GREASING_PUMP
                                  (SPECIF member_of
                                  F17_STARTUP_SEQUENCE))
                      date-1:   2008-10-16-08:26
                      date-2:

Behave:ActExplicitly (1.12)

*virt2.c33)  MOVE   SUBJ    AUXILIARY_LUBRICATION_PUMP_M202:
                                      (idle_)
                    OBJ     AUXILIARY_LUBRICATION_PUMP_M202:
                                      (running_)
                 CONTEXT (SPECIF SEQ1_GREASING_PUMP
                                      (SPECIF member_of
                                      F17_STARTUP_SEQUENCE))
                 date-1:    2008-10-16-08:26
                 date-2:

Move:ForcedchangeofState (4.12)

virt2.c30)       (GOAL  virt2.c32  virt2.c33)
```

We will then limit us to mention here that “reasoning” in NKRL ranges from the *direct questioning* of an NKRL knowledge base making use of “*search patterns*” (formal queries over the predicative occurrences included in the base), to *high-level inference procedures*.

Examples of these last are, e.g., the transformation rules that try to ‘adapt’, from a semantic point of view, the original query/queries (search patterns) that failed to the real contents of the existing knowledge bases. The principle employed consists in using rules to automatically ‘*transform*’ the original query (i.e., the original search pattern) into one or more different queries (search patterns) *that are not strictly ‘equivalent’ but only ‘semantically close’ to the original one*. Let us suppose a user, in a gas/oil context, asks whether a given oil extractor is running; in the absence (failure) of a direct answer, she/he can decide to activate the transformation mechanism. If an appropriate transformation rule is present in the rule repository, the system will be able to reply by supplying other related events stored in the knowledge base, e.g., information stating that the site leader has heard the working noise of the oil extractor. Expressed in natural language, this result can be paraphrased as: “The system cannot definitely assert that the oil extractor is running, but it can certify that the site leader has heard the working noise of this extractor”.

Another example of high-level inference procedures is represented by the hypothesis rules. These allow us to build up automatically ‘*reasonable*’ *logic/semantic connections among the data stored in an NKRL knowledge base* using a number of pre-defined reasoning schemata, e.g., ‘*causal*’ *schemata*. Let us suppose, for example, we have directly retrieved, in a querying-answering mode, information like: “The control room operators of the Asgard Let Down Station (ALDS) have carried out a piping segment isolation procedure in the context of an industrial accident”, which corresponds then to an elementary event to be ‘*explained*’. We can now suppose we have found in the rule base a hypothesis rule whose “*premise*” (triggering pattern) corresponds to a generalization of this event. Under these conditions, we could then be able (at least in principle) to automatically construct, using this hypothesis rule, a sort of ‘*causal explanation*’ of the triggering event (the isolation procedure) by retrieving in the knowledge base information (“reasoning steps”) in the style of: i) “someone has previously attempted to activate a (less drastic) corrective maintenance procedure”; ii) “this corrective maintenance has failed” and iii) “the accident is considered as a serious one”. We want also to verify that the person at the origin of the “corrective maintenance procedure” is a simple field operator, while the person/s that has/have implemented the “piping segment isolation procedure” is/are high-level control room operator/s. Note that – as usual in a ‘hypothesis’ context – the explication proposed by this rule *corresponds to only one of all the possible hypotheses about the ‘causes’ of the original event*: a particular hypothesis rule must always be conceived as a member of a ‘*family*’ of possible explications.

Transformations and hypotheses can also be used in an integrated way. For example, in case of an impossibility of directly finding in the knowledge base information corresponding to “the accident is considered as a serious one”, this reasoning step can be indirectly validated by retrieving by transformation information in the form: “the leakage has a gas cloud shape…”, “a growth of the risk level has been measured…”, “an alarm situation has been validated *and* the level of this alarm is 30% LEL (Low Explosion Level)”, etc.

7. Conclusion

In this paper, we have examined some problems linked with the passage from a “surface” to a “conceptual” level by trying to demonstrate that, because the original lexical knowledge is *not homogeneous* both from a syntactic and a semantic point of view, the recourse to a unique representation model is *methodologically erroneous*. We have then seen that, accordingly NKRL (the Narrative Knowledge Representation Language) makes use of several representation models. The usual “binary” model is utilized for the standard NKRL ontology, HClass (ontology of classes). An “*n*-ary” model, based on the notions of “conceptual predicate” and “functional roles” is used for representing the nodes of HTemp (ontology of templates, i.e., the NKRL “ontology of events”). Recursive lists of (reified) symbolic labels are used for modelling the “connectivity phenomena” and for representing correctly full narratives, complex events etc.; special representation schemes are employed for representing the temporal phenomena, etc. Several successful applications of NKRL in many different domains seem to confirm the correctness of this approach.

8. References

Bechhofer, S., van Harmelen, F., Hendler, J., Horrocks, I., McGuinness, D.L., Patel-Schneider, P.F., Stein, L.A., eds. (2004). *OWL Web Ontology Language Reference*, W3C Recommendation 10 February 2004. W3C, http://www.w3.org/TR/owl-ref/.

Brickley, D., Guha, R.V., eds. (2004). *RDF Vocabulary Description Language 1.0: RDF Schema*, W3C Recommendation 10 February 2004. W3C, http://www.w3.org/TR/rdf-schema/.

Bruce, B. (1975). Case Systems for Natural Language. *Artificial Intelligence* **6**, pp. 327--360.

Gangemi, A., Guarino, N. Masolo, C., Oltramari, A., Schneider, L. (2002). Sweetening Ontologies with DOLCE. In: *Knowledge Engineering and Knowledge Management, Ontologies and the Semantic Web – Proceedings of EKAW'2002*. LNCS, vol. 2473, pp. 166--182. Berlin: Springer.

Hitzler, P., Krötzsch, M., Parsia, B., Patel-Schneider, P.F., Rudolph, S., eds. (2009). *OWL 2 Web Ontology Language Primer*, W3C Recommendation 27 October 2009. W3C, http://www.w3.org/TR/owl2-primer/.

Liu, W., Liu, Z., Fu, J., Hu, R., Zhong, Z. '2010). Extending OWL for Modeling Event-oriented Ontology. In: *Proceedings of the 2010 International Conference on Complex, Intelligent and Software Intensive Systems*, pp. 581--586. Los Alamitos (CA): IEEE Computer Society Press.

Mizoguchi R., Sunagawa E., Kozaki K., and Kitamura Y. (2007). A Model of Roles within an Ontology Development Tool: Hozo. *Journal of Applied Ontology* **2**, pp. 159--179.

Noy, F.N., Fergerson, R.W., Musen, M.A. (2000). The Knowledge Model of Protégé-2000: Combining Interoperability and Flexibility. In: *Knowledge Acquisition, Modeling, and Management – Proceedings of EKAW'2000*. LNCS, vol. 1937, pp. 17--32. Berlin: Springer.

Palmer, M., Gildea, G., Xue, N. (2010). *Semantic Role Labeling*. San Rafael (CA): Morgan and Claypool Publishers.

Salguero, A.G., Delgado, C., Araque, F. (2009). Easing the Definition of N-Ary Relations for Supporting Spatio-Temporal Models in OWL. In: *Computer Aided Systems Theory, 12th International Conference, EUROCAST 2009*. LNCS, vol. 5717, pp. 271--278. Berlin: Springer.

Sowa, J.F. (2000). *Knowledge Representation – Logical, Philosophical, and Computational Foundations.* Pacific Grove (CA): Brooks/Cole.

Spärck Jones, K., Boguraev, B. (1987). A Note on a Study of Cases. *Computational Linguistics* **13**, pp. 65--68.

Zarri, G.P. (1998). Representation of Temporal Knowledge in Events: The Formalism, and Its Potential for Legal Narratives. *Information & Communications Technology Law – Special Issue on Formal Models of Legal Time: Law, Computers and Artificial Intelligence* **7**, pp. 213--241.

Zarri, G.P. (2009). *Representation and Management of Narrative Information, Theoretical Principles and Implementation*. London: Springer.

Zarri, G.P. (2011a). Knowledge Representation and Inference Techniques to Improve the Management of Gas and Oil Facilities. *Knowledge-Based Systems (KNOSYS)* **24**, pp. 989--1003.

Zarri, G.P. (2011b). A Structured Metadata Approach for Dealing in an 'Intelligent' Way with Complex 'Narrative' Information. *International Journal of Metadata, Semantics and Ontologies (IJMSO)* **6**, pp. 10--22.

Zarri, G.P. (2011c). Differentiating Between "Functional" and "Semantic" Roles in a High-Level Conceptual Data Modeling Language. In: *Proceedings of the 24th International Florida AI Research Society Conference, FLAIRS-24*, pp. 75--80. Menlo Park (CA): AAAI Press.

Supporting the identification of conceptual relations in semi-formal ontology development

Cristóvão Sousa$^{(1,2)}$, António Lucas Soares$^{(2,3)}$, Carla Pereira$^{(1,2)}$, Rute Costa$^{(4)}$

[1]CIICESI - ESTGF, Polytechnic Institute of Porto, Rua do Curral - Casa do Curral-Margaride, 4610-156, Felgueiras-Portugal;
[2]INESC Porto, Rua Dr. Roberto Frias, s/n 4200, Porto-Portugal;
[3]Department of Informatics Engineering, Faculty of Engineering, University of Porto, Rua Dr. Roberto Frias, s/n 4200, Porto-Portugal
[4]CLUNL - Universidade Nova de Lisboa, Avenida de Berna 26-C 1069 - 61 Lisboa

E-mail: cristovao.sousa@inescporto.pt, als@fe.up.pt, cpereira@inescporto.pt, rute.costa@fcsh.unl.pt

Abstract

Conceptualisation processes are pervasive to most technical and professional activities, but are seldom addressed explicitly due to the lack of theoretical and practical methods and tools. However, it seems not to be a popular research topic in knowledge representation or its sub-areas such as ontology engineering. The approach described in this paper is a contribution to the development of computer based tools supporting collaborative conceptualisation processes. The particularly challenging problem of conceptual relations elicitation is here tackled through a combination of ontological and terminological analysis, through a double theoretical perspective. A conceptual relations reference model was synthesised from a foundational ontological analysis and implemented through conceptual relations templates. The later are part of the conceptME system, a platform developed within this research line, providing knowledge and terminological tools and resources to support activities that involve collaborative conceptualisation processes. The work described in this paper adds more support to an area where this support is very scarce.

Keywords: Conceptualisation framework, conceptual relations, knowledge representation, terminology

1. Introduction

Research on knowledge representation has yet to address conceptual representations as designed, pragmatic artefacts whose validity and value is time, context and situation dependent. In fact, in several application areas, conceptual representations need to be created and recreated, used and reused, decomposed and synthesised and eventually disposed of, according to specific needs. The problems arising from this pattern of use are related with the heterogeneity of pragmatic conceptual representations and the social processes of developing them. Researching solutions for these problems has been surprisingly scarce in the knowledge representation literature, more specifically in the ontology engineering area.

This paper describes part of the research carried out in the cogniNET project[1]. The generic goal of this project is to develop the theory and practice of collaborative knowledge representation. More specifically, the project developed methods, models and tools to support domain specialists in collaboratively creating conceptual representations to be used in activities such as building knowledge organisation systems (for information management) or developing terminologies. The pragmatic properties of these conceptual representations (some may call them “lightweight ontologies” (Yu-liang, 2007)) - short-term validity, contextual and situational dependency - required an adequate information technology support materialised in a collaborative platform to support the above mentioned activities. That is what motivate the development of conceptME, a “conceptual Modelling Environment” where groups of specialists can find tools and resources to collaboratively develop conceptual representations (e.g. concept maps), organise them in libraries, share them with other colleagues and reuse them as needed. Tasks involving Conceptualisation 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 (Barros et. al, 2012 - this volume). It was then natural that conceptME integrates seamlessly methods and tools from terminology and knowledge representation. Furthermore, it is theoretically framed by the socio-semantics perspective in knowledge representation (Pereira and Soares, 2009) and the conceptually-based perspective in terminology (Costa, 2006).

From the research so far in cogniNET, it was concluded that the most difficult problem in a conceptualisation process is the elicitation of conceptual relations (Auger and Barriére, 2010: Elsayed, 2009). Thus, particular attention has been given to this subject in the development of conceptME. The problem has to be tackled from the double perspective of terminology and ontology development. In this paper the emphasis is on

[1] PTDC/EIA-EIA/103779/2008 finishing 10/2012: partners: INESC Porto and CLUNL

the ontology-based tools to support conceptual relations elicitation. Other paper in this volume emphasises the terminology tools (Barros et al., 2012). The rest of the paper overviews the conceptualisation process and the conceptME framework to support it, describes the ontological approach to conceptual relations elicitation, including a reference model and and application example. The paper finishes with an overview of the implementation of the reference model in conceptME.

2. Conceptualisation Processes

2.1 Definitions

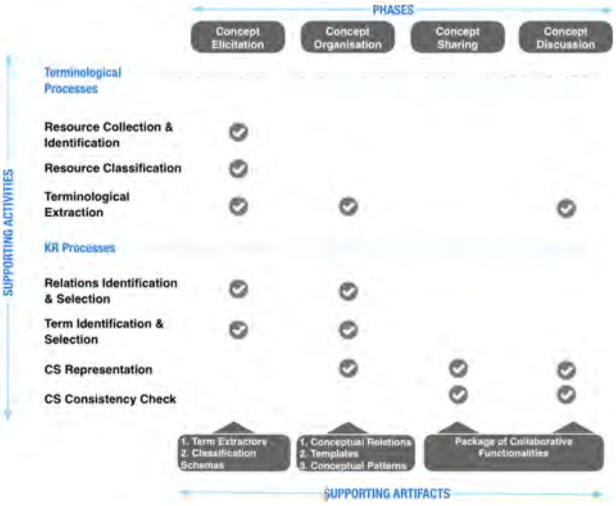
In this paper, the central notion of "conceptualisation process" (CP) is adopted following Pereira et al. (2012). In relation to an individual, a conceptualisation process of a given piece of reality is a collection of ordered cognitive activities that has as inputs information and knowledge internally or externally accessible to the individual, and as the output an internal or external conceptual representation. Furthermore, a "collaborative conceptualisation process" (CCP) is a conceptualisation process that involves more than one individual producing an agreed conceptual representation. In addition to an individual CP, the CCP involves social activities that include the negotiation of meaning and practical management activities for the collaborative process. In this paper "knowledge representation process" is also used to refer, in practical terms, to a CP.

The term conceptual structure (CS) is widely used in knowledge representation and conceptual modelling literature in general. According to Sowa (2000), conceptual structures express declarative knowledge by representing it as a connected bipartite oriented graph (conceptual graph). Mineau states that "every network of concepts, whether an hierarchy, ontology, partonomy or semantic network can be called a structure of concepts. More specifically, CS is a representation of the structure of concepts, which belong to a subject field or domain. Conceptual structures are related with rich aspects of perceptual and subjective experience." (Mineau et. al, 1999). The author considers that CS are models (or artefacts) representing a certain perception of reality. The construction of these modelling structures in a consistent way is a challenge, specially when it intends to follow an informal approach to representing conceptual structures (e.g, concept maps). As stated by Meena & Nagarjuna (2010), an informal representation of conceptual structures is harder than it may look like. It is in this context that the importance of the conceptual relations arises.

This paper considers "conceptual relation" as a relation linking meanings of concepts, and "lexical or semantic relation" as a relation linking linguistic units and the meanings they denotes". Although concentrating on the identification and representation of the former, the latter is important to achieve that goal.

2.2 Towards a conceptualisation framework

The conceptualisation framework[2] 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 Conceptualisation process in what regards to its main phases, activities and artefacts, tying together the terminological and knowledge representation view.



	Concept Elicitation	Concept Organisation	Concept Sharing	Concept Discussion
Terminological Processes				
Resource Collection & Identification	✓			
Resource Classification	✓			
Terminological Extraction	✓	✓		✓
KR Processes				
Relations Identification & Selection	✓	✓		
Term Identification & Selection	✓	✓		
CS Representation		✓	✓	✓
CS Consistency Check			✓	✓
	1. Term Extractors 2. Classification Schemas	1. Conceptual Relations 2. Templates 3. Conceptual Patterns	Package of Collaborative Functionalities	

Figure 1: Conceptualisation framework

The Conceptualisation process is divided into four main phases - concept elicitation, concept organization, and concept sharing and concept negotiation - whose execution depends on a set of supporting activities. Two main processes are considered: a) terminological processes and b) knowledge representation processes. The terminological processes encompass methods for identification/selection of lexical resources and its classification as well as terminological extraction techniques. Moreover, terminological processes are also used to assist the negotiation activities during concept discussion.

KR processes encompass activities to elicit, organise, share and discuss the conceptual representations. Concepts elicitation can be supported by the terminological processes helping to overcome concept identification difficulties (naming, meanings, contexts of use).

In order to accomplish domain structuring, general-purpose relations were identified and typified according to the characteristics of the terms that could be linked together. Those common basic conceptual relations will enable the representation of conceptual structures, made by entities from mental spaces or from reality, by means of the same patterns. For that a set of templates were instantiated and made available through a library of pre-defined reusable "term-relation-term" structures, which could either help to find the suitable

[2] Provides a structured and multidimensional view over the Conceptualisation process in which regards to its main phases, activities and artefacts, tying together the terminological and knowledge representation view.

relations to interlink selected terms or suggest the terms which fit a specific relation. However each domain has its own specificity, which asks for specific relations. Here, once again, terminological extraction plays an important role on the identification/selection of terms, which denote relations according to specific domain and task.

3. An ontological approach to conceptual relations elicitation

3.1 Related work

Although the specification of conceptual relations have been approached according to different theoretical perspectives, of conceptual relations stills as a relevant and interesting research topic, though hard to approach. The interest on relations specification is increasing within the research community when conceptualizing a specific domain, either for creating knowledge base systems or upper-level ontologies.

A closer look into the literature concerning the interplay between terminology and knowledge representation, revealed that ontology engineering together with the development of knowledge bases are the main research topics. Among the reported research, the elicitation of conceptual relations is addressed with variable emphasis and in different forms.

There are several researches addressing the construction of ontologies grounded in terminology (Gillam, L. et al., 2005) (Yu-liang, 2007) (Buitelaar, P. et al., 2009). In (Gillam, L. et al., 2005), terminology plays the role of term system provider which act as input for the construction of the ontology. The authors propose an automatic process to identify a tree of lexical related terms, which constitute a candidate conceptual structure.

(Yu-liang, 2007) motivated by the lack of reference models in the process of building ontologies, presented a three step process, grounded in extraction techniques and textual corpus analysis, comprising: i) recognize terminology in text (using statistical analysis and association rules created using TexAnalyst software, plus semantic network analysis, in order to overcome the problem of ignoring the terms with a low frequency) ; b) name tags in terminology (in order to face the synonyms or variance issue. Repertory Grid Technique was used); c) derive hierarchies (using Formal Conceptual Analysis). His stance is that “linguistic perspectives should be considered while building ontologies”. Further he underlines the need to develop a “lightweight ontology” which “is a schema like taxonomy which comprises a conceptual system used to model knowledge. Consequently, ontology editors must first construct a conceptual system, after which editors should identify hierarchical structures among concepts”. (Buitelaar, P. et al., 2009) argue - once again - that ontologies should be grounded in linguistics. The goal was to enrich current formalisms such as RDFS/OWL to include linguistic information such as “part-of-speech metadata of the lexical items”, morphological information and variations, expressed as RDFS/OWL properties.

One of the main areas where terminology interacts with ontology engineering is that of ontology learning. As mentioned by (Buitelaar, P. et al., 2005) “Term extraction is a prerequisite for all aspects of ontology learning from text”. However we consider that the use of terminology within knowledge representation contexts is wider than the use given by ontology learning field, that is, mainly corpus tagging for information extraction. Learning ontological relations is the most recent target in the scope of ontology learning. This is a fact that the identification of relations between concepts has a significant importance in the creation of artefacts to represent a specific domain. Other authors place conceptual relations as the core issue either on developing ontologies (Alvarez et al, 2007a) (Faber et al, 2009) or in representing a conceptual system in general (Storey, 2005) (Elsayed, 2009) (Auger, A. and Barrière, C., 2010)

3.2 Foundational ontological analysis

In order to accomplish the primary research goal and delineate a strategy to identify a set of domain-neutral conceptual relations, the focus has been placed on the main foundational upper-level ontologies, once ontologies describe the very general concepts that are the same across all knowledge domains. “Ontologies are often equated with taxonomic hierarchies of classes, classes definitions and the subsumption relations” (Grubber, 1993), however the aim is to identify other than only these ones. Hereupon, the approach was to follow through the main upper-level ontologies, namely: CyC[3], BFO[4], GFO[5], UFO (Guizzardi, 2005), SUMO[6], COSMO[7], DOLCE[8], PROTON[9].

When entering in a more detailed analysis of the ontologies it was found a considerable difference between the conceptual structures of ontologies (regarding the size and content). COSMO Ontology, for instance, has over 700 relations and 6400 Classes and its conceptual structure is translated into hierarchical relations. BFO is much smaller but contains only taxonomic relations. Yet, GFO and even UFO, provide a more interesting conceptual structure that goes beyond a taxonomy. Other issue on this study was the fact that some of these ontologies overlap each other. COSMO uses elements from CyC, SUMO, BFO and DOLCE. BFO, for example, overlaps DOLCE and SUMO. By its turn the second version of UFO combines elements from DOLCE.

[3] http://www.opencyc.org/
[4] http://www.ifomis.org/bfo
[5] http://www.onto-med.de/ontologies/gfo.html
[6] http://www.ontologyportal.org/
[7] http://micra.com/COSMO/
[8] http://www.loa.istc.cnr.it/DOLCE.html
[9] http://proton.semanticweb.org/

Summarising, and without going into the details, the following ontological categories of formal relations where selected: constitution and containment dependence, existential dependence, generic dependence, historical dependence (Thomasson, 1999). In this paper it will not be considered the Existential Dependence since it has to do with relations between entities and its examples and the intent, at this level, is to avoid mixing classes (concepts). But, in fact, the individuals which belongs to a specific category should be known in order to a new category/concept be added accurately. Constitution and Containment dependence was detailed as a Part-Whole conceptual relation as it is more common across literature. Following the same purpose, generic dependence was detailed into the Generic-Specific category. Historical dependence is related with temporal location relations. These kind of relations are treated differently (in terms of each taxonomy of categories used) in the available upper-lever ontologies. Historical dependence could have a space or time boundary considering physical or non-physical objects respectively, hence it was decided to detail it into two more specific conceptual relations, namely: Temporal Conceptual Relation and Spatial Conceptual Relation. Inspired mainly by GFO, it was decided to include Participation relation. Participation could be considered as an extension of historical dependence relation, however, in the context of collaborative networks of organizations, participation relation has an important role on offering an orthogonal view of an event or process. It can also offer a brief overview on the social interaction network around an event or process. Finally it was also considered the Cause-Effect Conceptual Relation. Casualty could easily be associated to space and time relations to describe events and consequently considered as not adding value for the current purpose. However, Cause-Effect Conceptual Relation is fundamental to add some dynamicity to conceptual representations on describing phenomenons and agents of change within some process or event or object state. Finally, it was achieved the following taxonomic:

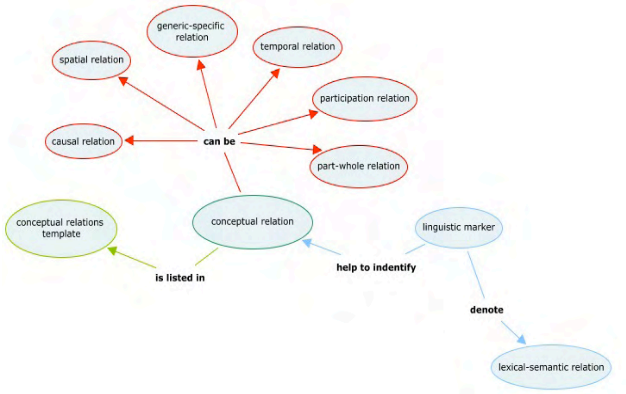
- Constitution and Containment Dependence
    - Part-Whole Conceptual Relation
- Generic Dependence
    - Generic-Specific Conceptual Relation
- Time and Space Dependence
    - Spatial Conceptual Relation
    - Temporal Conceptual Relation
- Cause-Effect Conceptual Relation
- Participation Conceptual Relation

3.3 Conceptual Relations Reference Model

To implement the support to specialists participating in a collaborative Conceptualisation process, in the specific activities regarding conceptual relations elicitation, a Conceptual Relations Reference Model (CRRM) was developed, providing a common baseline for Conceptual Structures construction.

Figure 2 shows the conceptual structure of the conceptual relations reference model and figure 3 shows its actual implementation.

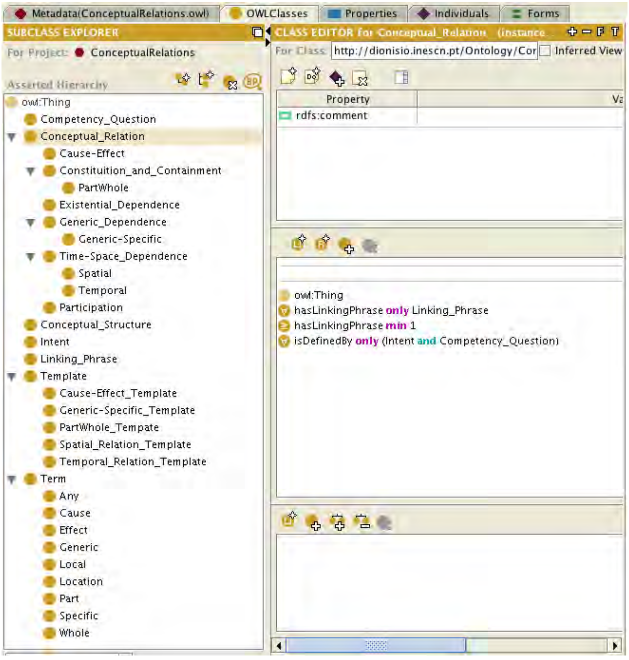
Figure 2: Conceptual structure of the CRRM



CRRM includes a taxonomy of conceptual relations under the foundational ontological relations umbrella. Additionally a set o templates were formalized - one for each conceptual relation. Together with conceptual relations taxonomy and templates, generic types of terms were provided characterizing the terms each conceptual relation could connect. Moreover, a class for competency questions and intents allow the definition of a conceptual relation, which could be represented by a linking phrase that designates the relation. Defined all restrictions and relations between CRRM concepts, the user has available a guidance to instantiate conceptual structures in the basic form of “concept - relation - concept”. CRRM is intended to be a baseline model, which could be extended either by adding new labels to designate conceptual relations or detailing it by means of domain-specific conceptual relations.

The identified domain-neutral relations among terms are defined by its intent and competency questions. The intent is the goal or application scenario of a certain type of relation, whereas competency questions purpose is to define the scope of a conceptual relation. In this case it is possible to define more than one question.

The way CCRM assist users along the Conceptualisation process is through templates, which are used to build conceptual structures. Nevertheless, if general conceptual relations could ease the Conceptualisation process in identifying the nature of the relation among top-level concepts of a domain, it is not so easy to find the appropriate naming for such relation. For that, a set of linking phrases were collected according to upper-level ontologies analysis and are then provided to the user. Yet, the user could not agree on the provided standard options to name a conceptual relation and he/she could use the “text” to get clues about new possible labels for a conceptual relation.


Metadata(ConceptualRelations.owl)
OWLClasses
Properties
Individuals
Forms
SUBCLASS EXPLORER
For Project: ConceptualRelations
Asserted Hierarchy
CLASS EDITOR for Conceptual_Relation (instance
For Class: http://dionisio.inescn.pt/Ontology/Cor
Inferred View
Property
Va

```
owl:Thing
  Competency_Question
  Conceptual_Relation
    Cause-Effect
    Constituition_and_Containment
      PartWhole
    Existential_Dependence
    Generic_Dependence
      Generic-Specific
    Time-Space_Dependence
      Spatial
      Temporal
    Participation
  Conceptual_Structure
  Intent
  Linking_Phrase
  Template
    Cause-Effect_Template
    Generic-Specific_Template
    PartWhole_Tempate
    Spatial_Relation_Template
    Temporal_Relation_Template
  Term
    Any
    Cause
    Effect
    Generic
    Local
    Location
    Part
    Specific
    Whole
```

rdfs:comment

```
owl:Thing
hasLinkingPhrase only Linking_Phrase
hasLinkingPhrase min 1
isDefinedBy only (Intent and Competency_Question)
```

Figure 3: CRRM implementation

The following example consists of a simple use case scenario for the proposed approach. From a corpus of urban rehabilitation, a list of terms was extracted. The next text snippet represents a small context containing three terms of the retrieved list.

"The diagnosis consists essentially in the process for identification or determination of the nature and the cause of the anomalies, through observation and investigation, using several tests, historical research and the expert opinion."

So far the user has been assisted by terminological extracting services, but then the challenge is on building the conceptual structure itself, that is, linking together the collected terms properly. Here, the templates can guide users to complete the process of concept organization. Browsing through the templates available the user is informed about the context of use (intent) of each template. Part-Whole, for instance, has the following intent: *"Used to represent relations between concepts in which a concept has another concept as its constituent forming a whole, which could be dependent or independent from its parts."* Following this, it is acceptable to consider that Part-Whole template is suitable to link diagnosis, observation and investigation according to the previous context. At the next step, the user will check the feasibility of the proposed link between the terms, according to a set of competency questions. Considering terms in analysis, examples of those questions could be:

1. *Observation* is a component/constituent or is attached to *Diagnosis*?
2. *Observation* and *Diagnosis* are nested?
3. *Observation* and *Diagnosis* are physically engaged?

By confirming the questions the user are able to select the appropriated linking phase between the terms. Here, the selected questions could indicate one or other linking phrase. For example, if the user agreed with question 2 and 3, the resulting conceptual structure could be: "Diagnoses includes Observation". On the other hand if the user agreed with question 1 and 3, the resulting conceptual structure could be: "Observation *isPartOf* Diagnosis".

4. A collaborative tool to support specialists to elicit conceptual relations

This section illustrates the use of the CRRM in a tool to support specialists in collaborative Conceptualisation processes. This tool - conceptME conceptual modeling environment - is under development and a short description of it can be found in this book (Barros *et. all*, 2012)

Besides the definition and inclusion of terminological extraction processes to retrieve relevant term candidates as input for structuring a specific domain, the knowledge acquisition "bottleneck" (Wagner, 2008) - characterized by: i) existence of unreliable sources of knowledge; ii) complexity on building transferable knowledge representations and; iii) the slowness of the process - was also addressed. CRRM was the starting point to define a workflow (see fig. 4) for aiding conceptual structures construction.

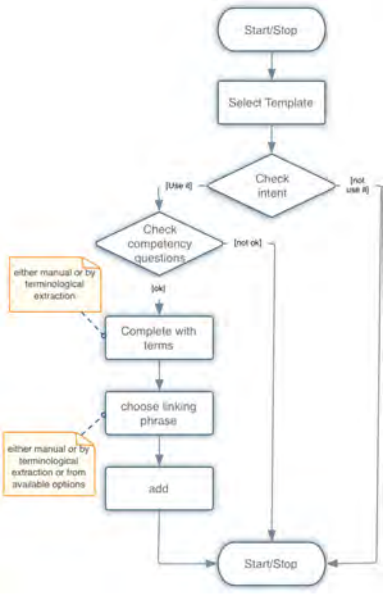
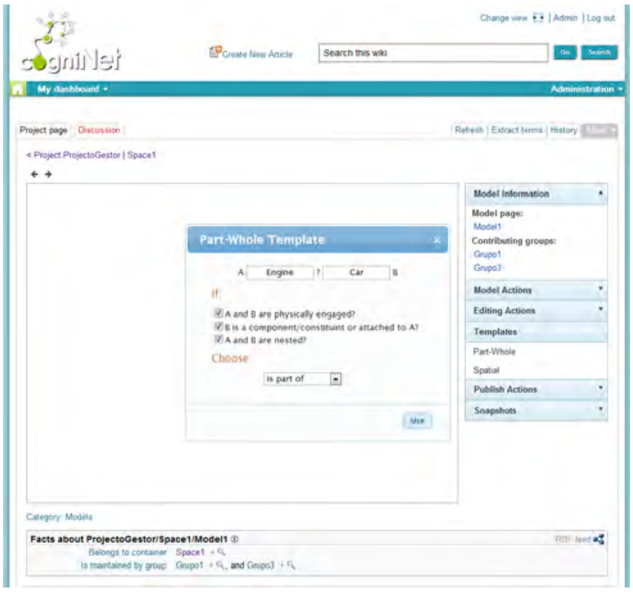

Start/Stop
Select Template
Check intent
[Use it]
[not use it]
Check competency questions
[not ok]
[ok]
either manual or by terminological extraction
Complete with terms
choose linking phrase
either manual or by terminological extraction or from available options
add
Start/Stop

Figure 4: Worlflow to add Conceptual Structures from templates

Conceptually, templates allow us to build statements about a domain of knowledge and represent those statements as conceptual structures. Within conceptME, templates could be used in three different scenarios: a) as a way to unblock the initial process of drawing conceptual structures; b) as a way to assist users to define a conceptual relation among two terms he/she has in mind and; c) in addition to terminological extraction. In other words, templates act as "skeleton" for relating two concepts to help users to create top-level domain statements based on the information contained in the generic or "skeleton" template. But, these conceptual relations templates (CRT) are not static, so, users could use it in conjuction with terminological features in order to discover candidate terms to populate or extend a conceptual relation template (see fig.5).



A Engine ? Car B

If:

☑ A and B are physically engaged?
☑ B is a component/constituint or attached to A?
☑ A and B are nested?

Choose

is part of

Model page: Model1
Contributing groups: Grupo1 Grupo3

Facts about ProjectoGestor/Space1/Model1
Belongs to container Space1
Is maintained by group Grupo1, and Grupo3

Figure 5: conceptME template usage

Basically the process run as follows: i) a library of CRT was made available from the conceptME user interface (UI). A mouse over action on each template and the main intent of a template is prompt. For the Part-Whole template, for example, the following intent sentence could arise: “*Part-Whole conceptual relation could be used to represent “Has-A” hierarchies (also known as partonomies) - it represent relations between concepts in which a concept has another concept as its constituent ou component forming a whole, which could be dependent or independent from its parts.*”. Afterwards, user decides if he/she wants to proceed, if so, a new pop-up screen arises (see fig.5) and the user is supported with more information about the template, more specifically the competency questions and the possible designation for the conceptual relation. At this stage users fill the gaps with the appropriate terms he/she has in mind or uses term extraction features.

5. Conclusion

This paper raised the attention to the problems and requirements in supporting the elicitation of conceptual relations in conceptualisation processes. Within the scope of a research line aimed at creating theory and practical tools to support collaborative conceptualisation processes by domain specialists, a way to overcome those difficulties was overviewed and its implementation outlined.

An on going work is described, but the results so far are sufficiently innovative to encourage the next steps which are, naturally, the empirical validation of the proposal and associated hypothesis. Moreover, the empirical studies are fundamental to clarify the roles and tasks within the process and the level of expertise needed to perform each task.

Future work will be focused on the running of experiments aimed at obtaining further feedback from specialists, in several domains, to improve and fine tune the methods and tools developed so far. Preliminary experiments were already run and the results showed that the CRT support improved the specialist performance in eliciting conceptual relations. But, as expected, as detail level of domain description increases, CRT revealed not so useful. It was also found that some templates were more easily understood than other. Nevertheless, these results are inconclusive so far.

Collaborative knowledge representation is a challenging research area, even more when theoretical and practical connections with terminology are established. In what concerns to conceptual relations elicitation this research is surely proposing a refreshing view on this subject.

.

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".

References

A. Auger and C. Barrière, “Probing semantic relations,” Probing Semantic Relations: Exploration and Identification in Specialized Texts, vol. 23, p. 1, 2010

Barros, S., Costa, C., Soares, AL., Silva, M. “Integrating terminological methods in a framework for collaborative development of semi-formal ontologies”. In Proceedings of the colabTKR - collaboration in Terminology and Knowledge, part of the LREC international conference on Language Resources and Evaluation, 2012

Buitelaar, P., Magnini, B., 2005. Ontology Learning from Text: An Overview. In Paul Buitellar, P., Cimiano, P., Mangnini B. (Eds.), Ontology learning from text: Methods, Applications and Evaluation, p.3-12.

Buitelaar, P. et al., 2009. Towards Linguistically Grounded Ontologies. In Proceedings of the 6th European Semantic Web Conference on The Semantic Web: Research and Applications. ESWC 2009 Heraklion. Berlin, Heidelberg: Springer-Verlag, pp. 111–125.

Costa, R. 2006. Terminology, Corpus Linguistics and Ontology, Constrastive Studies and Valency. Studies in Honor of Hans Ulrich Boas. Petra C. Steiner, Hans C. Boas. Stefan Scheirholz [eds.]. Berlin – Bern: Peter Lang Verlag

Elsayed, A., 2008. A Framework for Using Semantic Relations in Conceptual Structures. In 2008 Eighth IEEE International Conference on Advanced Learning

Technologies. 2008 Eighth IEEE International Conference on Advanced Learning Technologies. Santander, Cantabria, Spain, pp. 1069-1070.

F.J. Álvarez, A. Vaquero, F. Sáenz, M. Buenaga. “Neglecting Semantic Relations: Consequences and Proposals”. In Proceedings of the IADIS International Conference on Intelligent Systems and Agents (ISA 2007), part of the IADIS Multiconference on Computer Science and Information Systems (MCCSIS 2007), ISBN 978-972-8924-39-3, pp. 99-108, IADIS Press, July, 2007a

Faber, P., P. León, and J. A Prieto. 2009.a “Semantic relations, dynamicity and terminological knowledge bases.” Current Issues in Language Studies 1 (1): 1–23.

Gillam, L., Tariq, M. & Ahmad, K., 2005. Terminology and the Construction of Ontology. TERMINOLOGY, 11, p.55–81.

Gruber, Thomas R. (June 1993). "A translation approach to portable ontology specifications". Knowledge Acquisition 5 (2): 199–220.

Guizzardi, G. Ontological Foundations for Structural Conceptual Models, PhD Thesis (CUM LAUDE), University of Twente, The Netherlands. Published as the book “Ontological Foundations for Structural Conceptual Models”, Telematica Instituut Fundamental Research Series No. 15, ISBN 90-75176-81-3 ISSN 1388-1795; No. 015; CTIT PhD-thesis, ISSN 1381-3617; No. 05-74.

J. Sowa, Knowledge representation : logical, philosophical, and computational foundations. Pacific Grove: Brooks/Cole, 2000.

Meena Kharatmal & Nagarjuna G. (2010): 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. Doi: 10.1007/978-3-642-14197-3_22

Mineau, G., Stumme, G., Wille, R.: Conceptual Structures Represented by Conceptual Graphs and Formal Concept Analysis. Proc. ICCS ’99, LNAI 1640. Springer, Heidelberg 1999, 423–441

Pereira, Carla, Cristóvão Sousa, and António Lucas Soares, 2009. “A Socio-semantic Approach to Collaborative Domain Conceptualisation.” In On the Move to Meaningful Internet Systems: OTM 2009 Workshops, Robert Meersman, Pilar Herrero, and Tharam Dillon ed. 5872:524–533. Springer Berlin / Heidelberg. doi:10.1007/978-3-642-05290-3_66.

Pereira, C., Sousa, C., Soares, AL, 2012, A socio-semantic approach to support conceptualisation processes: a case study in an R&D project. accepted for publication in the International Journal of Computer Integrated Manufacturing.

Storey, V.C., 2005. Comparing relationships in conceptual modeling: mapping to semantic classifications. IEEE Transactions on Knowledge and Data Engineering, 17(11), pp.1478-1489.

Thomasson, A. L., “Fiction and Metaphysics”, Cambridge University Press, ISBN-13: 9780521065214,

Wagner, C., 2008. Breaking the knowledge acquisition bottleneck through conversational knowledge management. Innovative Technologies for Information Resources Management, p.200.

Yu-Liang, C., 2007. Elicitation synergy of extracting conceptual tags and hierarchies in textual document. Expert Systems with Applications, 32(2), pp.349–357