# Lexically-based Ontologies and Ontologically Based Lexicons

Antonio Toral, Valeria Quochi, Riccardo Del Gratta, Monica Monachini, Claudia Soria and Nicoletta Calzolari firstname.lastname@ilc.cnr.it

> Istituto di Linguistica Computazionale Consiglio Nazionale delle Ricerche Via G. Moruzzi 1 - 56124 Pisa, Italy

**Abstract.** This paper deals with the relations between ontologies and lexicons. We study the role of these two components and their evolution during the last years in the field of Computational Linguistics. Subsequently, we survey the current lines of research at ILC-CNR which tackle this topic. They involve (I) the reuse of already existing Lexical Resources to derive formal ontologies, (II) the conversion and combination of terminologies into rich and formal Lexical Resources and (III) the use of formal ontologies as the backbone of multilingual Lexical Resources.

#### 1 Introduction

During the last two decades, the Computational Linguistics community has dedicated considerable effort to research on and development of Language, and more specifically Lexical Resources (LRs), especially Computational Lexicons. Examples of large and informationally rich LRs include WordNet [2], EuroWordNet [16] and SIMPLE [7]. WordNet is a monolingual LR for English inspired by current psycholinguistic theories of human lexical memory. EuroWordNet follows the WordNet model and consists of a set of lexicons for many European languages cross-linked by means of a flat interlingual index. Finally, SIMPLE is a LR partially based on the Generative Lexicon theory and made up of a set of lexicons for twelve European languages.

These LRs, even though belonging to different linguistic approaches and theories, share a common element; all of them contain, explicitly or implicitly, an ontology as the means of organizing their structure. Therefore, these LRs can be seen as ontologically based lexicons, in that at least part of the lexicon architecture reflects the typical structure of an ontology. As noted by Hirst [6], word senses can be seen as the equivalent of ontological categories, and lexical relations would correspond to ontological relations (for example, hyperonymy would stand for subsumption).

In recent years, there has been increasing interest in formal ontologies within the Computational Linguistics community. In fact, formal ontology languages have become common within the area; the most widely spread are KIF and OWL which belong respectively to the First Order

Logic and Description Logics paradigms. The main advantage of formal ontologies is that they allow for reasoning and inferencing and, therefore, offer a great potential for applications related to the Semantic Web and to Natural Language Understanding/Processing.

From all these experiences, it becomes clear that LRs would benefit from incorporating formal ontologies and for ontologies in their turn would benefit from an easy access to typical linguistic information. As stated by Niles and Pease [9], "in a lexicon, the meaning of words relies on human interpretation rather than on a precise mathematical specification". Thus, especially for multilingual scenarios, "a purely linguistic representation is not enough to enable correct cross-linguistic mappings whereas a formal definition can function as an universal language and enable creators of lexicons for different languages to verify cross-language links against formal, logical definitions".

Among formal ontologies nowadays widely used in Computational Linguistics, two well-known examples are the Suggested Upper Merged Ontology (SUMO) [10] and DOLCE [5]. Moreover, researchers have begun to transform existing LRs into formal ontologies: a clear example is the conversion of WordNet into OWL [15]. LRs encoded as formal ontologies can be defined as lexically based ontologies, because on the one hand, they take the form of formal ontologies whereas on the other, they represent lexical knowledge.

The rest of the paper presents some current research activities of the ILC-CNR involving relations between ontologies and lexicons. Next section discusses the exploitation of existing lexicon data to automatically derive formal ontologies. In section 3, we report on a case study in the biology domain: the design and population of a lexical-teminological resource interlinked to a parallel formal ontology for gene regulation. Section 4 presents two big projects related to the use of formal ontologies in multilingual LRs (4). We end up by presenting some conclusions in 5.

## 2 Deriving Formal Ontologies from Existing Lexicon Data

The increasing importance of formal ontologies in LRs together with the availability of high quality and broad coverage, but non formal, computational lexicons developed during the last decades lead us to the question: can the lexical knowledge contained in these LRs be exploited to derive formal knowledge?

We decided to test this hypothesis by studying the conversion of the SIM-PLE ontology<sup>1</sup> into the formal language OWL [14]. The SIMPLE ontology is not defined in a formal language, but contains cardinal semantic constraints regarding relations and features encoded in a systematic way. For our purposes, we considered each element of the SIMPLE ontology

<sup>&</sup>lt;sup>1</sup> Acronym for Semantic Information for Multipurpose Plurilingual Lexicons; a Language Engineering project funded by the Eurpean DG-XIII. This project has been continued at the national level with the CLIPS project. Within SIMPLE lexicons for 12 European languages were built according to a shared top ontology.

and translated it into an OWL appropriate correspondent. Furthermore, we enriched the formalised ontology with semantic information from the SIMPLE Italian lexicon ([12]) which was first generalised and afterwards included in the ontology.

The systematic organization of SIMPLE, thanks mainly to the template objects, together with the high quality of the information it contains, allowed us to implement a fully automatic procedure to convert its ontology in OWL and to enrich the resulting ontology with information extracted from the lexicon. Such information consists of quantifier restrictions, predicates and constraints regarding relations and features. By means of an automatic generalisation of such information, the original ontology has been expanded and enriched. The resulting ontology is formal, semantically rich and yet it preserves the multi-dimensionality of the original SIMPLE ontology.

Let us clarify the procedure by means of presenting a sample of the transformation. Consider the information contained in the template of the SIMPLE ontology for the node "Artifact food" (see table 1). It consists of a set of cardinal restrictions regarding relations and features. For example, it states that an entry that belongs to this node should instantiate at least one relation of the type "Createdby", while it should instantiate exactly once a feature of the type "PLUS\_EDIBLE".

Table 1. Template for the ontology node Artifact\_Food

| Item                | Type     | constraint | value |
|---------------------|----------|------------|-------|
| Createdby           | relation | >=1        |       |
| Madeof              | relation | >=0        |       |
| Objectoftheactivity | relation | >=1        |       |
| PLUS_EDIBLE         | feature  | =1         |       |

Figure 1 presents the resulting transformation of this ontology node. As it can be seen, not only information from the template (cardinality restrictions) is included, but also knowledge obtained from the lexicon by means of generalisation: quantifier restrictions. Now, not only we can state that an entry should instantiate the "Createdby" relation, but we are also able to state which nodes the target of the relation should belong to: in this case "Cause\_Change\_of\_State". It is also important to mention that ontology nodes inherit the knowledge introduced at higher levels. We can see in the figure additional knowledge to that explicitly encoded for the node "Artifact\_food"; there is also inherited knowledge from the ancestor nodes "Food" and "Entity".



Fig. 1. Asserted conditions for the class "Artifact\_Food" in the resulting OWL ontology

# 3 From Domain Terminologies to Semantically Rich and Formal LRs

Formally, terminology is the study of terms - i.e. words used in specific domains - which aims at systematising and promoting correct usage, to help technicians developing documentation or translators optimize their translations. As such, terminologies are more concerned about the conceptual organisation of terms in special fields than on their linguistic behavior in texts and are therefore close to ontologies, or better taxonomies. However, while an ontology generally needs to represent formally consistent and coherent relations among terms/concepts because its main use is to reason about properties of the domain, a terminology needs not do so.

Terminologies, moreover, present shortcomings when compared to structured LRs (formal ontology plus lexicons), especially if we want to use them to automatise some steps of the information extraction/knowledge acquisition process. On the one hand, terminologies lack the linguistic information that is typically encoded in lexicons and used by text mining and NLP applications. On the other, they do not need to represent knowledge formally, whereas formal ontologies do, which allows for automatic reasoning.

Despite the aforementioned disadvantages, terminologies are widely applied in specialised domains. In biology/biochemistry, for example, there are many different terminologies -often incompatible one another-which are characterised by the two problems already pointed out. These limitations, together with the fact that biology/biochemistry is a very dynamic domain in which papers and written material with vital knew knowledge are continuously being produced, generates an increasing demand for comprehensive LRs that allow for semantic interoperability and text-based knowledge harvesting. The challenge here is to build large resources which integrate properties of computational lexicons, terminologies and formal ontologies, so that new technology can be applied and/or developed combining various NLP techniques and applications.

At ILC, we tackled this challenge within the BOOTStrep project<sup>2</sup>, in which one of the main goals is to organise, transform and combine ex-

<sup>&</sup>lt;sup>2</sup> BOOTStrep (Bootstrapping Of Ontologies and Terminologies STrategic Project) is a Specific Targeted Research Project of the European Unions 6th Framework

isting terminologies of the biology/biochemistry domain into LRs which will support information extraction and text mining. This is achieved by combining the richness of information of lexicons and the power of formal ontologies. Within this project two parallel resources have been created, the BioLexicon, which follow lexical and ontological standards, and the BioOntology. The two resources are linked together via explicit references, thus bringing together the formal rigor and semantic expressiveness of formal ontologies and the high coverage and the linguistic information (term variants, synonyms, parts-of-speech, verb subcategorisation, etc.) of lexicons (for details see [11], [8]). A reusable BioLexicon with sophisticated linguistic information, linked to a bio-ontology, will enable the bio-informatic community to develop information extraction tools of higher quality.

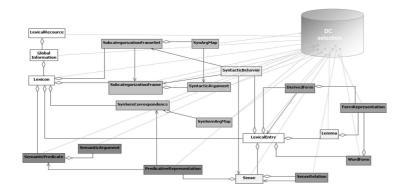
The BioLexicon should be considered as a customization of the LMF meta-model based on the requirements gathered both from the biomedical and the text mining community. It is modeled in an XML DTD according to LMF and is realised as a MySQL database: it implements the core LMF model plus objects from the NLP extensions for the representation of morphological, syntactic and lexical semantics aspects of words and terms. The model consists of a number of independent lexical objects (or classes) and a set of Data Categories (see fig. 2). In conformity to the ISO philosophy, the Data Category Selection for the BioLexicon is partially drawn from the ISO 12620 Data Category Registry [1], and partially defined for the specific purposes of the project and the special domain.

The BioLexicon encodes almost 2 millions (English) terms related to the bio-domain and represent their morphological, syntactic and lexical semantic properties. Among these terms, around 3K are biologically relevant verbs and nominalizations, i.e. predicates typically used in biomedicine to refer to bio-events. For such lexical items a full explicit representation of their syntactic complementation and semantic argument structure is also represented.

### 3.1 The Semantic Extension

The semantic module of the lexicon is made of lexical objects related to the Sense class. Sense represents lexical items as lexical semantic units. Semantic relatedness among terms is expressed through the SenseRelation class, which encodes (lexical) semantic relationships among instances of the Sense class. The SemanticPredicate class, instead, is independent from specific entries and represents an abstract meaning together with its associated semantic "arguments". It represents a meaning that may be shared by more senses that are not necessarily considered as synonyms. It is referred to by the PredicativeRepresentation class, which represents the semantic behavior of lexical entries and senses in context.

Programme within IST call 4. Six partners from four European countries (Germany, U.K., Italy, France) and one Asian partner from Singapore are involved in the project. www.bootstrep.eu



 ${f Fig.\,2.}$  The BioLexicon abstract model

A direct link to the BioOntology is established at Sense level by an URI pointing to either a same-level concept, or to a higher one, in the ontology. Such links are derived automatically from the input data- elaborated by the group of data and information extractors.

Additionally, the SemanticPredicate may also point to a concept in the BioOntology, so that predicates with their semantic argument information can be linked to the formal properties represented in the Ontology, thus allowing for complex inferencing.

# 4 Formal Ontologies as the Backbone for Multilingual LRs

We are currently working on two projects which, although dealing with different languages and goals, share a common aspect: the use of a formal ontology as the means to connect multilingual lexicons. These projects are: KYOTO ("Knowledge-Yielding Ontologies for Transition-Based Organization"<sup>3</sup>), funded by the European Union and "Developing International Standards of Language Resources for Semantic Web Applications" [13], funded by the International Joint Research Program of the New Energy and Industrial Technology Development Organization (NEDO) programme. The following paragraphs, discuss the role of formal ontologies in these two projects.

### **4.1 KYOTO**

The goal of KYOTO is to construct a language-independent information system for a specific domain (environment/ecology) anchored to a language-independent ontology that is linked to wordnets in several languages. KYOTO will allow both experts and the general public to access large amounts of distributed multimedia data from wide-spread sources

<sup>&</sup>lt;sup>3</sup> www.kyoto-project.eu

in a number of culturally diverse languages by means of deep semantic search technology. Figure 3 presents the general architecture of this project.

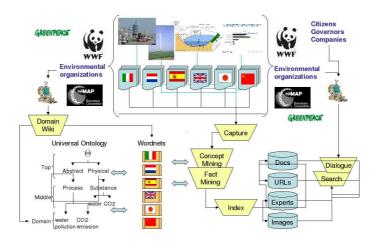


Fig. 3. Kyoto architecture

The project can be seen as a testcase on which to build further interoperable lexical resources based on the lexicon-ontology duality, as it implies putting together wordnets for seven languages from different language families and connecting them via a formal ontology. The KYOTO ontology is, thus, the core of the resource architecture. It is a language and culture neutral object that connects the lexicons for different languages. Besides, it encodes semantic constraints which can be used to perform fact extraction from text. This ontology is made up of three layers (top, middle and domain):

- The top layer will be based on existing top level ontologies, among them SUMO and DOLCE.
- The middle layer will be derived from existing wordnets, where concepts are mapped to lexical units.
- The domain layer will be derived from domain source documents.
   Terms are semi-automatically extracted from these sources. This layer can be also extended manually by domain experts by using a graphical interface.

### 4.2 NEDO

The aim of NEDO is to create a common standard for Asian LRs (Chinese, Japanese, Korean and Thai; plus Italian) by adapting the ISO Lexical Markup Framework (LMF)[4], which was originally designed taking in to account mainly Indo-European languages. NEDO developed a conceptual core for a multilingual ontology with the main focus on Asian language diversity, and a multilingual LMF-conformant core lexicon.

Central activities in the project are thus the building of sample lexicons, an upper-layer ontology to which the lexicons are linked, and an evaluation of the proposed framework through an application (as in 4).

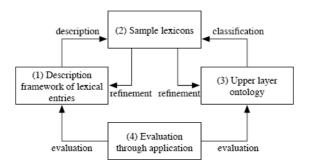


Fig. 4. Nedo Research Cycle

The lexical specifications are based on and compliant with LMF which, among other properties, supports the linking of lexical entries to ontology elements both in monolingual and multilingual scenarios.

Each language team has built sample lexicons following this standard. The starting point to build the lexicons has been the Swadesh list and its translations in the languages involved in the project. The list can be seen as a least common denominator for vocabulary. The coverage of the Swadesh list has been compared with the one of the Base Concept Set (BCS) as it is proposed by the Global WordNet Association. The reason for considering the Swadesh as a potential core lexicon resides in the fact that some of the languages considered suffer from lack of available resources.

Due to the fact that the Swadesh list has been mapped to SUMO, we have been able to further rely on the latter for developing the core ontology. Apart from SUMO, the middle layer ontology MILO was also considered because the first was too unspecific for some areas of the lexicon. The Swadesh-SUMO/MILO ontological structure was then pruned in order to obtain a proper ontology for representing the concepts in the Swadesh list.

The experiments performed within the NEDO project yielded very promising results and suggest that a similar approach, or the addition of lexica for the other asian languages would be highly advantageous for the whole community.

### 5 Conclusions

Through a series of project on which we have been working at ILC-Pisa, this paper has surveyed the important role that ontologies play for lexical resources and their evolution in this area. In a nutshell, it can be said that LRs are evolving from ontologically based lexicons to lexically based ontologies.

The paper explored the current trends of using formal ontologies as a core module of LRs which present important advantages, especially in multilingual and terminological environments.

The article has introduced the experience of two projects in which formal ontologies play a central role in the context of multilingual LRs: the ontology in fact acts as the glue for connecting different monolingual lexicons and provides cross-lingual reasoning capabilities.

Apart from this, the work briefly presented in 2 proves that the knowledge already present in *non formal* LRs can be exploited to derive formal ontologies without much manual effort if the LRs were systematically built. In fact, the structure of the SIMPLE model, and especially its templates, has proven useful to convert its ontological information, and even lexical data by means of generalisation procedures, to the formal language OWL.

In 3 we described the structure of a rich lexical-terminological resource for biology, which is directly linked to a parallel domain ontology. The combined resource will allow for intelligent content access in a domain where it is of utmost importance for researchers to be able to access the new knowledge and facts that are being experimentally discovered on a daily basis.

Finally, section 4 presented two projects aiming at bringing together multilingual resources of different and distant languages, to foster real interoperability and exchange.

In spite of the widely usage of ontologies in LRs and of the obvious advantages it entails, some issues remain open. The main issue concerns the boundaries between ontology and lexicon, which nowadays are still not clear. What kind of information should be in the ontology and what instead in the lexicon? A recent paper by Pease and Fellbaum [3] addresses this topic and concludes that, in principle, conceptual relations should be in the ontology whereas lexical relations (word-to-word) such as synonymy should be in the lexicon.

#### References

- Gil Francopulo et al. Lexical markup framework (lmf). In Proceedings of the LREC 2006, Genova, Italy, 2006.
- 2. Christiane Fellbaum. WordNet: An Electronic Lexical Database (Language, Speech, and Communication). The MIT Press, May 1998.
- 3. Christiane Fellbaum and Piek Vossen. Challenges for a global wordnet. In *Proceedings of the First International Workshop on Global Interoperability for Language*, 2008.
- G. Francopoulo, N. Bel, M. George, N. Calzolari, M. Monachini, M. Pet, and C. Soria. (forthcoming) Multilingual resources for NLP in the Lexical Markup Framework (LMF). Language Resources and Evaluation Journal, 2008.
- Aldo Gangemi, Nicola Guarino, Claudio Masolo, Alessandro Oltramari, and Luc Schneider. Sweetening ontologies with dolce. pages 166–181. Springer, 2002.
- Graeme Hirst. Ontology and the lexicon. In Steffen Staab and Rudi Studer, editors, *Handbook on Ontologies*, International Handbooks on Information Systems, pages 209–230. Springer, 2004.
- Alessandro Lenci, N. Bel, F. Busa, N. Calzolari, E. Gola, M. Monachini, A. Ogonowski, I. Peters, W. Peters, N. Ruimy, M. Villegas, and A. Zampolli. Simple: A general framework for the development of multilingual lexicons. *International Journal of Lexicography*, 13(4):249–263, 2000.
- 8. Monica Monachini, Nilda Ruimy, Valeria Quochi, and Nicoletta Calzolari. Lexical relations and domain knowledge: The biolexicon meets the qualia structure. In *Proceedings of the GL2007 Conference*, Paris, France, May 2007.
- 9. I. Niles and A. Pease. Linking lexicons and ontologies: Mapping WordNet to the suggested upper merged ontology. In *Proceedings of the 2003 International Conference on Information and Knowledge Engineering (IKE '03)*, Las Vegas, Nevada, 2003.
- 10. Adam Pease, Ian Niles, and John Li. The Suggested Upper Merged Ontology: A large ontology for the semantic web and its applications. In Working Notes of the AAAI-2002 Workshop on Ontologies and the Semantic Web, Edmonton, Canada, 2002.
- Valeria Quochi, Monica Monachini, Riccardo Del Gratta, and Nicoletta Calzolari. A lexicon for biology and bioinformatics: the bootstrep experience. In European Language Resources Association (ELRA), editor, Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), Marrakech, Morocco, may 2008.
- Nilda Ruimy, M. Monachini, R. Distante, E. Guazzini, S. Molino, M. Ulivieri, N. Calzolari, and A. Zampolli. Clips, a multi-level italian computational lexicon: A glimpse to data. In *Proceedings of the Third International Conference on Language Resources and Evalua*tion (LREC'02), Las Palmas de Gran Canaria, Spain, 2002.
- Tokunaga Takenobu, Dain Kaplan, Chu-Ren Huang, Shu-Kai Hsieh, Nicoletta Calzolari, Monica Monachini, Claudia Soria, Kiyoaki Shirai, Virach Sornlertlamvanich, Thatsanee Charoenporn, and Xia

- YingJu. Adapting international standard for asian language technologies. In European Language Resources Association (ELRA), editor, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may 2008.
- Antonio Toral and Monica Monachini. SIMPLE-OWL: a Generative Lexicon Ontology for NLP and the Semantic Web. In Workshop on Cooperative Construction of Linguistic Knowledge Bases (AIIA 2007), 2007.
- 15. Mark van Assem, Aldo Gangemi, and Guus Schreiber. Conversion of WordNet to a standard RDF/OWL representation. In *Proceedings* of the Fifth International Conference on Language Resources and Evaluation (LREC'06), Genoa, Italy, May 2006.
- Piek Vossen. Eurowordnet a multilingual database with lexical semantic networks, 1998.