Interfacing Ontologies and Lexical Resources

Laurent Prévot, Stefano Borgo and Alessandro Oltramari
Laboratory for Applied Ontology, ISTC-CNR, Trento, Italy
{prevot,borgo,oltramari}@loa-cnr.it

Abstract

During the last few years, a number of works aiming at interfacing ontologies and lexical resources have been initiated. This paper aims at clarifying the current picture of this domain. It compares ontologies built following different methodologies and analyses their combination with lexical resources. A point defended in the paper is that different methodologies lead to very different characteristics for the resulting resources. We classify these methodologies show how actual projects fit into this classification.

1 Introduction

During the last few years, ontologies and lexical resources have been put under the spotlight for dealing with various NLP tasks such as word sense disambiguation and bridging resolution. Interfacing ontology and computational lexicon[1] has been presented as a promising approach for Human Language Technologies (HLT), from classical NLP tasks to meaning negotiation in multi-agent systems. In this paper we aim at clarifying the populated landscape of the on-going initiatives in the domain. We will introduce in section 2 our methodology classification for combining ontologies and lexical resources. Then we will survey some of the most popular top-level ontologies, namely DOLCE (Masolo et al., 2003), OPENCYC[2] and SUMO (Niles and Pease, 2001). These ontologies are quite different although this might not be evident to the newcomer. The purpose of the section 3 is to highlight the methodologies used for building them. In section 4, on the ground of the first two sections we will show how actual initiatives fit into our classification. The lexical resources considered in the paper are basically those of the WORDNET family (Fellbaum, 1998). We will conclude with some comments on multi-linguality issues.

2 Classifying experiments in ontologies and lexical resources

The main aim of interfacing ontologies and lexical resources is the development of *machine-understandable knowledge bases* to be used in Human Language Technologies. The need for such integrated knowledge resources is a central issue for the next generation tools envisaged by the Semantic Web, where knowledge sharing, information integration, interoperability and semantic adequacy are main requirements.

Different methods may guide the linking of ontologies and lexical resources, depending on the final result one intends to achieve, namely to enhance the coverage of an ontology or to build a system comprising properties of an ontology and a lexical resource. A generalization of these tasks may suggest the following methodological options:

(i) *restructuring* a computational lexicon on the basis of ontological-driven principles;

(ii) *populating* an ontology with lexical information;

(iii) *aligning* an ontology and a lexical resource

[1] The terms "computational lexicon" and "lexical resource" are often used as synonyms in the literature.

[2] See *http://www.opencyc.org/releases/doc/*

The first option (i) concentrates on the lexical resource and involves the ontology only at the "meta-level": the ontological restructuring is carried out following formal constraints of ontological design (Guarino and Welty, 2004), for instance introducing the ontological distinction between *role* or *type* for concepts.

On the other side, (ii) requires to map lexical units to ontological entries, focusing on the "object-level" (Niles and Pease, 2001): in this case the formal constraints correspond to ontological categories and relations already implemented in an existing ontology. In this simplifying view, a computational lexicon and an ontology are taken as bare taxonomies of terms, the former contains only lexicalised concepts (i.e. `substance`)[3] and linguistic relations (i.e. *hyponymy*) while the latter provides formal structure of both lexicalised and not-lexicalised concepts (i.e. AMOUNT-OF-MATTER) and relations (*part-of*). It is clear then that this second method has to include a comparative analysis of the ontology and the lexical resource in order to find bridging synonymous terms, including the search for cases of possible homonyms too.

Finally (iii), the most complete of the proposed approaches, collects both the "meta-level" and "object-level" character of the previous approaches in order to produce a system that is ontologically sound and linguistically motivated (Oltramari et al., 2002).

The experimental perspectives focused in this paper will show that ontologies and lexical resources generally keep their own peculiarities in the process of integration: in other words, neither (ii) nor (iii) bring to an actual *merging* of ontological properties and lexical information.[4] Although it is possible for different ontologies to be coherently merged in a new one - associating semantically similar concepts and finding the points of intersection (Taboada et al., 2005) - the real benefit of integrating ontologies and computational lexicons follows from keeping them as *distinct layers of semantic information*, albeit improved by their mutual linkings and features. This is the main reason why we choose the term *alignment* (and not *merging*) for the most advanced interfacing method, i.e. (iii).

As stated above, both ontologies and lexical resources may be built around a taxonomic structure; however they often include information of different type as well. Consider an *axiomatic ontology* like DOLCE (Masolo et al., 2003): it provides an axiomatisation of *part-of, constitution, dependence, participation*, that is, it characterizes several non-hierarchical relations. Similarly, a lexical resource like the Princeton WORDNET (Fellbaum, 1998) is organised as a *semantic network*, whose nodes (sets of synonym terms) are bound together by several lexical and conceptual relations (besides *hyponymy/hyperonymy* we have meronymy, antonymy, causation, entailment and so on). This fact suggests the introduction of another dimension here called *contraint density*, which, as far as we know, has not been considered in the literature.

Constraints density captures the density of the "network of constraints" that holds between the concepts. It can be opposed to the *concept density* that situates ontologies from top-level to domain-level (see Fig. 1). *Constraint density* deals with non-hierarchical features of ontologies and lexical resources, like extension with axioms for dependence, participation and constitution, formalization of meronymy relation, translation of glosses into axioms and consistency checks (See for instance (Gangemi et al., 2003b)).

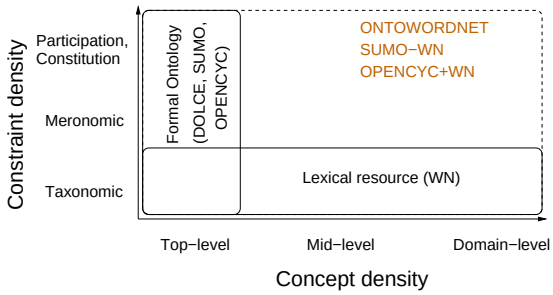

Figure 1: Concept and constraint density

To make an analogy with the ontology development terminology, resources having very dense constraint network correspond to *heavyweight* ontologies while loose constraint network can be asso-

[3] In the paper we will stick to the following font convention: `typewriter` for WORDNET synsets, SMALLCAPS for ontology concepts and *italics* for relations.

[4] Of course method (i) is not considered since it provides only "ontological-driven principles" without any real ontological category or relation.

ciated to *lightweight* ontologies (Guarino, 1998a; Gómez-Pérez et al., 2004). Lexical resources are conceptually very dense but they do not have a dense network of constraints. On the other hand, ontologies, specially top-level ones, are not densely populated but offer a dense network of constraints for their concepts.

A final remark needs to be done about the nature of the lexical resources we look at. Although the experiments we consider in sections 4 and 5 concern the interfacing of ontologies with Princeton WORDNET, the methodologies we present here are general and apply to other resources as well, like computational lexicons built on the basis of the original Princeton resource (i.e. EUROWORDNET[5] (Vossen, 1998) modules). The three methods we isolated have not been applied to other types of lexical resources, for example FRAMENET[6]. We believe this is a question of time since nowadays in the literature "Wordnets" is the *de facto* standard for interfacing. We expect that further experience with different kinds of lexical resources will shed new light on the advantages and drawbacks of the three methodologies.

3 Ontologies and their construction

Ontology, as understood in the area of knowledge representation, is a young research area with several weaknesses among which the lack of established methodologies as well as of evaluation criteria. Thus, one should not be surprised if the ontologies today available have been built following disparate approaches resulting in quite different systems. This is particularly evident in the area of *top-level ontology* by which, in this paper, we mean the research in *formal* and *foundational* ontologies. These ontologies are knowledge structures that (1) adopt a rich formal language (generally some kind of first-order logic) and (2) aim at classifying basic notions of general interest like process, event, object, quality, and so on.

Here we concentrate on three top-level ontologies, namely DOLCE, OPENCYC, and SUMO, that well indicate this variety of approaches. However, the main reason to focus on these is their attention to linguistic resources: these are the systems that have been used explicitly in relationship WORDNET.

3.1 DOLCE

DOLCE[7] (a Descriptive Ontology for Linguistic and Cognitive Engineering (Masolo et al., 2003)) was released in its actual version in 2003 and has been constructed according to well documented philosophical principles. The content of the ontology is motivated from a cognitive viewpoint since the overall aim is to capture the ontological categories underlying everyday language and human commonsense. This view explains the adoption in DOLCE of a multiplicative approach which justifies the existence of co-localized (yet different) objects. For instance, DOLCE claims that a statue and the clay of which it is made, are different entities which share the same spatial (and possibly temporal) location. Co-localized entities are needed to consistently model linguistic expressions in which incompatible properties seem to be referred to the same object: a scratched statue is different (since scratched) and yet it is the same statue it was before. In DOLCE this is possible since the statue itself might not be affected by (minor) scratches, but the clay does because parts of it break up. DOLCE includes very basic and general notions only providing a total of about 40 categories which are richly axiomatized by using about a 100 relations and 80 axioms.

In the paper, we consider the "lite" version of DOLCE (aka DOLCE-LITE+), namely an extension of the axiomatic ontology that do not consider modality, temporal indexing, and relation composition. This version contains more concepts and allows for the implementation of DOLCE-based resources (i.e. the alignment of DOLCE and WORDNET called ONTOWORDNET) in languages that are less expressive than FOL e.g. OWL-DL, OWL-Lite, and RDF.

DOLCE is public resource and is released under the Lesser GNU Public License.

3.2 OPENCYC

OPENCYC is the ontology of CYC, a project initiated in 1984 with the aim of building a knowledge base comprising both scientific and common-

[5] See *http://www.illc.uva.nl/EuroWordNet/*

[6] Based on frame semantics (Fillmore, 1976). See *http://framenet.icsi.berkeley.edu/*

[7] See *http://www.loa-cnr.it/DOLCE.html*

sense knowledge. CYC grow to include hundreds of thousand elements between atomic terms, concepts, and axioms. To overcome consistency issues, CYC is now subdivided in hundreds of "microtheories". Microtheories are, roughly speaking, bundles of assertions and rules in a specific domain of knowledge and are supposed to be locally consistent although not official claim is made in this sense. OPENCYC is a byproduct of CYC and was not part of the original project. Unfortunately the OPENCYC ontology has not been constructed according to philosophical principles nor following an ontological tested methodology. Indeed, still today the focus is on coverage: in the website one reads that OPENCYC includes "an upper ontology whose domain is all of human consensus reality", which explains the 47,000 concepts and more than 300,000 assertions it contains, but makes one wonder what "upper" means here! Initially, it was obtained by isolating the taxonomy of the most general notions in CYC (perhaps with minor adjustments) but it was never followed by an ontological analysis and study of these notions. One can observe that OPENCYC adopts (at least in part) a cognitive viewpoint since some categories capture naïve conceptions of "reality". For this reason, OPENCYC is compatible with the multiplicative approach (as seen in DOLCE) although this has not been followed in a systematic way. Since we lack a characterization of the ontological commitment and an analysis of the ontological choices embedded in the OPENCYC hierarchy there is not much to say about its ontological relevance. A further problem is the scarce axiomatization of OPENCYC which makes impossible to analyze the adequacy of the system in formal ontology.

OPENCYC is publicly available under the GNU Lesser General Public License.

3.3 SUMO

SUMO (Suggested Upper Merged Ontology (Niles and Pease, 2001)) began as a potpourri of theories in the knowledge representation area among which (Sowa, 1995; Borgo et al., 1996; Allen, 1984; Smith, 1996). The ontology was created for computer applications (data interoperability, information retrieval, etc.) with no philosophical concerns and did not adopt ontological principles. This attitude is still present today (notwithstanding sporadic claims that SUMO is "rooted in metaphysical naturalism"), and the overall system is ontologically unclear as pointed out several times in the SUO mailing list.[8] Still, one can recognize some ontological choices in the system like the distinction between objects and events, and the adoption of a realistic approach. However, there is no guarantee that these have been consistently exploited in the whole ontology. The last version was released in 2005 and consists of about 4,000 assertions and 1,000 concepts. Several domain ontologies, linked to SUMO, are also available.

In the paper, we consider also the middle level ontology called MILO. MILO is written in the same language of SUMO and is provided as a "bridge" system between the general ontology and a number of domain ontologies. The latest version available on the web has been released in July 2004 and is marked "provisional and incomplete". We consider it since it is an integral part of the research in ontology and linguistic resources based on SUMO.

SUMO was initially distributed under the GNU Licence. Now it is subject to other restrictions;[9] in particular, SUMO "must not be utilized for any conformance/compliance purposes" and "[...] entities seeking permission to reproduce this document, in whole or in part, must obtain permission." However, these restrictions should not apply to research work.

4 How actual resources fit the classification

Generally speaking, projects interfacing ontologies and lexical resource are not easy to compare since beside generic statement, objectives are rarely addressed and the results are not homogeneously evaluated. Our classification of methodologies is simply a framework to put some order and to situate these resources. It is not meant to be a measure for ranking these resources. This section shows how actual resources fit into the classification.

4.1 ONTOWORDNET

The work underlying the ONTOWORDNET project is rooted in early proposals about upper levels of lexical resources (Guarino, 1998b). More recent presentations can be found in (Oltramari et al., 2002;

[8] See *http://suo.ieee.org/index.html*

[9] See *http://ontology.teknowledge.com/IEEE_license.htm*

Gangemi et al., 2003a). The program of ONTOWORDNET includes:

1 to reengineer WORDNET lexicon as a formal ontology, and in particular:
   1.1 to distinguish synsets that can be formalized as classes from those that can be formalized as individuals;
   1.2 to interpret lexical relations from WORDNET as ontological relations.
2 to align the top-level of WordNet to a foundational ontology, to allow for re-interpretation of hyperonymy when it is the case;
3 to check the consistency of the overall result, and to correct the cases for inconsistency;
4 to learn and revise formal domain relations (either from glosses or from corpora).

The first point clearly addresses the restructuring task we presented in section 2 while points (2) and (3) deal with populating an ontology. The last point addresses the orthogonal issue of *constraint density* (axiomatizing the glosses).

ONTOWORDNET project highly relies on the ONTOCLEAN methodology (Guarino and Welty, 2004). This methodology proposes to determine which meta-properties hold for a given property. Very roughly, a *rigid* property is a property that is essential to all its instances while a *non-rigid* property is not and a *anti-rigid* is essential to none of them. Some properties (called sortals) are carrying with them an *identity* criterion. A property (ϕ) can be said to be *dependent* on another one (ψ) if for all instance of ϕ some instance of ψ must exist (and it is not a part or a constituent of it).[10] Finally, another meta-property we will use in our example (section 5) is *unity*. "A property (ϕ) is said to carry *unity* (+**U**) if there is a *common* unifying relation R such that all the instances of ϕ are essential wholes under R. A property carries *anti-unity* ($\sim$**U**) if all its instances can possibly be non-wholes" (Gangemi et al., 2001).

The second step of the methodology consists of checking that a series of constraints involving these meta-properties are actually satisfied. For example, it is required for unitarian properties to not subsumes anti-unitarian ones, or properties that subsume rigid properties to be rigid themselves. This prevents, for example, roles to subsume types. More accurately we found in (Guarino and Welty, 2000) that roles are *non-rigid*, they do not supply their *identity criterion* but might carry one and they are *dependent* on other properties. Types, on the other side, are *rigid* and supply their own *identity criterion*. The first version of ONTOWORDNET was extreme on this point in requiring the removal of roles from the ontology. The last version softens the position and requires only to label roles for separating them from types.

This constraint checking is a crucial aspect of the ONTOWORDNET project. It is at this step that the lexical resource benefits from some ontological cleaning. ONTOWORDNET does not simply populate the top-level ontology by attaching WORDNET terms under ontology concepts. ONTOWORDNET determines which constraints have to be satisfied for integrating a WORDNETsynset in an ontology in order to preserve its properties. ONTOWORDNET also claims that WORDNET itself benefits from the reorganization and from the application of the constraints. A full description of these constraints can be found in (Oltramari et al., 2002; Gangemi et al., 2003a). However, the re-structuration has been performed systematically only on the third-fourth upper levels of WORDNET. The current ONTOWORDNET is therefore made of a re-structured, cleaned, upper level and of a simple copy of WORDNET at the lower levels (without any ONTOCLEAN check).

Finally, axiomatizing WORDNET glosses (in the spirit of XWN described in section 4.4) is a research objective of ONTOWORDNET project as well and it is currently pursued as shown in (Gangemi et al., 2003b).

In conclusion, ONTOWORDNET is a costly methodology that hasn't been apply of the totality of WORDNET but that offers general rules to clean the lexical resource and populate the ontology. This methodology fully corresponds to our third category, an *alignment* between a lexical resource and an ontology.

[10] This is a very rough introduction, for a detailed account see (Guarino and Welty, 2000) which provides an insightful account on properties and meta-properties and (Guarino and Welty, 2004) for an overview of ONTOCLEAN.

	Level	Examples
Re-structurating	Meta	ONTOCLEAN
Populating	Object	OPENCYC, SUMO-WN
Aligning	Object&Meta	ONTOWORDNET

Figure 2: Methodology classification

4.2 OPENCYC and WORDNET

The next proposal we present is the integration of OPENCYC with WORDNET. The integration is obtained by adding in OPENCYC synonym relationship between OPENCYC concepts and WORDNET synsets (Reed and Lenat, 2002). The purpose is simply to enrich the ontology with WORDNET information.

In our classification, this work falls into the *populating an ontology* option since there is no interest in restructuring the lexical resource nor in merging the two systems.

4.3 SUMO-WN

We call the third approach SUMO-WN (Niles and Pease, 2003), i.e. the integration of SUMO with WORDNET. This integration has been performed for nouns, verbs, adverbs and adjective synsets. The result is a new resource whose entries are WORDNET synsets tagged by SUMO categories. At first sight, this project seems to address the three methodologies we identified: (i) Re-structuring a lexical resource (tagging WORDNET entries with SUMO categories might constitute a first step for re-structuring WORDNET), (ii) Populating an ontology (tagging also allows to present WORDNET synsets as synonyms, hyponyms and instances of SUMO concepts), (iii) Aligning an ontology and a lexical resource because SUMO-WN concerns both methodologies.

This brief description of SUMO-WN integration makes it sound very complete. However we need to look closer at the methodology in order to understand exactly what is done in SUMO-WN.

The result of the interfacing between SUMO and WORDNET is a list of synset annotated with SUMO concepts. The main task is therefore the annotation. In (Niles and Pease, 2003) three unproblematic annotation cases are presented:

- the WORDNET synset is a *synonym* of an existing SUMO concept
- the WORDNET synset is an *hyponym* of an existing SUMO concept
- the WORDNET synset is an *instance of* of an existing SUMO concept

Unfortunately, the examples given in (Niles and Pease, 2003) are a bit confusing as we will see in our discussion of practical example (section 5). We will see later that ontology has been recently improved but since our focus is on the methodology we look at problems that arise from its application disregarding *ad hoc* solutions.

Another global problem for SUMO-WN is the absence of verification during the integration process. The quality of the resulting resource relies totally on the quality of WORDNET and SUMO. This is problematic since structural problems of WORDNET are now well-known and we saw in section 3 that the methodology for building SUMO jeopardize its use as a well-founded reference for annotating the resource. We believe that a more careful restructuring of WORDNET is required before populating the ontology, and only then an annotation with SUMO concepts might have its interest. SUMO-WN links are rather *ad-hoc* and it is difficult to figure out how such an approach can extend the accuracy of WORDNET or SUMO.

In conclusion, SUMO-WN addresses only the second category of our classification (*populating*) although the annotation of WORDNET entries could be seen as a preliminary step for re-structuring the resource. Moreover, since there is no clear methodology for determining how to perform the tagging, it would be dangerous to use this tagging for modifying the resource.

4.4 Axiomatizing glosses (EXTENDED WORDNET)

The EXTENDED WORDNET(XWN) project started with the objective of improving several weaknesses

of WORDNET. These weaknesses are described in (Harabagiu and Moldovan, 1998) and include in particular the need for more conceptual relations such as *causation* and *entailment* which are absent or not enough developed in WORDNET.

The proposal (Harabagiu et al., 1999) consists in *"translating"* WORDNET glosses into logical formulas with the help of natural language analysis. WORDNET glosses are in a first step parsed to produce "logical forms". The second step consists in the transformation of the "logical form" into "semantic forms" by taking into account finer semantic aspects such as thematic relations. WORDNET glosses eventually become axioms that can be manipulated in a more precise and efficient way than current natural language glosses. Moreover the disambiguation of the terms in the glosses and its systematic linking to other WORDNET entries and to terms present in other glosses augment dramatically the connectivity between WORDNET synsets.

This work is very promising and is complementing the previous approaches presented in sections 4.1 and 4.3 which at this point provides mainly taxonomic axioms.

In our terminology, XWN wants to increase the constraint density since the axioms derived from this method are potentially of all types. XWN is not properly speaking proposing to interface an ontology and a lexical resource because it does not involve explicitly an existing ontology. Since the ontological input is implicit XWN does not enter into our classification. However, if this ontological input was coming from an existing ontology, XWN would belong to the *re-structuring* methodological option.

4.5 Summary

The result of our classification is summarized in table 2. Among the initiatives we looked at, OPENCYC is a clear example of a *populating* methodology, SUMO-WN fall also into this category while ONTOWORDNET includes both the *re-structuring* methodology through the application of ONTOCLEAN and the *populating* one by linking WORDNET synsets to DOLCE-LITE+ categories. Finally, SUMO-WN offers a complete integration of WORDNET while ONTOWORDNET and OPENCYC are, for different reasons, incomplete.

5 Two practical examples

5.1 `Christian_Science` and `Underground_Railroad` examples

The first example comes from the SUMO-WN presentation (Niles and Pease, 2003). It concerns the *hyponym* case. It is claimed that the SUMO concept RELIGIOUSORGANIZATION is a hypernym of WORDNET synset `Christian_Science` *(gloss: "religious system based on the teachings of Mary Baker Eddy emphasizing spiritual healing").* Since RELIGIOUSORGANIZATION are ORGANIZATIONS, there is no clear reason for setting `Christian_science` to be an organization because SUMO organizations are "*corporate or similar institutions (...)"*. The corresponding category for `Christian_Science` should be something like *Christian_Science_Church*. There are actually another WORDNET synset for `Christian_Science` *(gloss: "Protestant denomination founded by Mary Baker Eddy in 1866.").* But even accepting this conceptual shortcut, it is still not clear why `Christian_Science` is an *hyponym* of RELIGIOUSORGANIZATION and not an *instance-of* it. For `Christian_Science` to be a sub-type of ORGANIZATION, there must be at least two instances of `Christian_Science`. The WORDNET gloss describes it more as a general doctrine and therefore as an instance of something like a *Religious_System*. The example provided fits better the second WORDNET synset.

We don't have enough information about the notions of religious systems and organization to pursue further the investigation but it is clear to us that the choices that have been made in SUMO on these topics are dubious. The current version we can find online[11] corrected some of these problems as we can see in figure 3. In the downloadable file, we can find both synsets. The first one is now an *instance-of* RELIGIOUSORGANIZATION while the second one is a sub-type of SUMO's PROPOSITION.

The example provided for illustrating the *instance-of* case is very similar to the previous one. In this case, `UndergroundRailroad` *(gloss: abolitionist secret aid to escaping slaves)* is taken to be an *instance-of* and not an *hyponym*

[11] See *http://ontology.teknowledge.com/*

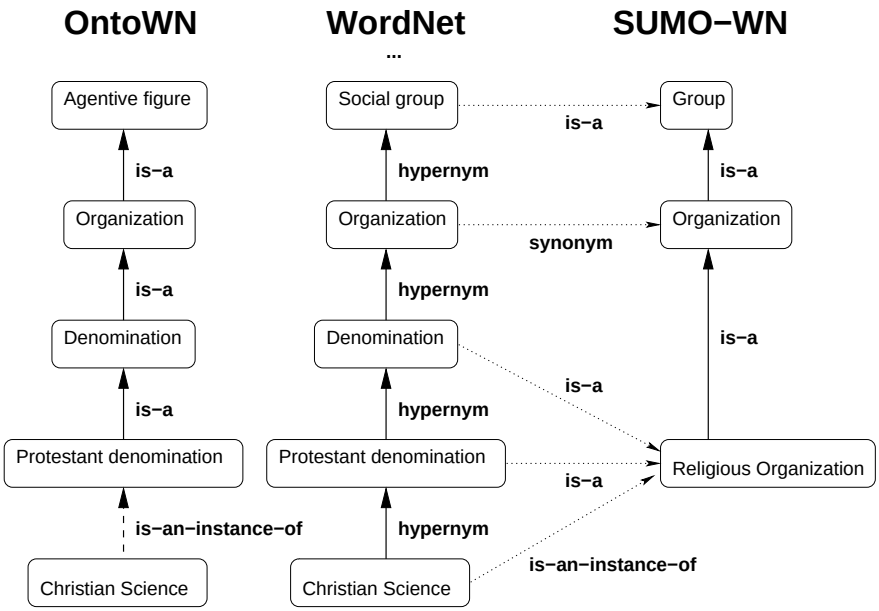

Figure 3: `Christian_Science` example

of SUMO ORGANIZATION. In the end, it remains difficult to understand the methodology adopted to classify terms in these two examples one wonders if there is something more than the intuitions of the SUMO-WN developers.

In ONTOWORDNET `Christian_Science` and its hyperonyms are integrated in the resource as shown in figure 3. The hierarchy corresponds to the WORDNET one until the top-level of the hierarchy. About the first sense, the last WORDNET hypernym is `Organization` and there is ORGANIZATION present in DOLCE-LITE+. The second sense is more tricky because of a double inheritance in the WORDNET hierarchy.

Regarding the `Underground_Railroad`, the ONTOWORDNET version proposed it as a subtype of `Escape`. It is a clear example that shows that the application of the methodology is incomplete in the current version of ONTOWORDNET. Because of its development cost, the checking and the restructuration of WORDNET couldn't go deeper than the first four upper levels of the hierarchy. As a result, `Underground_Railroad` hasn't been checked and therefore not corrected yet in ONTOWORDNET.

5.2 `Cement` example

The second example concern the need for WORDNET restructuration. In WORDNET `cement` (*gloss: "a building material that is a powder made of a mixture of calcined limestone and clay") is situated under* `building_material` *and further under* `artifact` *(see Fig. 4). On the meta-properties level,* `Artefact` *presents therefore both unitarian concept such as regular artefacts (*`chair, hammer,...`*)* and non-unitarian object such as `cement`. This constitutes a formal violation in terms of ONTOCLEAN methodology.

In SUMO-WN this violation is repeated since `building_material` *is-a* SELFCONNECTEDOBJECT, which is an unitarian concept (+**U**) and SELFCONNECTEDOBJECT include FOOD which subsumes itself BEVERAGE, that is clearly non-unitarian (∼**U**).

On the other hand, ONTOWORDNET performs a re-structuration at this level which forces to distinguish unitarian and non-unitarian concepts as explained in (Oltramari et al., 2002). `building_material` is therefore removed from the `artefact` category and put under FUNCTIONAL-MATTER which *is-subsumed* by AMOUNT_OF_MATTER (∼**U**). The `artefact` synset is put under ORDINARY_OBJECT. Finally,

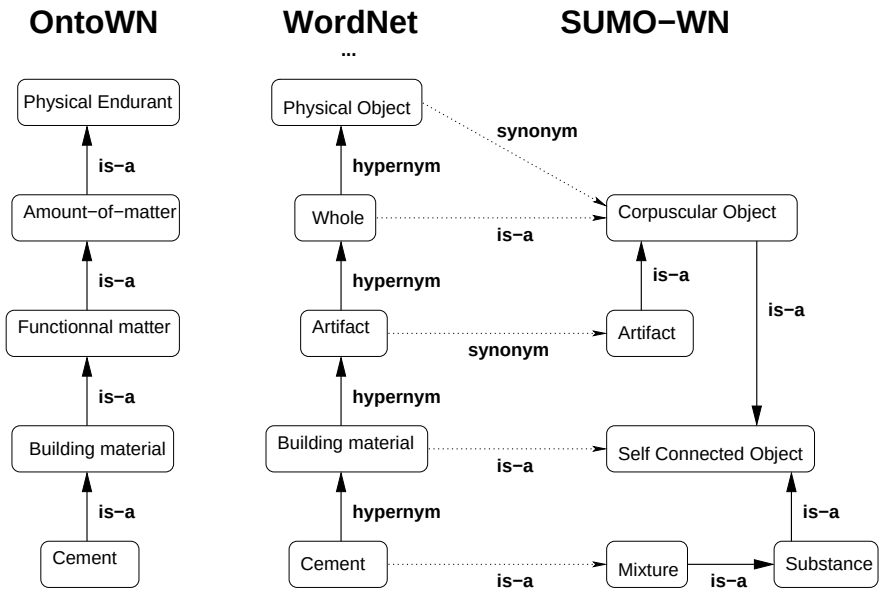

Figure 4: `Cement` example

we do not discuss, specific examples involving OPENCYC for lack of public material.

6 One or many resources?

The previous sections of this paper have emphasized the differences between existing proposals for interfacing top-level ontologies with lexical resources. These strong differences lead unsurprisingly to very different lexical resources or enhanced ontologies. The next natural step is to ask for what these tools are more adequate. Is each specific tool adequate to a given number of NLP tasks, or on the opposite, can we claim that some tool is suited for most of them? The soundness of the top-level ontologies seems to be crucial. Afterall this property is already essential for the quality of the top-level on which the whole architecture relies. However, soundness is very complex and costly to preserve when the size of the ontology increases. One might wonder if such hard constraint is absolutely necessary when dealing with lexical resources for performing NLP tasks. One motivation for this work is to determine whether these efforts for preserving the soundness of the resources are worth in all applications, and if not for which applications soudness should be guaranteed. Unfortunately, the lack of benchmark on these issues prevent us from answering these questions at this time. We can only emphasize the need of comparative evaluations on different tasks such as word sense disambiguation, bridging resolution, metaphor resolution or translation. Here we remain at the level of theoretical justifications while quantitative evaluations will be considered in future work.

As we saw, many differences between the approaches we presented can be explained in terms of granularity or what we called *constraint density*. Taxonomic axioms (coarse-grained) are easy to obtain, while meronomic (mid-grained), dependance and participation (fine-grained) axioms are much more difficult to get. This classification has to be related to the knowledge engineering distinction proposed in (Guarino, 1998a) in which coarse-grained ontologies are opposed to fine-grained one. The former corresponds to *reference*, *off-line* or *heavy-weight* ontologies and the later to *shareable*, *on-line* or *light-weight* ones. Performing some tasks can be more efficient or even provide better results with only taxonomic axioms while other tasks might require more fine-grained axioms. Finally, each developer should be able to choose the level of *constraint density* of the tool he develops according to the required level of precision.

A general coherent model, both for applications and psycho-linguistic adequacy, may consist of only one resource in which the constraints are classified according to their level of granularity.

7 Application to multilinguality

The previous section suggests that different applications need different characteristics from the resource. As an illustration we will consider now multilingual resources and, more particularly, how the different approaches presented above can be used in a multilingual context. The general purpose of this kind of work is to provide lexical resources in several languages aligned with a shared ontology like the Inter-Lingual-Index in EUROWORDNET (Vossen, 1998). Such resources are obviously highly valuable for most applications dealing with translation and multi-linguality. Linguistic and psycho-linguistic fundamental research is also concerned in such experiments. They might allow the practical comparison of lexical structure of different languages and thus constitute an important aspect in the research about the relativist/universalist debate (Gumperz and Levinson, 1996).

A first approach, taken for example in (Huang et al., 2004) consists in starting from an existing lexical resource (in this case the original English WordNet) aligned with an ontology (here SUMO) and a translation database (here the English-Chinese Translation Equivalents Database). The general idea is to translate the original resource into the new language. In order to do so, a bottom-up approach has been used. First Chinese lemmas were linked to English WORDNET synsets. Then the English WORDNET semantic relations were automatically inserted into the Chinese resource. The next crucial step consisted in the manual checking of the result. The correspondence is made through the terminology and the taxonomic axioms. This approach is fast and does not require a pre-existing lexical resource for each language. The main drawback seems to be the level of precision of the alignment. The accuracy of the new resource will be at most the one of the starting resource.

Moreover it might be the case, specially in lower levels of the lexical resource, that lexical structures do no fit. This methodology commits to lexical organization derived from a given language and tries to fit another language into its actual structure. In (Huang et al., 2003), an accurate manual check might reduce this tendency but the methodology to fix the problems in the alignment is not fully discussed and it is costly.

A second approach consists in developing parallel alignment with the same methodology for both languages. Of course, it requires to have a lexical resource for each language. In this case the correspondence between the languages can be made not only through the terms and taxonomic axioms but also through more fine-grained axioms potentially derived from original glosses. This approach is obviously less efficient than the previous one. Moreover the need to have computational lexicons for both languages prevents its use for most of natural languages.

These "foundational" multilingual lexical resources are still to come and won't replace the fast developing multilingual resources since these latter are essential for languages that did not benefit of extended and systematic efforts comparable to those on English language.

However, the careful integration of two rich and comparable (because axiomatized in a same logical language) lexical resources is expected to provide a more detailed and accurate tool than those currently developed. Also, this method allows a cross-linguistic investigation on languages and might constitute a significant step toward understanding lexical organization.

Overall the creation of authentic lexical resource tailored for each language and built according to language specificities and their careful integration seems complementary to the current efficient and fast-developing multilingual resource proposed in (Huang et al., 2004).

8 Conclusion

We proposed a way to classify the work done in interfacing ontologies and lexical resources. It consists in a clear separation between the restructuration of a lexical resource on the ground of an existing ontology hosting ontological principles, and the process of populating an ontology with lexical resources terms. A third option, called *alignment*, is a combination of these two aspects for the benefit of both the lexical resource and the ontology. We have shown how actual on-going work fits this classification through some examples. Finally, in the light of these clarifications, we discussed the issue of *con-*

straint density for lexical resources and related it to the light-weight/heavy-weight distinction established in kwnowledge representation. In this paper, we showed that different construction methodologies leads to different features in the resulting resources. Finally we emphasized the need for selecting top-level ontologies and lexical resources according to their reliability.

As explained in the paper, future work concerns in particular the practical evaluation of the resources developed with the different methods that have been presented. This evaluation has to be done task by task in order to understand better which task requires which features. Such an evaluation constitute a crucial step for the integration of ontological enhancement for lexical resources.

Acknowledgments

The authors would like to thank the ONTOLEX reviewers for their comments and suggestions.

References

J. F. Allen. 1984. Towards a general theory of action and time. *Artificial intelligence*, 23:123–154.

S. Borgo, N. Guarino, and C. Masolo. 1996. A pointless theory of space based on strong connection and congruence. In M. Kaufmann, editor, *Proceedings of Knowledge Representation and Reasoning (KR96)*.

C. Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

C. Fillmore. 1976. Frame semantics and the nature of language. In *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech*.

A. Gangemi, N. Guarino, and A. Oltramari. 2001. Conceptual analysis of lexical taxonomies: The case of wordnet top-level. In C. Welty and B. Smith, editors, *2nd International Conference on Formal Ontology in Information Systems*, pages 285–296.

A. Gangemi, N. Guarino, C. Masolo, and A. Oltramari. 2003a. Sweetening WordNet with DOLCE. *AI Magazine*, 3(24):13–24.

A. Gangemi, R. Navigli, and P. Velardi. 2003b. The ontowordnet project: extension and axiomatisation of conceptual relations in wordnet. In *International Conference on Ontologies, Databases and Applications of SEmantics (ODBASE)*, Catania, (Italy).

A. Gómez-Pérez, M. Fernández-López, and O. Corcho. 2004. *Ontological Engineering*. Springer.

N. Guarino and C. Welty. 2000. A formal ontology of properties. In *in 12th International Conference on Knowledge Engineering and Knowledge Management: Methods, Models and Tools*.

N. Guarino and C. Welty. 2004. An overview of ontoclean. In S. Staab and R. Studer, editors, *Handbook of ontologies*, chapter An overview of ontoclean, pages 151–159. Springer Verlag.

N. Guarino. 1998a. Formal ontology in information systems. In IOS Press, editor, *Proceedings of FOIS'98*, pages 3–15, Trento, Italy.

N. Guarino. 1998b. Some ontological principles for designing upper level lexical resources. In A. Rubio, N. Gallardo, R. Castro, and A. Tejada, editors, *Proceedings of First International Conference on Language Resources and Evaluation*, pages 527–534. ELRA.

J. J. Gumperz and S. C. Levinson. 1996. *Rethinking Linguistic Relativity*. Studies in the social and Cultural Foundations of Language. Cambridge University Press.

S. M. Harabagiu and D. I. Moldovan. 1998. Knowledge processing on an extended wordnet. In Christiane Fellbaum, editor, *WordNet, An electronic lexical Database*, chapter 16, pages 379–406. The MIT Press.

S. M. Harabagiu, G. A. Miller, and D. I. Moldovan. 1999. Wordnet 2 - a morphologically and semantically enhanced resource. In *SIGLEX 1999*.

C-R. Huang, E. I. J. Tseng, D. B. S. Tsai, and B. Murphy. 2003. Cross-lingual portability of semantic relations: bootstrapping chinese wordnet with english wordnet relations. *Language and Linguistics*, 4(3):509–532.

C.-R. Huang, R.-Y. Chang, and S.-. Lee. 2004. Sinica BOW (bilingual ontological wordnet): Integration of bilingual WordNet and SUMO. In *4th International Conference on Language Resources and Evaluation (LREC2004)*, Lisbonne.

C. Masolo, S. Borgo, A. Gangemi, N. Guarino, and A. Oltramari. 2003. Wonderweb deliverabled18, ontology library (final). Technical report, LOA-ISTC, CNR.

I. Niles and A. Pease. 2001. Towards a standard upper ontology. In Chris Welty and Barry Smith, editors, *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001)*, Ogunquit, Maine.

I. Niles and A. Pease. 2003. 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.

A. Oltramari, A. Gangemi, N. Guarino, and C. Masolo. 2002. Restructuring wordnet's top-level: The ontoclean approach. In K. Simov, editor, *Workshop Proceedings of OntoLex'2, Ontologies and Lexical Knowledge Bases, LREC2002*, pages 17–26, Las Palmas, Spain.

S. Reed and D. Lenat. 2002. Mapping ontologies into cyc. In *AAAI 2002 Conference Workshop on Ontologies For The Semantic Web*, Edmonton, Canada.

B. Smith. 1996. Mereotopology: A theory of parts and boundaries. *Data and Knowledge Engineering*, 20:287–303.

J. Sowa. 1995. *Knowledge Representation*. Brooks/Cole, Pacific Grove, CA.

M. Taboada, D. Martinez, and J. Mira. 2005. Experiences in reusing knowledge resources using Protégé and PROMPT. *International Journal of Human-Computer Studies*, 62:597–618.

P. Vossen, editor. 1998. *EuroWordNet: a multilingual database with lexical semantic networks*. Kluwer Academic Publishers.