Exploiting Linguistic Resources for building linguistically motivated ontologies in the Semantic Web

Maria Teresa Pazienza, Armando Stellato

AI Research Group, DISP, University of Rome, Tor Vergata
Via del Politecnico 1 00133 ROMA (ITALY)
{pazienza,stellato}@info.uniroma2.it

Abstract

Ontologies provide formal models for representing domain knowledge, which reveal to be useful in several contexts where efficient organization of available data and an shared understanding of its content reveals to be crucial. The Semantic Web offers the most appropriate scenario for exploiting ontologies' potentialities, due to the large amount of information which is to be exposed and accessed. The Semantic Web is however not a controllable and easy to manage knowledge base, and is instead characterized by huge quantities of documents accessed by thousands of users. Though machine readability is a primary demand for automatic exchange of data, several SW services (Intelligent Q&A, Semantic Search Engines etc..) still need to access knowledge expressed in the primary way humans can easily understand it: natural language. Moreover, the role of different cultures and languages is fundamental in a real World aWare Web, so that multilingualism becomes of great interest in this boiling cultural cauldron. These premises suggest that ontologies as we know them now, should be enriched to cover formally expressed conceptual knowledge as well as to expose its content in a linguistically motivated fashion. This paper presents our approach in establishing a framework for semi-automatic linguistic enrichment of ontologies, which led to the development of OntoLing, a plug-in for the popular ontology development tool Protégé. We describe here its features and design aspects which characterize its current release.

1. Introduction

The scenario offered by the SW (and by the Web in general) is however characterized by huge quantities of documents and by users willing to access them. Though machine readability is a primary aim for allowing automatic exchange of data, several SW services like Intelligent Q&A, Semantic Search Engines etc.. still need to understand and expose knowledge expressed in the sole way humans can easily understand it: natural language. Moreover, the role of different cultures and languages is fundamental in a real World aWare Web and, though English is recognized de facto as a "lingua franca" all over the world, much effort must be spent to preserve other idioms expressing different cultures. As a consequence, multilinguality has been cited as one of the six challenges for the Semantic Web (Benjamins et al., 2004). These premises suggest that ontologies as we know them now, should be enriched to cover formally expressed conceptual knowledge as well as to expose its content in a linguistically motivated fashion.

In this paper we introduce our work in establishing a framework for semi-automatic linguistic enrichment of ontologies, which has run through the identification of different categories of linguistic resources and planning their exploitation to augment the linguistic expressivity of ontologies. This effort has lead to the development of OntoLing, a plug-in for the popular ontology editing tool Protégé (Gennari et al., 2003) which allows for linguistic enrichment of ontologies. We describe here the features characterizing its current release and discuss some of the innovations we are planning for the near future. In particular, Section 2 describes the motivations for a linguistically-aware approach to ontology development, and lists the main objectives which guided the development of OntoLing. Section 3 provides some background on linguistic resources, their availability and how they are characterized. Section 4 describes a general interface for accessing the content of these resources, introducing the concept of Linguistic Watermark. In Section 5 we describe the architecture of OntoLing, its functionalities and its adaptive behavior towards different lexical resources. Section 6 describes how linguistic enrichment has been modeled in Protégé and Protégé OWL. Section 7 concludes this document with considerations on the work done so far, adding some hints on future research directions.

2. Ontologies meet language

Ontology Development is a task requiring considerable human involvement and effort, at a large extent with the objective of providing a shareable perspective over domain related knowledge. What "shareable" means, depends on the nature of the task(s) the ontology is thought for. The scenario offered by the Semantic Web is in fact characterized by distributed services which must both realize and rely on a proper connection of machine-accessible formal semantics and more traditional Web content.

For this connection to be true, a complete Ontology Development process should consider the formal aspects of conceptual knowledge representation, as well as guarantee that the same knowledge be recognizable amongst its multiple expressions which are available on real data: that means language.

To achieve such a deeper expressivity, we should reconsider the process of Ontology Development to include the enrichment of semantic content with proper lexical expressions in natural language. Ontology Development tools should reflect this need, supporting users with dedicated interfaces for browsing linguistic resources: these are to be integrated with classic views over knowledge data such as class trees, slot and instance lists, offering a set of functionalities for linguistically enriching concepts and, possibly, for building new ontological knowledge starting from linguistic one.

By considering some of our past experiences (Atzeni et al, 2004, Pazienza et al. 2003, 2005) with knowledge

based applications dealing with concepts and their lexicalizations, a few basic functionalities for browsing linguistic resources (from now on, LRs) emerged to be mandatory:

- Search term definitions (glosses)
- Ask for synonyms
- Separate different sense of the same term
- Explore genus and differentia
- Explore resource-specific semantic relations as well as some others for ontology editing:
- Add synonyms (or translations, for bilingual resources) as additional labels for identifying concepts
- Add glosses to concepts description (documentation)
- Use notions from linguistic resources to create new concepts

While ontologies have undergone a process of standardization which culminated, in 2004, with the promotion of OWL (Dean et al, 2002) as the official ontology language for the semantic web, linguistic resources still maintain heterogeneous formats and follow different models, which make tricky the development of such an interface. The next sections address this problem and discuss our approach in defining the model of OntoLing, the Plug-in for Protégé dedicated to linguistic enrichment of ontologies.

3. Linguistic Resources, an overview

"The term linguistic resources refers to (usually large) sets of language data and descriptions in machine readable form, to be used in building, improving, or evaluating natural language (NL) and speech algorithms or systems" (Cole et al, 1997). Examples of linguistic resources are written and spoken corpora, lexical databases, grammars, treebanks and field notes. In particular, this definition includes lexical databases, bilingual dictionaries and terminologies (which can all be indicated as lexical resources), which may reveal to be necessary in the context of a more linguistic-aware approach to KR. In past years several lexical resources were developed and made accessible (a few for free), and a wide range of resources is now available, ranging from simple word lists to complex MRDs and thesauruses. These resources largely differentiate upon the explicit linguistic information they expose, which may vary in format, content granularity and motivation (linguistic theories, task or system-oriented scope etc…).

Multiple efforts have been spent in the past towards the achievement of a consensus among different theoretical perspectives and systems design approaches. The Text Encoding Initiative [OR5] and the LRE-EAGLES (Expert Advisory Group on Linguistic Engineering Standards) project (Calzolari et al., 1996) are just a few, bearing the objective of making possible the reuse of existing (partial) linguistic resources, promoting the development of new linguistic resources for those languages and domains where they are still not available, and creating cooperative infrastructure to collect, maintain, and disseminate linguistic resources on behalf of the research and development community.

However, at present time, with lack of a standard on existing LRs, it appears evident that desiderata for functionalities which we described in section 2, would depend upon the way these resources had been organized. Often, even a local agreement on the model adopted to describe a given (a series of) resource does not prevent from an incorrect formulation of its content. This is due to the fact that many resources have been initially conceived for humans and not for machines, As an example, on existing available dictionaries words' definitions and synonyms are not always managed the same way: in some cases synonyms are clustered upon the senses which are related to the particular term being examined (among others, Babylon [OR1] and Dict [OR2] dictionaries, where the senses are separated by a ";" symbol), other simply report flat lists of terms without even identifying their different meanings (as for Freelang dictionaries [OR3]). In several dictionaries, synonyms are mixed with extended definitions (glosses) in a unpredictable way and it is not possible to automatically distinguish them. Terms reported as synonyms may sometimes not be truly synonyms of the selected term, but may represent more specific or general concepts (this is the case of Microsoft Word synonymy prompter). Of course, the ones mentioned above represent mere dictionaries not adhering to any particular linguistic model, though they may represent valuable resources on their own.

A much stronger model is offered by Wordnet (Fellbaum, 1998), which, being a structured lexical database, presents a neat distinction between words, senses and glosses, and is characterized by diverse semantic relations like hypernymy/hyponymy, antonymy etc… Though not being originally realized for computational uses, and being built upon a model for the mental lexicon, WordNet has become a valuable resource in the human language technology and artificial intelligence. Due to its vast coverage of English words, WordNet provides with general lexico-semantic information on which open-domain text processing is based. Furthermore, the development of WordNets in several other languages (Vossen, 1998) extends this capability to trans-lingual applications, enabling text mining across languages.

It is impossible to foresee all the features which could be exposed by different resources, from simple word lists to complex multilingual Wordnets: a trade-off must be found, to outline the shape of an interface with sufficient level of generality to be exploited automatically, while leaving space for introducing custom functionalities, to be considered as resource specific services and thus exploited upon discovery.

4. A General Interface for Lexical Resources: The Linguistic Watermark

Along with the analysis of a general interface for linguistic resources, it emerged the logical independence which it could maintain with respect to its possible embedding applications. Our experience pointed out usefulness in diverse natural language related applications like Ontology Mapping, Question&Answering and Information Extraction, where support for multilinguality and a wider linguistic awareness could be, if not necessary, at least useful for improving performances. Moreover, the interface could also act as a sort of unique fingerprint for describing the underlying resource for which access is provided, its information being exploitable in many application-dependant contexts.

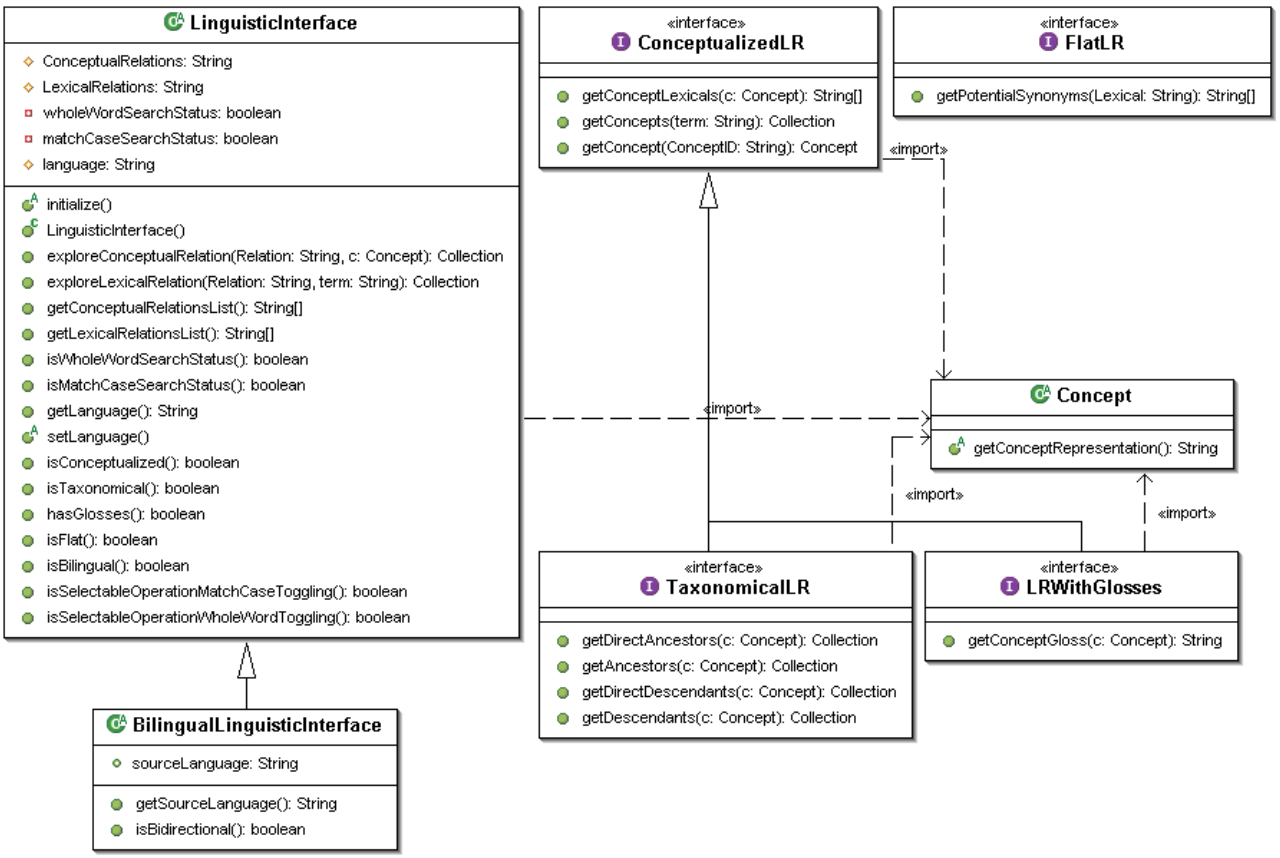

Figure 1: A Class diagram depicting part of Linguistic Watermark classes and interfaces

For this reason, we introduced the notion of Linguistic Watermark, as the series of characteristics and functionalities which distinguish a particular resource inside our framework. As we can observe from the Class Diagram in Fig. 1, we sketched a sort of ontology of linguistic resources, with the addition of operational aspects. Linguistic resources are in fact structured and described in terms of their features and how their lexical information is organized; the ontology has then been completed with query methods for accessing resource's content. We thus implemented this operational ontology as a java package on its own, which can externally be imported by any application willing to exploit natural language resources like lexicons and terminologies. The core of the package is composed of an Abstract Class, named `LinguisticInterface`, which is both the locus for a formal description of a given linguistic resource and a service-provider for exposing the resource specific methods. The other abstract classes and interfaces in the package, which can be implemented or not, depending on the profile of the resource being wrapped, provide instead the signatures for known interface methods.

We have currently developed several implementations of the Linguistic Watermark. Two of them, the WordNet Interface and the last DICT Interface, being related to freely available resources, have been made publicly available on the OntoLing site.

The first one is an almost totally complete implementation of the Linguistic Watermark. The WordNet Interface is in fact a `ConceptualizedLR`, because its linguistic expressions are clustered upon the different senses related to the each term. These senses – "synsets", in WordNet terminology – have been implemented through the Concept interface, which we see bounded by the import statement in the class diagram. WordNet is a `LRWithGlosses`, as glosses are neatly separated from synonyms and organized in a one-to-one relation with synsets. Finally, WordNet Interface implements `TaxonomicalLR`, as its indexed word senses are organized in a taxonomy of more specific/more generic objects.

The other one, DICT Interface, is based on the Dictionary Server Protocol (DICT) [OR2], a TCP transaction based query/response protocol that allows a client to access dictionary definitions from a set of natural language dictionary databases. The DICT interface is conceptualized too, though its word senses are not indexed as in WordNet (that is, it is not possible to correlate senses of two different terms upon the same meaning). DICT Interface is also a `BilingualLinguisticInterface`, as its available word-lists provide translations for several idioms.

Other available interface classes denote Flat resources (as opposed to Conceptualized ones), which contain flat lists of linguistic expressions for each defined term, and `BidirectionalTranslators`, which represent a further specialization of Bilingual Linguistic Interfaces providing bidirectional translation services. Other interfaces (`ApproximateSearchToggling`) are not directly related to the characteristics of the wrapped LR, but to search functionalities which have been provided for it.

As previously mentioned, we defined two classes of methods for browsing LRs: those defined in advance in the interfaces, which can thus be exploited inside automatic processes, and other very specific resource-

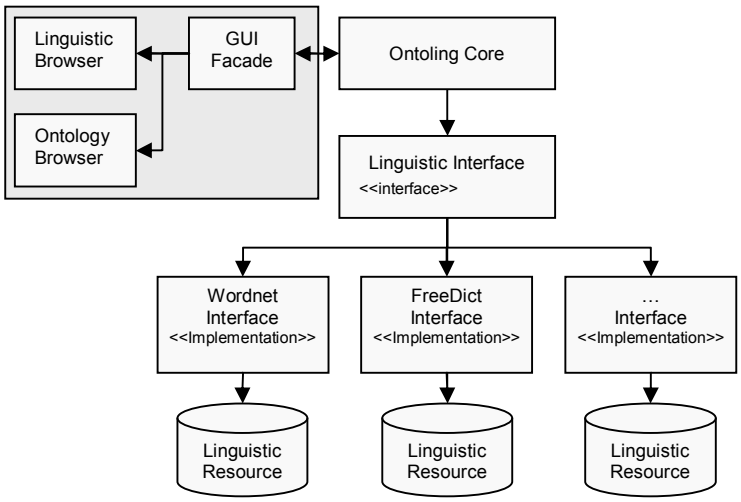

Figure 2: OntoLing Architecture

dependent methods, which are loaded at run-time when the LR is interfaced to some browsing application (e.g. OntoLing). Two methods available in LinguisticInterface: `getLexicalRelationList` and `getConceptualRelationList` act thus as service publishers, the former providing different methods for exploring lexical relations among terms or relating terms to concepts, the latter reporting semantic relations among concepts. Through these methods, the WordNet Interface makes available to the user all the semantic relations contained in WordNet.

5. OntoLing Architecture

The architecture of the Ontoling plugin (see Fig. 2) is based on three main components:

1. the GUI, characterized by the Linguistic Resource browser and the Ontology Enrichment panel
2. the external library Linguistic Watermark, which has been presented in the previous section, providing a model for describing linguistic resources
3. the core system

and an additional external component for accessing a given linguistic resource. This component, which can be loaded at runtime, must implement the classes and interfaces contained in the Linguistic Watermark library, according to the characteristics of the resource which is to be plugged. In the following sections we provide details on the above components.

5.1. OntoLing Core Application

The core component of the architecture is responsible for interpreting the Watermark of linguistic resources and for exposing those functionalities which suit to their profile. Moreover, the behavior of the whole application is dependant on the nature of the loaded resource and is thus defined at run-time. Several methods for querying LRs and for exposing results have been encapsulated into objects inside a dedicated library of behaviors: when a given LR is loaded, the core module parses its Linguistic Watermark and assigns specific method-objects to each GUI event.

With such an approach, the user is provided with a uniform view over diverse and heterogeneous linguistic resources, as they are described in the Linguistic Watermark ontology, and easily learns how to interact with them (thus familiarizing with their peculiarities) by following a policy which is managed by the system.

For example, with a flat resource, a search on a given term will immediately result in a list of (potential) synonyms inside a dedicated box in the GUI; instead, with a conceptualized resource, a list of word senses will appear in a results table at first, then it will be browsed to access synonymical expressions related to the selected sense. Analogous adaptive approaches have been followed for many other aspects of the Linguistic Watermark (mono or bidirectional Bilingual Translators, presence of glosses, Taxonomical structures and so on…) sometimes exploding with combinatorial growth.

Future development of Ontoling will go in the direction of considering supervised techniques for automatic ontology enrichment; selecting and modeling the right strategies for the adopted LRs is another task the core module is in charge for.

5.2. OntoLing User Interface

Once activated, the plug-in displays two main panels, the Linguistic Browser on the left side, and the Ontology Panel on the right side (see Fig. 3).

The Linguistic Browser is responsible for letting the user explore the loaded linguistic resource. Fields and tables for searching the LR and for viewing the results, according to the modalities decided by the core component, are made available. The menu boxes on the left of the Linguistic Browser are filled at run time with the methods for exploring LR specific Lexical and Conceptual relations.

The Ontology Panel, on the right, offers a perspective over ontological data in the classic Protégé style. By right-clicking on a frame (class, slot or instance), the typical editing menu appears, with some further options provided by OntoLing to:

1. search the LR by using the frame name as a key
2. change then name of the selected frame to a term selected from the Linguistic Browser
3. add terms selected from the Linguistic Browser as additional labels for the selected frame
4. add glosses as a description for the selected frame
5. add IDs of senses selected from the linguistic browser as additional labels for the frames

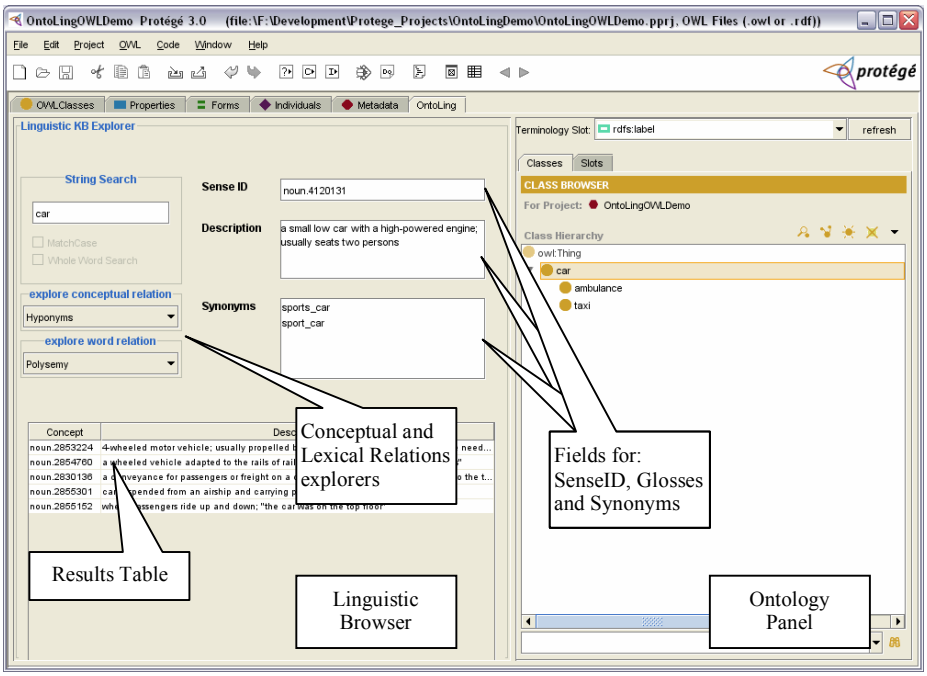

Figure 3: A screenshot of the OntoLing Plug-in

6. create a new frame with a term selected from the Linguistic Browser as frame name (identifier)
7. only in class and slot browser: if the LR is a `TaxonomicalLR`, explore hyponyms (up to a chosen level) of the concept selected on the Linguistic Browser and reproduce the tree on the frame browser, starting from the selected frame, if available

These functionalities allow not only for linguistic enrichment of ontologies, but can be helpful for Ontologists and Knowledge Engineers in creating new ontologies or in improving/modifying existing ones.

How terms and glosses are added to the description of ontologies concepts, depends on the ontology model which is being adopted and is explained in detail in the following section.

6. Using OntoLing with Protégé and Protégé OWL

When a frame-based approach was first adopted in Protégé as a knowledge model for representing ontologies and knowledge bases, no explicit effort was dedicated to the representation of possible alternate labels (synonyms) for concepts neither to support the idea of multilingualism in Ontologies. Frame names were almost as equivalent as IDs, and people were only encouraged, as it is common practice in computer programming when addressing variable names, to adopt "meaningful and expressive names" to denote these IDs. The Protégé model was indeed quite strong and expressive, so that every ontology developer could deal with his linguistic needs at a meta-ontological level and find the right place for them, though no official agreement was yet established.

Later on, with the advent of OWL as a KR standard for the Semantic Web, and with the official release of the Protégé OWL plug-in (Knublauch et al., 2004), things started to converge towards a minimal agreement for the use of language inside ontologies. When we first started working on OntoLing, the OWL plug-in had just been released, and the majority of users continued to use Protégé in the usual way, so we had to find a solution that was quite easy (for the user) to make do with this lack in the standard Protégé model.

To this end, we defined the notion of terminological slot, as a slot which is elected by the user to contain different linguistic expressions for concepts. Any string-typed slot with cardinality set to multiple, can potentially be selected as a terminological slot, and, for easiness of use, OntoLing prompts the user only with this class of slots. This way, to use Ontoling with standard Protégé, a user only needs to define a proper metaclass and metaslot, containing the elected terminological slot; naturally, the same slot can be dedicated to instances at class level. Multilingual ontologies can also be supported by creating different slots and selecting each of them as terminological slots during separate sessions of Linguistic Enrichment, with diverse LRs dedicated to the different chosen languages. Concerning glosses, these can be added to the common "documentation" slot which is part of every frame by default.

Conversely, Linguistic Enrichment of OWL Ontologies follows a more predictable path, thanks to OWL's language dedicated Annotation Properties, such as *rdfs:label* and *owl:comment*. When Ontoling recognizes a loaded ontology as expressed in the OWL language, the terminological slot is set by default (though modifiable) to *rdfs:label*. In this case the *xml:lang* attribute of the label property is automatically filled with the language declared by the Linguistic Interface.

7. Conclusions and future work

As it has been widely described and discussed in the literature on Ontology Development (Noy & McGuinness, 2001, Fernandez et al, 1997), the role of language must not be underestimated. In this work we contributed to the linguistic aspects of ontology development, by identifying functionalities for augmenting the linguistic expressivity of existing ontologies and by implementing these functionalities in the OntoLing Protégé plug-in.

OntoLing, with WordNet as its first exploitable resource, has been adopted by a community of users coming from diverse research areas, from pure linguists approaching ontologies, to ontology developers exploiting specific parts of WorldNet's taxonomical structure as a basis for creating their own domain ontology, up to users needing its main functionalities to enrich ontological concepts of existing ontologies with greater linguistic emphasis. With the recent release of the DICT Interface we added a little step in assisting multilingual ontology development and we now look forward other freely available resources to be added to Ontoling plug-in library: two extensions for MultiWordNet (Pianta et al., 2002) and EuroWordNet (Vossen, 1998) are being developed and will be released in the next months. Moreover, we are currently examining the possibility of extending the interface beyond traditional lexical resources, embracing other type of linguistic resources, such as FrameNet (Baker et al., 1998) and VerbNet (Kipper et al., 2000).

Another explored research direction (see Pazienza & Stellato, 2006) is related to automatization of the process, in order to reduce human effort to a fully supervised methodology for linguistic enrichment of ontologies. We are improving our conceived techniques and testing their quality against real available ontological data, as the results of this specific research will contribute to extend the possibilities offered by the whole framework.

Finally, an important aspect we will address in the future is to better express the relations between ontology and language. Adherence to nowadays standards for ontology representation has been in fact a limit for our research on linguistic enrichment, where a more structured and close bridging between conceptual and linguistic knowledge, with respect to the one we have provided, would be expected. The link we establish in this work between conceptual knowledge and its associated linguistic representation is characterized by simple references between concepts and labels (as offered by the standard *owl:comment* and *rdfs:label* properties), while more sophisticated relationships between lexical entries and ontological objects are required to address the complex conceptualizations which characterize a significant fraction of every ontology.

8. Online Resources

[OR1] Babylon: www.babylon.com
[OR2] DICT: www.dict.org/bin/Dict
[OR3] Freelang: www.freelang.com
[OR4] WordNet: http://www.cogsci.princeton.edu/~wn/
[OR5] Text Encoding Iniziative: www.tei-c.org

9. References

Atzeni, P., Basili, R., Hansen, D. H., Missier, P., Paggio, P., Pazienza, M. T. and Zanzotto, F. M.: Ontology-based question answering in a federation of university sites: the MOSES case study. *9th International Conference on Applications of Natural Language to Information Systems (NLDB'04)* Manchester (United Kingdom), June 2004

Baker, C.F., Fillmore, C.J and Lowe., J.B.: The Berkeley FrameNet project. *In Proceedings of the COLING-ACL*, Montreal, Canada, 1998

Benjamins V. R., Contreras, J., Corcho O. and Gómez-Pérez, A. *Six Challenges for the Semantic Web*. SIGSEMIS Bulletin, April 2004

Calzolari, N., McNaught, J. and Zampolli, A.: *EAGLES Final Report: EAGLES Editors Introduction*. EAG-EB-EI, Pisa, Italy 1996

Cole, R. A., Mariani, J., Uszkoreit, H., Zaenen, A. and Zue, V. Eds. *Survey of the State of the Art in Human Language Technology*, Cambridge University Press, Cambridge, UK, 1997

Dean M. and Schreiber, G., Editors. *OWL Web Ontology Language Reference*. W3C Recommendation, 10 February 2004, http://www.w3.org/TR/2004/REC-owl-ref-20040210/. Latest version available at http://www.w3.org/TR/owl-ref/

Fellbaum, C.: *WordNet - An electronic lexical database*. MIT Press, (1998).

Fernandez, M., Gómez-Pérez, A. and Juristo, N. (1997) METHONTOLOGY: From Ontological Art Towards Ontological Engineering, *AAAI-97 Spring Symposium on Ontological Engineering*, Stanford University, March 24-26th

Gennari, J., Musen, M., Fergerson, R., Grosso, W., Crubézy, M., Eriksson, H., Noy, N., and Tu, S.: The evolution of Protégé-2000: An environment for knowledge-based systems development. *International Journal of Human-Computer Studies*, 58(1):89–123, 2003.

Kipper, K., Trang Dang, H. and Palmer, M.: Class-Based Construction of a Verb Lexicon. *AAAI-2000 Seventeenth National Conference on Artificial Intelligence*, Austin, TX, July 30 - August 3, 2000

Knublauch, H., Fergerson, R. W., Noy, N. F. and Musen M.A.: The Protégé OWL Plugin: An Open Development Environment for Semantic Web Applications. *Third International Semantic Web Conference - ISWC 2004*, Hiroshima, Japan. 2004

Noy, N. F., McGuinness, D. L: *Ontology Development 101: A Guide to Creating Your First Ontology*. Stanford Knowledge Systems Laboratory Technical Report KSL-01-05. March 2001.

Pazienza, M.T. and Stellato, A.: Linguistic Enrichment of Ontologies: a methodological framework. *Second Workshop on Interfacing Ontologies and Lexical Resources for Semantic Web Technologies (OntoLex2006)*, held jointly with LREC2006, Magazzini del Cotone Conference Center, Genoa, Italy, 24-26 May 2006

Pazienza, M.T., Stellato, A., Vindigni, M., Valarakos, A. and Karkaletsis, V.: Ontology integration in a multilingual e-retail system. *HCI International 2003*, Crete, Greece, 2003

Pazienza, M. T., Stellato, A., Henriksen, L., Paggio, P., Zanzotto, F. M.: Ontology Mapping to support ontology-based question answering. *Proceedings of the second MEANING workshop*. Trento, Italy, February 2005

Pianta, E., Bentivogli L., & Girardi, C.: MultiWordNet: Developing an aligned multilingual database. *In Proceedings of the 1st International Global WordNet Conference* (pp. 293--302), Mysore, India, January 21-25, 2002

Vossen, P.: EuroWordNet: *A Multilingual Database with Lexical Semantic Networks*, Kluwer Academic Publishers, Dordrecht, 1998