This is the final version of the following article:

Buendía, Miriam, and José Manuel Ureña. 2010. Towards a Methodology for Semantic Annotation: The Case of Meteorology. In *Traducción y modernidad. Textos científicos, jurídicos, económicos y audiovisuales*, ed. R. López-Campos, C. Balbuena, and M. Álvarez, 27-36. Córdoba: Universidad de Córdoba.

You can find more articles authored by LexiCon Research Group members at <http://lexicon.ugr.es>.

Towards a Methodology for Semantic Annotation: The Case of Meteorology

Miriam Buendía Castro and José Manuel Ureña Gómez-Moreno

University of Granada (Spain)

Abstract

Internet has brought with it a new way of organizing and obtaining information. It provides the possibility of accessing any type of information at any time and at any place. Although search engines such as Google have emerged to deal with this *information overload*, searches are often frustrating because users are unable to find what they are looking for.

The main reason for this lack of satisfaction is the fact that the *World Wide Web* generally uses HTML for the codification of documents. HTML allows the annotation of documents, but unfortunately, this particular kind of annotation only deals with the visual presentation of texts, and has no semantic links. In order to overcome this difficulty and improve the results obtained in searches, it has become necessary to create semantic machine-readable contents that computers can understand and process. This is the reason why the *Semantic Web* was conceived. The idea is to help computers to understand contents, which humans can easily process. In this respect, *semantic annotation* has become a fundamental part of the development of the Semantic Web.

This paper describes a methodology for the semantic annotation of specialised texts and web documents. This approach applied to the domain of Meteorology, through the use of a subcorpus of texts extracted from the EcoLexicon corpus of the Ecosystem research project[i]. The tagging system proposed is based on an incipient ontology currently being built within EcoLexicon. With the software application SMORE[ii] we

have imported the ontology and generated semantic annotations, thus linking different sections of the texts (words, phrasemes, etc.) to the concepts of the ontology.

This preliminary linguistic analysis shows that these *semi-automatic* annotation systems help us to semantically annotate texts through the use of an ontology, but do not allow us to see the results of the annotation, and benefit from them. Consequently, we are now working on the implementation of a query agent that will exploit the semantic annotations, map the queries to the ontology in order to identify only those projects related to the Semantic Web area, and then offer semantic information.

Keywords: semantic annotation, corpus linguistics, ontology

1. Introduction

The semantic web is an extension of the WWW in that it tags information with a well-defined meaning. Its main objective is to formally specify the meaning of data contents on the net, and to obtain a network of contents which a computer can understand and process. The idea is to help computers to understand contents, which humans can easily process. This can be achieved through semantic tagging.

Ontologies have been proposed as a knowledge representation model capable of formally describing web resources and their vocabulary. An ontology can be used to make explicit the underlying meaning of concepts included in web pages. Ontological Semantics (Niremburg & Raskin, 2001) is a theory that studies the meaning and natural language processing. It uses an abstract model of the world, or ontology, as the main resource for extracting and representing text meaning (Aguado de Cea et al., 2002).

This study is part of the EcoSystem project, whose main objective is the conceptual representation of the specialized domain of Environmental Engineering in the form of a visual thesaurus of specialized concepts, organized in a constellation of interrelated dynamic knowledge frames (http://manila.ugr.es/visual/). This frame structure is based on information extracted from a corpus, which was the fundamental resource used to obtain naturally occurring data with a view to specifying the conceptual structure of the domain. The corpus data also provided the basis for the elaboration of terminological definitions, which represent each specialized concept.

This paper describes a system of semantic tagging created for a subcorpus of texts within the domain of Meteorology. This subcorpus was extracted from the corpus designed for the Ecosystem research project. This small group of texts was used as

bootstrapping for the subsequent implementation of semi-automatic tagging tools based on ontologies.

Due to the lack of standardization and tools for automatic semantic annotation[iii], we first analyzed the corpus and extracted the conceptual relations of TROPICAL CYCLONE, which was useful for the subsequent implementation of the ontology that we are building. From the ontology we extracted the entity classes to be codified in the text, and assigned each of them a unique identifier. Afterwards, we semantically enriched the corpus with the software SMORE.

This preliminary analysis showed how these entities were hierarchically related to each other. Semantic annotation facilitates the search, recovery and extraction of information, and is valuable in machine translation, the learning of ontologies, and web mining.

2. Why annotate a corpus?

Plain corpora have certain limitations in comparison to annotated corpora. For example, when studying a linguistic phenomenon in a plain text, searches can only be carried out by exact matching since it is impossible search by grammatical category.

As Leech (1997: 4) points out: "The fact is that to extract information from a corpus, we often have to begin by building information in". And that is exactly what annotation involves, enriching corpus texts with information in order to make explicit, implicit characteristics of textual elements.

2.1. Annotation: definition, types and applications

Annotation can be defined as: "[…] the practice of adding interpretative linguistic information to a corpus" (Leech 2004). Initially, Leech (1997) considered the following types of annotation: orthographic, phonetic, prosodic, grammatical, syntactic, semantic, discoursal and pragmatic/stylistic. However, in a later publication (Leech 2004), Leech established the following categories: grammatical, phonetic, semantic, pragmatic, discourse, stylistic and lexical.

On the basis of Leech's proposals (2004, 1997), we considered six types of annotation:

- **Part of speech tagging (POS**) identifies the different parts of a sentence, such as nouns, verbs or articles.

- **Lexical annotation or lemmatization** consists of adding the lemma identification of each word of a text. In English, it might be considered redundant, but in languages such as Spanish or German, it can be very useful for information extraction.
- **Syntactic annotation** consists of adding syntactic information to a corpus through the incorporation in the text of syntactic structure indicators.
- **Semantic annotation** consists of adding information about the semantic categories of words.
- **Discourse annotation** entails the addition of information about anaphoric links in a text.
- **Pragmatic annotation** consists of adding information regarding the kinds of speech act that occur in a spoken dialogue.

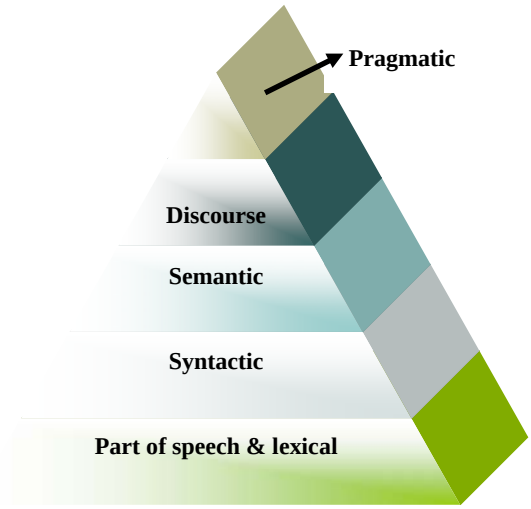

Fig. 1. Levels of linguistic annotation

Annotation has a wide range of applications: it makes extraction and recovery of information easier; it allows the re-usability of corpora; and it is multi-functional (Leech 2004, 1997).

2.1.1. Standards for corpus annotation

Leech (2004, 1997) suggests some *standards of good practice* that should be applied to any corpus annotation project:

1. Annotations should be separated from the text. It must always be easy to separate annotations from the raw corpus so that the raw corpus can be retrieved exactly in its original form.

2. Detailed and explicit documentation of the corpus should be provided. Burnard (2004) highlights the importance of providing detailed and explicit documentation about the corpus and its constituent texts. In the same way, it is important to provide such documentation about the annotation itself, i.e., the answers to questions such as *how*, *where*, *when* or *by whom* were the annotations carried out, as well as the type of annotation and coding scheme used. An annotation scheme is meant to be an explanatory system supplying information about the annotation practices followed. A coding scheme, on the other hand, refers to the set of symbolic conventions employed to represent the annotations themselves.
3. Annotation practices should be linguistically consensual and respect *de facto* standards. *De facto* standards reflect the standardization that has already begun to take place, and on which the research community tends to agree.

3. The Semantic Web

According to Berners-Lee, Hendler & Lassila (2001): "The Semantic Web is not a separate Web, but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation".

The main objective of the semantic web is to formally specify the meaning of data contents on the net, and to obtain a network of contents which a computer can understand and process. The idea is to help computers to understand contents, which humans can easily process. In this respect, semantic annotation has become a fundamental to the development of the semantic web.

3.1. Semantic annotation

Semantic annotation consists of assigning a semantic description to text entities. If a tag is added to each word in a text to indicate its semantic field, this helps to extract all the related terms of a semantic field in particular. Fig. 2 (Kiryakov et al. 2003: 484) shows an example of the process of semantic annotation:

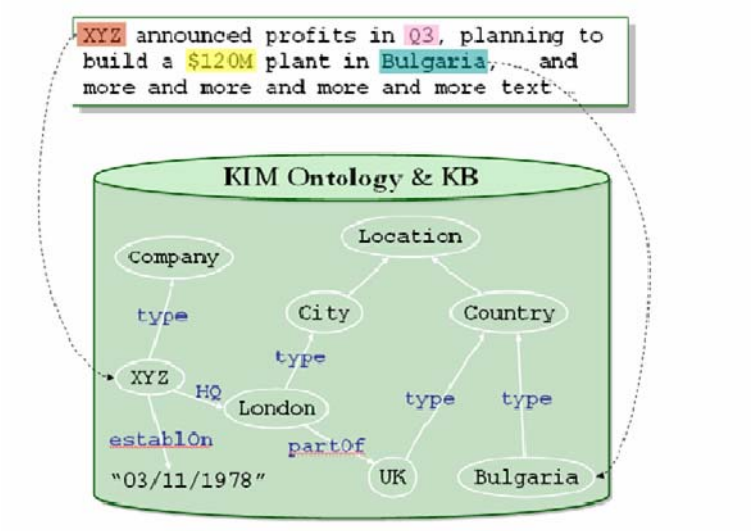

Fig. 2. Semantic annotation

This type of annotation is often referred to as *semantic annotation*, *entity annotation* or *semantic tagging*. However, there is no well established term for semantic annotation. Nor is there consensus on what this kind of annotation actually entails. Nevertheless, it is generally agreed that semantic annotation is the basis of a wide range of applications, such as the indexation, recovery and categorization of information or the generation of more advanced metadata.

Artificial Intelligence researchers have found in ontologies the ideal knowledge model to formally describe web resources and their vocabulary. Aguado de Cea et al. (2002: 38) opt for combining and identifying complementary features of semantic annotation models in Artificial Intelligence and annotations proposed by Corpus Linguistics.

3.2. Ontologies

In Computational Linguistics, the standard definition of *ontology* is "[…] a formal, explicit specification of a shared conceptualization" (Gruber 1993: 199).

Formal refers to the fact that an ontology should be machine-readable; *explicit* means that the type of concepts used and the restrictions on their use are explicitly defined. *Shared* reflects the notion that the ontology captures consensual knowledge. In other words, it is not the privilege of some individual, but accepted by a group. *Conceptualization* refers to an abstract model of some phenomenon in the world by having identified the relevant concepts of that phenomenon.

Weigand (1997) gives a more specific definition when he defines ontology as "a database describing the concepts in the World or some domain, some of their properties and how the concepts relate to each other".

From the previous definitions we can conclude that an ontology is composed of concepts, attributes, relations and restriction rules.

3.2.1. Web Ontology Language (OWL)

OWL (Web Ontology Language) is designed for applications that process the content of information instead of just presenting information to humans. OWL can be used to explicitly represent the meaning of terms in vocabularies and the relationships between those terms, through ontologies. OWL facilitates greater machine interpretability of Web content than that supported by XML, RDF, and RDF Schema[iv] by providing additional vocabulary along with a formal semantics.

OWL has three sublanguages: OWL Lite, OWL DL, and OWL Full, which incorporate different functionalities and which go, from a more simplified version — OWL-Lite—, used in the representation of simple hierarchies, to OWL-Full.

There are already a large number of OWL ontologies on the Web, such as an ontology of cancer developed by the National Cancer Institute (NCI), distributed as a component of the NCI Center for Bioinformatics of the USA[v] and the OWL version of the medical ontology GALEN developed by the University of Manchester[vi].

4. The methodology

Our methodology includes the following steps: (i) analysis of the corpus; (ii) extraction of conceptual relations; (iii) implementation of the ontology; (iv) semantic annotation of texts with the software SMORE.

4. 1. The corpus

The EcoLexicon corpus is a specialized corpus of texts that was compiled manually. It is composed of complete texts, belonging to the domains of Environmental Engineering and Coastal Management. It is a comparable corpus since it contains original texts in Spanish, English and German, selected according to the EAGLES criteria (1996a, 1996b, 1999). The texts in the corpus belong to the same domain, but are not translations of each other.

The English corpus contains about six million words. For our study, we extracted a mini corpus of approximately 1.000,000 words within the domain of Meteorology. This corpus was used to obtain naturally occurring data with a view to specifying the conceptual structure of the domain. We analysed the concordances[vii] of different terms related to Meteorology such as *wind, tropical cyclone*, etc. with the lexical analysis program, WordsmithTools[viii]. The corpus data also provided the basis for the elaboration of terminological definitions, which represent each specialized concept.

4.2. Extraction of the conceptual relations

The linguistic description of any concept should achieve the following: (1) make category membership explicit; (2) reflect its relations with other concepts within the same category; (3) specify its essential attributes and features (Faber et al., 2005). In a definition there are two major parts, the *genus* or nuclear part (which is indicative of the IS_A relationship) and the adverbial modification or *differentiae* that provides the characteristics that distinguish one concept from another within the same category (Faber et al., 2005).

Once the conceptual and terminological information was extracted, the next step was to abstract a schema or template which, according to the concordances, encodes the definitional structure for all of the concepts belonging to this domain. The definitional template for TROPICAL CYCLONE is the following:

Table 1. Definitional structure of TROPICAL CYCLONE

tropical cyclone	
IS_A	non frontal cyclone
HAS_CORE	warm
HAS_SIZE	synoptic scale
HAS_MOVEMENT	organized deep convection
HAS_ORIGIN	tropical or subtropical waters
HAS_INTENSITY	33m/s= 64kt= 74mph
HAS_TIME PERIOD	from 1 June to 30 November
CAUSES	-severe local storms -tornado (type of storm) -storm surge -high wind -heavy rainfall -floods -landslide

This template helps to extract all the conceptual relations activated in order to implement a preliminary ontology with all the concepts belonging to this domain.

4.3. The ontology

The domain ontology is implemented by using the ontology editor, *Protégé*[ix], after analysing the concepts and their relations. An ontology is composed principally of the following components: (i) classes and individuals; (ii) properties; (iii) rules.

Classes and individuals represent relevant concepts to the domain. For example, as can be seen in Fig. 3, *tropical cyclone*, *extratropical cyclone* and *subtropical cyclone* are classes of the ontology. Individuals are entities that belong to a particular class. Classes are hierarchically organized in such a way as to show that individuals of a subclass belong to the class.

Properties are the conceptual relations that link concepts of the ontology. Our example represents a specific TYPE_OF *wind*, related to *cyclones*. *Cyclones* are divided into *tropical cyclone*, *extratropical cyclone* and *subtropical cyclone*. Types of *tropical cyclone* are *tropical depression* and *tropical storm* which vary according to the HAS_INTENSITY relation. On the other hand, *hurricane*, *typhoon*, *severe tropical cyclone*, and *severe cyclonic storm* are *tropical cyclones* that vary in the HAS_LOCATION relation. All these features are specified in the properties of the ontology.

Rules are model logic sentences that are always verified, and are frequently used to model knowledge that cannot be represented by means of classes, individuals and properties.

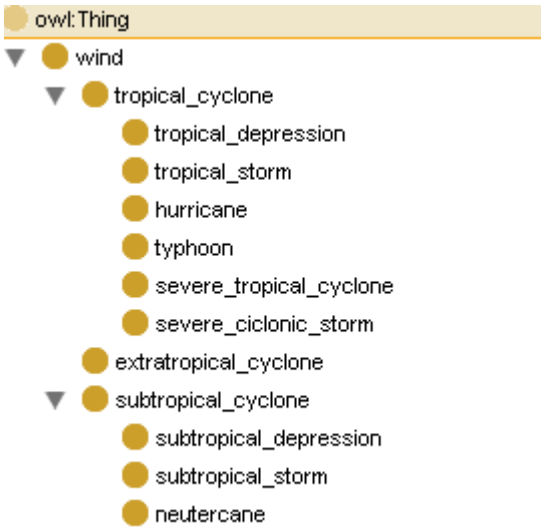

Fig. 3. Display of the incipient ontology

4.4. Semantic annotation with SMORE

SMORE is designed to allow users to tag documents in HTML with OWL through the use of web ontologies. In this respect, it provides a way of using classes, properties and individuals of existing ontologies with a view to editing them, or even creating a new ontology. For the purposes of our study, SMORE was useful because it permitted us to: (i) enrich our ontology with concepts extracted from texts and web documents that were not present previously in the ontology; (ii) annotate the corpus semantically.

The ontology can thus be enhanced simply by selecting a text from any web site or document, and clicking on the corresponding toolbar button to create OWL classes, properties and individuals. Apart from incorporating new concepts into the ontology, texts and web documents were annotated in order to subsequently study new relations between them.

Afterwards, a RDF/XML file was generated in which semantic annotations were established. As shown in Fig. 4, relations between the concepts in the ontology and the corpus were specified. We annotated *tropical cyclone* and incorporated new concepts related to the advisories and warning centres such as the *RSMC Tokyo Typhoon Centre Warning Area* linked to *typhoon*. The result was that each term in the text, instead of constituting a simple chain of characters, was located conceptually in a part of the ontology. This facilitated the recovery of information and the inference of knowledge.

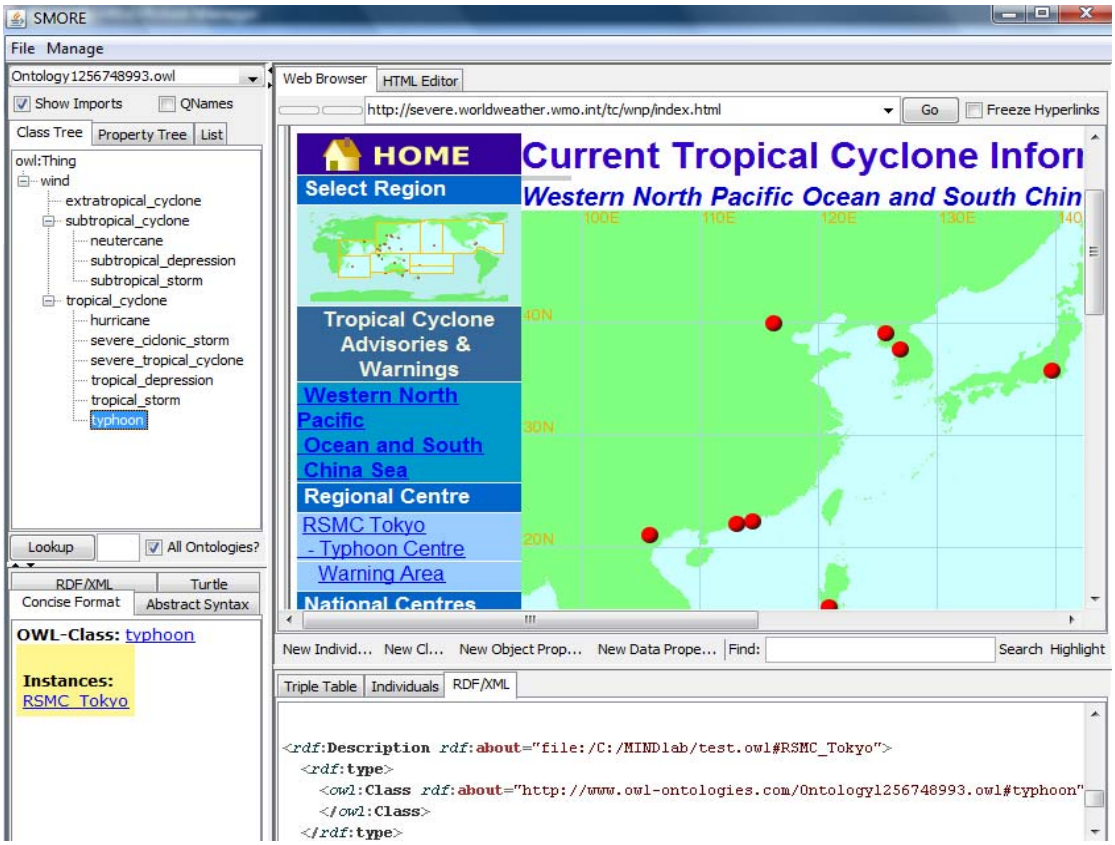

Fig. 4. Semantic annotation through SMORE

5. Conclusions

Although the Semantic Web is not as yet a reality, we are beginning to see the great advantages it will one day offer, such as permitting large-scale knowledge sharing.

In this preliminary study we have described a methodology for semantic annotation in the domain of Meteorology. The tagging system proposed is based on an incipient ontology currently being built. With the software application SMORE we have imported the ontology and performed semantic annotations, linking different sections of the texts (words, phrasemes, etc.) to the concepts of the ontology.

This preliminary linguistic analysis highlights the fact that these *semi-automatic* annotation systems help us annotate texts by using an ontology. Without a doubt, there is still a lot of work to do on the Semantic Web, mainly regarding *query agents* that will exploit the semantic annotations, map the queries to the ontology in order to identify only the most relevant data related to the Semantic Web area, and then offer semantic information, using inference mechanisms.

In a subsequent phase we plan to develop a query agent for the domain of Meteorology that would standardize all the concepts related to this particular domain

and whose methodology could be applied to other particular domains. The query agent would use inference mechanisms so that we would be able to not only consult explicit data, but also implicit information. This would resolve anomalies in searches, and eliminate irrelevant information.

Searches related to annotated resources linked to the concepts of the ontology would be performed through interfaces that allow us to access to information that permit the user to choose the more appropriate search criteria. For example, *wind* would lead to searches for TYPE_OF, INTENSITY, LOCALISATION (*geographical* and *the time of year*). This type of multidimensional searches, takes into account the semantic relations defined in the ontology where they were annotated.

It is our assertion that the methodology proposed could be used for the subsequent implementation of (semi)automatic tools based on ontologies.

REFERENCES

Aguado de Cea, G., Álvarez de Mon, I., Pareja-Lora, A., & Plaza-Arteche, R. (2002). RDF(S)/XML linguistic annotation of semantic web pages. Paper presented at the *International Conference on Computational Linguistics. Proceedings of the 2nd workshop on NLP and XML*, Taipei, Taiwan, *17*, 1-8. Retrieved from <http://portal.acm.org/citation.cfm?id=1118809> [25/10/09]

Aguado de Cea, G., Álvarez de Mon, I., & Pareja-Lora, A. (2003) Primeras aproximaciones a la anotación lingüístico-ontológica de documentos de web semántica: OntoTag. *Revista Iberoamericana De Inteligencia Artificial, 1*, 37-49.

Berners-Lee, T., Hendler, J. A., & Lassila, O. (2001). The semantic web. *Scientific American, 284*(5), 34-43. Retrieved from <http://www.sciam.com/article.cfm?id=the-semantic-web> [30/08/09]

Burnard, L. (2004). Metadata for corpus work. In M. Wynne (Ed.), *Developing linguistic corpora: A guide to good practice* (pp. 30-46). Oxford: Oxbow Books. Retrieved from <http://ahds.ac.uk/linguistic-corpora> [12/10/09]

EAGLES. (1996a). EAGLES: Recommendations for the morphosyntactic annotation of corpora. *EAGLES document EAG--TCWG--MAC/R*. Retrieved from <http://www.ilc.cnr.it/EAGLES/annotate/annotate.html> [10/05/09]

EAGLES. (1996b). EAGLES: Recommendations for the syntactic annotation of corpora. *EAGLES document EAG-TCWG-SASG/1.8.* Retrieved from <http://www.ilc.cnr.it/EAGLES/pub/eagles/corpora/sasg1.ps.gz> [10/05/09]

EAGLES. (1999). *EAGLES LE3-4244: Preliminary recommendations on lexical semantic encoding, final report.* Retrieved from <http://www.ilc.cnr.it/EAGLES/EAGLESLE.PDF> [10/05/09]

Faber, P., Márquez, C., & Vega, M. (2005). Framing terminology: A process-oriented approach. *META, 50*(4). Retrieved from <http://www.erudit.org/livre/meta/2005/000255co.pdf> [03/06/09]

Gruber, T. R. (1993). A translation approach to portable ontologies. *Journal on Knowledge Acquisition, 5*(2), 199-220.

Kiryakov, A., Popov, B., Ognyanoff, D., Manov, D., Kirilov, A., & Goranov, M. (2003). Semantic Annotation, Indexing and Retrieval. Paper presented at the *2nd International Semantic Web Conference (ISWC2003)*, Florida, USA, *2870* 484-499. Retrieved from <http://www.ontotext.com/publications/SemAIR_ISWC169.pdf> [22/10/09]

Leech, G. (1997). Introducing corpus annotation. In R. Garside, G. Leech & T. McEnery (Eds.), *Corpus annotation: Linguistic information from computer text corpora* (pp. 1-18). London: Longman.

Leech, G. (2004). Adding linguistic annotation. In M. Wynne (Ed.), *Developing linguistic corpora: A guide to good practice* (pp. 17-29). Oxford: Oxbow Books. Retrieved from <http://ahds.ac.uk/linguistic-corpora> [03/05/09]

Niremburg, S., & Raskin, V. (2001). *Ontological semantics.* Retrieved from <http://crl.nmsu.edu/Staff.pages/Technical/sergei/book/index-book.html> [20/09/09]

Weigand, H. (1997). Multilingual Ontology-Based Lexicon for News Filtering —The TREVI Project. Paper presented at the *15th International Joint Conferences on Artificial Intelligence (IJCAI-97)*, Nagoya, Aichi, Japan. 160-165. Retrieved from <http://crl.nmsu.edu/Events/IJCAI> [15/09/09]

[i] This research is part of the Project Ecosystem: Single Information Space for Frame-based Environmental Data and Thesaurus (FFI2008-06080-C03-01/FILO) funded by the Spanish Ministry for Science and Innovation.

[ii] <http://www.mindswap.org/2005/SMORE/>

[iii] Some recommendations regarding lexical-semantic annotation have been suggested, EAGLES (1999), but no standardization has been published for semantic annotation of corpora, contrary to other kinds of annotation, such as part-of-speech annotation (EAGLES, 1996a) or syntactic annotation (EAGLES, 1996b), that already have their standards.

[iv] For further information: <http://www.w3.org/TR/rdf-schema/>

[v] For further information: <http://www.mindswap.org/2003/CancerOntology/>

[vi] For further information: <http://www.opengalen.org/index.html>

[vii] A concordance is an example of a given word or phrase, which shows its context and combinatory potential.

[viii] http://www.lexically.net/wordsmith/

[ix] <http://protege.stanford.edu>