OntoTag: A Semantic Web Page Linguistic Annotation Model

Guadalupe Aguado de Cea[1], Inmaculada Álvarez de Mon[2], Antonio Pareja-Lora[3], Rosario Plaza-Arteche[4]

Abstract. Although with the *Semantic Web* initiative much research on web page semantic annotation has already been done by AI researchers, linguistic text annotation, including the semantic one, was originally developed in Corpus Linguistics and its results have been somehow neglected by AI. The purpose of the research presented in this proposal is to prove that integration of results in both fields is not only possible, but also highly useful in order to make Semantic Web pages more machine-readable. A multi-level (possibly multi-purpose and multi-language) annotation model based on EAGLES standards and Ontological Semantics, implemented with last generation Semantic Web languages is being developed to fit the needs of both communities.

1. INTRODUCTION.

All of us are by now used to making extensive use of the so-called World Wide Web (WWW) which we might consider a great source of information, accessible through computers but, hitherto, only understandable to human beings. In its beginning, web pages were hand made, intended and oriented to the exchange of information among human beings. All of these documents contained a huge amount of text, images and even sounds, meaningless to a computer. In this way, they put the burden of extracting and interpreting the relevant information on the reader. Due to the astonishing growth of Internet use, new technologies emerged and, with them, machine-aided web page generation appeared.

Currently, web page presentation in the WWW is being handled independently from its content, mainly through the use of XML [1] or other resource-oriented languages as XOL [2], SHOE [3], OML [4], RDF [5], RDF Schema [6], OIL [7] or DAML+OIL [8]. But even though the automatic process of information is being eased, still the above-mentioned tasks –relevant information access, extraction and interpretation– cannot be wholly performed by computers. Hence, the goal of enabling computers to understand the meaning (the semantics) of written texts and web pages is the main pillar sustaining the development of the *Semantic Web* [9]. In this context, the *semantic annotation of texts*, since it makes meaning explicit, has become a relevant topic and, therefore, advanced design and application of models and formalisms for the semantic annotation of web pages are needed.

Lately, much research has already been carried out by ontologists on the semantic annotation of web pages [3], [10], [11], [12]. However, such works have somehow neglected the results obtained on corpus annotation in the field of *Corpus Linguistics*, not only in the semantic level, but also in other linguistic levels. These other linguistic levels, whilst not being intrinsically semantic, can add extra semantic information to help a computer understand a text or, in our case, web pages.

The goal of this paper is to present the results of our research in which special efforts are being devoted to finding a way of bringing together and identifying complementarities between the semantic annotation models from AI and the annotations proposed by Corpus Linguistics.

This paper is organised as follows: firstly, an introduction to the state of the art in semantic annotation in corpus linguistics is presented (section 2). In section 3, some brief notes on the use of ontologies in semantic annotation are sketched. In section 4, an example of the integration of both paradigms (AI's and Corpus Linguistics') is presented in the scope of our project goals. The main advantages of this integration are then analysed –section 5– and, finally, further work to be done is included –section 6–.

2. SEMANTIC ANNOTATION IN CORPUS LINGUISTICS.

The idea of *text annotation* was originally developed in Corpus Linguistics. Traditionally, linguists have defined *corpus* as "a body of naturally occurring (authentic) language data which can be used as a basis for linguistic research" [13]. From this point of view, **Corpus Linguistics** [14] may not be considered a branch of Linguistics in itself, like syntax or semantics. The latter are focused on describing or explaining an aspect of language use; the former is rather a methodology or an approach which can be taken by these branches to explain or describe their particular aspect of language use. Following the same authors, Corpus Linguistics was first applied to research on language acquisition, to the teaching of a second language, to the elaboration of descriptive grammars, etc.. With the arrival of computers, the number of potential studies to which corpora could be applied increased exponentially.

[1] Department of Applied Linguistics to Science and Technology (DLACT), Computer Science Faculty, UPM, Madrid, Spain. lupe@fi.upm.es.

[2] DLACT, Telecommunications Engineering College, UPM, Madrid, Spain. ialvarez@euitt.upm.es.

[3] Department of Computer Systems and Programming (DSIP), Computer Science Faculty, UCM, Madrid, Spain. apareja@sip.ucm.es.

[4] Department of Applied Linguistics to Science and Technology (DLACT), Computer Science Faculty, UPM, Madrid, Spain. rplaza@fi.upm.es.

So, nowadays, the term **corpus** is being applied to "a body of language material which exists in electronic form, and which may be processed by computer for various purposes such as linguistic research and language engineering" [13]. An **annotated corpus** "may be considered to be a repository of linguistic information [...] made explicit through concrete annotation" [14]. The benefit of such an annotation is clear: it makes retrieving and analysing information about what is contained in the corpus quicker and easier. Let us now see the recommendations stated in Corpus Linguistics for text semantic annotation.

As asserted in [14], two broad types of semantic annotation may be identified, related to:

1. Semantic relationships between items in the text (i.e., the agents or patients of particular actions). This type of annotation has scarcely begun to be applied.
2. The semantic features of words in a text, essentially the annotation of word senses in one form or another. There is no universal agreement in semantics about which features of words should be annotated[5].

Although some preliminary recommendations on lexical semantic encoding have already been posited [15], no EAGLES semantic corpus annotation standard has yet been published; nevertheless, for choosing or devising a corpus semantic field[6] annotation system (second type of semantic annotation above mentioned) a set of reference criteria has been proposed by Schmidt and is presented in [16]. These criteria are:

1. *It should make sense in linguistic or psycholinguistic terms.* It is known from psycholinguistic experiments that certain basic categories exist in the mind. At present, in general, there is a good agreement between many basic categories we already know about from neuropsychology (for example colours, body parts, topography and so on); but still an exhaustive set of categories is to be determined. Overabstraction must be avoided, in any case.
2. *It should be able to account exhaustively for the vocabulary in the corpus, not just for a part of it.* If a term cannot readily be classified in the existing annotation system, then the system clearly needs to be amended.
3. *It should be sufficiently flexible to allow for those emendations that are necessary for treating a different period, language, register or textbase*. The treatment of specialised texts (such as computer-related, commerce, etc.) may require considerably more detailed subclassification of the domain in question than other texts.
4. *It should operate at an appropriate level of granularity (or delicacy of detail)* –related to criteria (3). What level of granularity is correct for an annotation system is an open question and depends partly on the aims of the end user. For this reason, the next criterion is posited.
5. *It should, where appropriate, possess a hierarchical structure*. If a semantic category system has a hierarchical structure, based on increasingly general levels of relatedness between terms, the end user can look at all the different levels and decide which one must employ, simply by moving up or down to the next level in the hierarchy.
6. *It should conform to a standard, if one exists*. A hard-and-fast system of categories, even being the result of a consensual work, may be rejected by many researchers. However, a standard in this level could lay, like EAGLES standards have done in other levels, a broad framework of principles and *major* categories. Such a standard would facilitate comparability and, at the same time, could be modified as necessary for individual needs[7].

3. ONTOLOGIES AND SEMANTIC WEB ANNOTATIONS.

AI researchers have found in *ontologies* [17], [18] the ideal knowledge model to formally describe web resources and its vocabulary and, hence, to make explicit in some way the underlying meaning of the terms included in web pages. With Ontological Semantics [19] as a support theory[8], the annotation of these web resources with ontological information should allow intelligent access to them, should ease searching and browsing within them and should exploit new web inference approaches from them. Many systems and projects have been developed: SHOE [3]; the $(KA)^2$ initiative [10]; PlanetOnto [11] and the Semantic Community Web Portals project [12]. Semantic annotation tools have also been developed so far: COHSE [20], MnM [21], OntoMat-Annotizer [22], SHOE Knowledge Annotator [23] and AeroDAML [24].

4. INTEGRATION OF PARADIGMS: AN EXAMPLE.

As we have already mentioned, the goal of this paper is to present the complementarities of linguistic and ontological annotation for the Semantic Web. The purpose of the project we are presenting, *ContentWeb*, is the creation of an ontology-based platform to enable users to query e-commerce applications by using natural language, performing the automatic retrieval of information from web documents annotated with ontological and linguistic information. *ContentWeb* objectives can be enunciated as follows:

1. Semi-automatic building of ontologies in the domains of e-commerce and of entertainment, reusing existing ontologies and international e-commerce standards and joint initiatives.
2. Elaboration of *OntoTag*, a model and environment for the hybrid –linguistic and ontological– annotation of web documents.
3. Development of *OntoConsult*, a natural language interface based on ontologies.

---

[5] See, for example, the controversies within the SENSEVAL initiative meetings – [30], [31].

[6] A **semantic field** (sometimes also called a conceptual field, a semantic domain or a lexical domain) is a theoretical construct which groups together words that are related by virtue of their being connected –at some level of generality– with the same mental concept [16].

[7] Once again the SENSEVAL initiatives [30], [31] must be mentioned: they reveal the demand for semantic standardization in the field of word sense disambiguation.

[8] Ontological Semantics [19] uses a constructed world model –the ontology– as the central resource for extracting and representing meaning of natural language texts, reasoning about knowledge derived from texts as well as generating natural language texts based on representations of their meaning.

```
<contentWeb:FilmReview>
    <contentWeb:text>Tras cinco años de espera y después de
            muchas habladurías, llega a nuestras pantallas la película
            más esperada de los últimos tiempos.</contentWeb:text>
</contentWeb:FilmReview>

<!-- Morpho-syntactic annotation excerpt -->

<morphAnnot:Word rdf:ID="1_16">
    <morphAnnot:surface_form>la</morphAnnot:surface_form>
    <morphAnnot:TradAnnot rdf:about="#trad_ann_info_1_16"/>
    <morphAnnot:MBTAnnot rdf:about="#mbt_ann_info_1_16"/>
    <morphAnnot:ConstrAnnot rdf:about="#constr_ann_info_1_16"/>
</morphAnnot:Word>

<morphAnnot:TradAnnot rdf:ID="trad_ann_info_1_16">
    <trad:tag> ARTDFS </trad:tag>
    <morphAnnot:lemma> el </morphAnnot:lemma>
</morphAnnot:TradAnnot>

<morphAnnot:MBTAnnot rdf:ID="mbt_ann_info_1_16">
    <mbt:tag> TDFS0 </mbt:tag>
    <morphAnnot:lemma> el </morphAnnot:lemma>
</morphAnnot:MBTAnnot>

<morphAnnot:ConstrAnnot rdf:ID="constr_ann_info_1_16">
    <constr:tag> DET </constr:tag>
            <constr:genus>FEM</constr:genus>
            <constr:numerus>SG</constr:numerus>
    <morphAnnot:lemma>la</morphAnnot:lemma>
    <constr:synfunction>DN&gt;</constr:synfunction>
</morphAnnot:ConstrAnnot>
```

Figure 1: Morphosyntactic annotation of the article *"la".*

4. Creation of *OntoAdvice*, an ontology-based system for querying and retrieving information from annotated web documents in the entertainment domain.

One of the tasks performed to reach goal 2 is the manual annotation of a Spanish sentence *"Tras cinco años de espera y después de muchas habladurías, llega a nuestras pantallas la película más esperada de los últimos tiempos."* ("After five years of expectation and gossiping, here comes the most expected film for the time being.") on the languages XML and RDF(S). The RDF(S) annotation of this sentence in the first three levels is shown in Figure 1, Figure 2 and Figure 3.

In the morphosyntactic level (Figure 1) every word or lexical token is given a different Uniform Resource Identifier (URI). The morphosyntactic annotation of the article *"la"*, according to three different tagsets and systems is presented. Each tagset has been assigned a different class in the morphAnnot namespace: *TradAnnot* (CRATER tagset), *MBTAnnot* (MBT tagset [25]) and *ConstrAnnot* (Constraint Grammar - CONEXOR tagset [26]). For the sake of space, just the annotation of the article "la" has been included in the figure.

In the syntactic level (Figure 2) every syntactic relationship between morpho-syntactic items is given a new URI, so that it can be referenced in higher-level relationships or by other levels of the annotation model (i.e. *<synAnnot:Chunk rdf:ID="1_510">*). The annotation of the phrase *"la película más esperada de los últimos tiempos"* has been included in the figure.

In the semantic level (see Figure 3) some components of lower level annotations are tagged with semantic references to the concepts, attributes and relationships determined by our (domain) ontology, implemented in the language DAML+OIL.

5. ADVANTAGES OF THE INTEGRATED MODEL.

As shown in the example from section 4, it seems that AI and Corpus Linguistics, far from being irreconcilable, can join together to give birth to an integrated annotation model. This conjunct annotation scheme would be very useful and valuable in the development of the Semantic Web and would benefit from the results of both disciplines in many ways. Let us now see the benefits at the semantic level of a hybrid annotation model, first from a linguistic point of view and, then, from an ontological point of view.

```
<!-- Syntactic annotation excerpt -->

<synAnnot:Chunk rdf:ID="1_510">
    <synAnnot:synfunction>NP</synAnnot:synfunction>
    <synAnnot:hasChild rdf:about="#1_21">los</synAnnot:hasChild>
    <synAnnot:hasChild rdf:about="#1_22">últimos</synAnnot:hasChild>
    <synAnnot:hasChild rdf:about="#1_23">tiempos</synAnnot:hasChild>
</synAnnot:Chunk>

<synAnnot:Chunk rdf:ID="1_511">
    <synAnnot:synfunction>PP</synAnnot:synfunction>
    <synAnnot:hasChild rdf:about="#1_20">de</synAnnot:hasChild>
    <synAnnot:hasChild rdf:about="#1_510"> los últimos tiempos
            </synAnnot:hasChild>
</synAnnot:Chunk>

<synAnnot:Chunk rdf:ID="1_512">
    <synAnnot:synfunction>AdjP</synAnnot:synfunction>
    <synAnnot:hasChild rdf:about="#1_18">más</synAnnot:hasChild>
    <synAnnot:hasChild rdf:about="#1_19">esperada</synAnnot:hasChild>
    <synAnnot:hasChild rdf:about="#1_511">de los últimos tiempos
            </synAnnot:hasChild>
</synAnnot:Chunk>

<synAnnot:Chunk rdf:ID="1_513">
    <synAnnot:synfunction>NP</synAnnot:synfunction>
    <synAnnot:hasChild rdf:about="#1_16">la</synAnnot:hasChild>
    <synAnnot:hasChild rdf:about="#1_17">película</synAnnot:hasChild>
    <synAnnot:hasChild rdf:about="#1_512">más esperada de los últimos
            tiempos </synAnnot:hasChild>
</synAnnot:Chunk>
```

Figure 2: Syntactic annotation of the chunk *"la película más esperada de los últimos tiempos"* in RDF(S).

```
<!-- Semantic annotation excerpt -->

<onto:PremiereEvent rdf:ID="_anon27">
    <semSynAnnot:includes rdf:about="#1_13">llega</semSynAnnot:includes>
    <semSynAnnot:includes rdf:about="#1_509">a nuestras pantallas</semSynAnnot:includes>
    <onto:hasFilm rdf:about="#_anon30"/>
</onto:PremiereEvent>

<onto:Film rdf:ID="_anon30">
    <semAnnot:includes rdf:about="#1_18">película</semAnnot:includes>
    <onto:comment rdf:about="#_anon40">
    <onto:comment rdf:about="#_anon41">
</onto:Film>

<onto:ControversialFilm rdf:ID="_anon40">
     <semSynAnnot:includes rdf:about="#1_506">después de muchas habladurías</semSynAnnot:includes>
</onto:ControversialFilm>

<onto:AwaitedFilm rdf:ID="_anon41">
     <semSynAnnot:includes rdf:about="#1_503">Tras cinco años de espera</semSynAnnot:includes>
     <semSynAnnot:includes rdf:about="#1_512">más esperada de los últimos tiempos</semSynAnnot:includes>
</onto:ControversialFilm>

<onto:Film rdf:about="#_anon30">
     <semSynAnnot:includes rdf:about="#3_507">El Señor de los Anillos</semSynAnnot:includes>
     <onto:filmTitle>El Señor de los Anillos</onto:filmTitle>
</onto:Film>
```

Figure 3: Semantic annotation of *"Tras cinco años de espera y después de muchas habladurías, llega a nuestras pantallas la película más esperada de los últimos tiempos."* in RDF(S).

5.1. Regarding ontology-based annotations from a linguistic point of view.

The first result of our work is that the use of ontologies as a basis for a semantic annotation scheme fits perfectly and accomplishes the criteria posited by Schmidt. Clearly, its mostly hierarchical structure fulfils by itself criterion (5) and, as a side effect, criteria (2) and (4), since an ontology can grow horizontally (in breadth) and vertically (in depth). Criterion (3) is also satisfied by an ontology-based semantic annotation scheme, since we can always specialise the concepts in the ontology according to specific periods, languages, registers and textbases. Ontologies are, by definition, consensual and, thus, are closer to becoming a standard than many other knowledge models, as criteria (6) requires. Concerning criterion (1), quite a lot of groups developing ontologies are characterized by a strong interdisciplinary approach that combines Computer Science, Linguistics and (sometimes) Philosophy; then, an ontology-based approach should also make sense in linguistic terms.

5.2. Regarding linguistic annotations from an ontological point of view.

The main drawback for AI researchers to adopt a linguistically motivated annotation model would lie on the fact that (section 2) "there is no universal agreement in semantics about which features of words should be annotated" or on Schmidt's criterion (1): "still an exhaustive set of categories is to be determined". But ontology researchers are trying to fill this gap with initiatives such as the UNSPSC [27] or RosettaNet [28] in specific domains (i.e. e-commerce). In any case, linguistic annotations at the semantic level are more ambitious and potentially wider than the strictly ontology-based ones. Establishing a link between semantic annotation and discourse annotation and text construction following the RST approach, which has already been applied in text generation [29], seems a fairly promising linguistic enhancement.

6. CONCLUSIONS AND FURTHER WORK.

This paper has shown the results of the research carried out on how linguistic annotation can help computers understand the text contained in a document –a Semantic Web page– bringing together semantic annotation models from AI and the annotations proposed for every linguistic level from Corpus Linguistics.

Further elements susceptible of semantic annotation are presently being sought and research is being done towards their determination by the team of linguists in our project. The pragmatic counterpart of OntoTag has not yet been tackled at this phase of the project.

Still, much work must be done in order to fully specify, implement and assess the whole model. Besides, many efforts are being devoted to developing OntoAdvice, the ontology-based information retrieval system, in order to validate this model.

ACKNOWLEDGEMENTS.

The research described in this paper is supported by MCyT (Spanish Ministry of Science and Technology) under the project name: ContentWeb: "PLATAFORMA TECNOLÓGICA PARA LA WEB SEMÁNTICA: ONTOLOGÍAS, ANÁLISIS DE LENGUAJE NATURAL Y COMERCIO ELECTRÓNICO" – TIC2001-2745 ("ContentWeb: Semantic Web Technologic

Platform: Ontologies, Natural Language Analysis and E-Business"). We would also like to thank Socorro Bernardos, Óscar Corcho and Mariano Fernández for their help with the ontological aspects of this paper.

REFERENCES.

[1] Bray, T., Paoli, J., Sperberg, C. (1998) *Extensible Markup Language (XML) 1.0.* W3C Recommendation. http://www.w3.org/TR/REC-xml

[2] Karp, R., Chaudhri, V., Thomere, J. (1999) *XOL: An XML-Based Ontology Exchange Language*. Technical Report. http://www.ai.sri.com/~pkarp/xol/xol.html

[3] Luke S., Heflin J. (2000) *SHOE 1.01. Proposed Specification*. SHOE Project. http://www.cs.umd.edu/projects/plus/SHOE/spec1.01.htm

[4] Kent, R. (1998) *Conceptual Knowledge Markup Language (version 0.2).* http://sern.ucalgary.ca/KSI/KAW/KAW99/papers/Kent1/CKML.pdf

[5] Lassila, O., Swick, R. (1999) *Resource Description Framework (RDF) Model and Syntax Specification*. W3C Recommendation. http://www.w3.org/TR/PR-rdf-syntax

[6] Brickley, D., Guha, R.V. (2000) *Resource Description Framework (RDF) Schema Specification*. W3C Candidate Recommendation. http://www.w3.org/TR/PR-rdf-schema.

[7] Horrocks, I., Fensel, D., Harmelen, F., Decker, S., Erdmann, M, Klein, M. (2000) OIL in a Nutshell. In *12th International Conference in Knowledge Engineering and Knowledge Management, Lecture Notes in Artificial Intelligence*, 1–16. Berlin, Germany: Springer-Verlag. http://www.cs.vu.nl/~ontoknow/oil/downl/oilnutshell.pdf

[8] Horrocks, I., Van Harmelen, F. (2001) *Reference description of the DAML+OIL ontology markup language.* Draft report, 2001. http://www.daml.org/2000/12/reference.html

[9] Berners-Lee, T., Fischetti, M. (1999) *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by its Inventor.* San Francisco: Harper.

[10] Benjamins, V.R., Fensel, D., Decker, S., Gómez-Pérez, A. (1999) $(KA)^2$: Building Ontologies for the Internet: a Mid Term Report. *IJHCS, International Journal of Human Computer Studies*, 51: 687–712.

[11] Motta, E., Buckingham Shum, S. Domingue, J. (1999) Case Studies in Ontology-Driven Document Enrichment. In Proceedings of the 12th Banff Knowledge Acquisition Workshop, Banff, Alberta, Canada.

[12] Staab, S., Angele, J., Decker, S., Erdmann, M., Hotho, A., Mädche, A., Schnurr, H.-P., Studer, R. (2000) *Semantic Community Web Portals.* WWW´9. Amsterdam.

[13] Leech, G. (1997a) Introducing corpus annotation. In Garside R., Leech, G., McEnery, A. M. (eds.) *Corpus Annotation: Linguistic Information from Computer Text Corpora*. London: Longman.

[14] McEnery, A. M., Wilson, A. (2001) *Corpus Linguistics: An Introduction.* Edinburgh: Edinburgh University Press.

[15] EAGLES (1999) *EAGLES LE3-4244: Preliminary Recommendations on Semantic Encoding*, Final Report. http://www.ilc.pi.cnr.it/EAGLES/EAGLESLE.PDF

[16] Wilson, A., Thomas, J. (1997) Semantic Annotation. In R. Garside, G. Leech & A. M. McEnery, (eds.) *Corpus Annotation: Linguistic Information from Computer Text Corpora*. London: Longman.

[17] Gruber, R. (1993) A translation approach to portable ontology specification. *Knowledge Acquisition.* #5: 199-220.

[18] Studer, R., Benjamins, R., Fensel, D. (1998) *Knowledge Engineering: Principles and Methods*. DKE 25(1-2): 161-197.

[19] Nirenburg, S. and Raskin, V. (2001) *Ontological Semantics (Draft)* http://crl.nmsu.edu/Staff.pages/Technical/sergei/book/index-book.html.

[20] COHSE (2002) http://cohse.semanticweb.org/

[21] Vargas-Vera, M., Motta, E., Domingue, J., Shum, S. B., Lanzoni, M. (2001) Knowledge Extraction by Using an Ontology-based Annotation Tool. In Proceedings of the K-CAP'01 Workshop on Knowledge Markup and Semantic Annotation, Victoria B.C., Canada.

[22] OntoMat (2002) http://annotation.semanticweb.org/ontomat.html

[23] SHOE (2002) http://www.cs.umd.edu/projects/plus/SHOE/KnowledgeAnnotator.html

[24] AeroDAML (2002) http://ubot.lockheedmartin.com/ubot/hotdaml/aerodaml.html

[25] MBT (2002) http://ilk.kub.nl/~zavrel/tagtest.html

[26] Conexor OY (2002) http://www.conexoroy.com/products.htm

[27] UNSPSC (2002) *Universal Standard Products and Services Classification (UNSPSC).* http://www.unspsc.org/

[28] RosettaNet (2002) *RosettaNet: Lingua Franca for eBusiness.* http://www.rosettanet.org/

[29] Mann, W., Thomson, S. (1988) *Rhetorical Structure Theory: Toward a functional theory of text organization.* Text Vol.18, 3: 243–281.

[30] Kilgarriff, A. (1998) SENSEVAL: An Exercise in Evaluating Word Sense Disambiguation Programs. In *Proceedings of LREC*, Granada, Spain, pp. 581–588.

[31] Kilgarriff, A. & Rosenzweig, J. (2000) English SENSEVAL: Report and Results. In *Proceedings of LREC*. Athens, Greece.