A SEMANTIC WEB PAGE LINGUISTIC ANNOTATION MODEL

Guadalupe Aguado de Cea*, Inmaculada Álvarez de Mon**, Asunción Gómez-Pérez***,
Antonio Pareja-Lora****, Rosario Plaza-Arteche*

* Dep. Lingüística Aplicada a la Ciencia y a la Tecnología (DLACT), Facultad de Informática, UPM, Madrid, Spain, 28660
** DLACT, Escuela Universitaria de Ingenieros de Telecomunicación, UPM, Madrid, Spain, 28031
*** Dep. Inteligencia Artificial (DIA), Facultad de Informática, UPM, Madrid, Spain, 28660
**** Dep. Sistemas Informáticos y Programación (DSIP), Facultad de Informática, UCM, Madrid, Spain, 28040
{lupe, asun, rplaza}@fi.upm.es , ialvarez@euitt.upm.es , apareja@sip.ucm.es

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 on the reader the burden of extracting and interpreting the relevant information in them. 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 (Bray et al., 1998) or other resource-oriented languages as XOL (Karp et al., 1999), SHOE (Luke et al., 2000), OML (Kent, 1998), RDF (Lassila et al., 1999), RDF Schema (Brickley et al., 2000), OIL (Horrocks et al., 2000) or DAML+OIL (Horrocks et al., 2001). 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* (Berners-Lee et al., 1999). In this context, the *semantic annotation of texts*, since it makes meaning explicit, has become a relevant topic. Thus, 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 (Luke et al., 2000), (Benjamins et al., 1999), (Motta et al., 1999), (Staab et al., 2000). However, such works have been neglecting, somehow, the results obtained in the field of *Corpus Linguistics* on corpus annotation, 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 on how linguistic annotation can help computers understand the text contained in a Semantic Web document. Special efforts are being devoted to finding a way of conjugating 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 text annotation in corpus linguistics will be presented (section 2). In subsection 2.1, high-level recommendations given for the main levels of annotation are also included, together with (subsection 2.2) a presentation of these linguistic annotation levels, namely: lemma, morphosyntactic, syntactic, semantic and discourse annotation. In subsection 2.3, the EAGLES standards on morphosyntactic and syntactic annotation will be enunciated. In section 3, some brief notes on the use of ontologies in semantic annotation will be sketched. In section 4, an example of the integration of both paradigms (AI's and Corpus Linguistics') will be presented in the scope of our project goals. The main advantages of this integration will be analysed afterwards –section 5– and, finally, some conclusions will be stated –section 6–.

2. Text 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" (Leech, 1997a). From this point of view, **Corpus Linguistics** (McEnery & Wilson, 2001) 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 or 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. 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" (Leech, 1997a). An **annotated corpus** "may be considered to be a repository of linguistic information [...] made explicit through concrete annotation" (McEnery & Wilson, 2001). 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 annotation and the different levels to which it is applicable.

2.1. General Recommendations for Text Annotation.

In (Leech, 1997a) and (McEnery & Wilson, 2001) a set of practical guidelines, standards or recommendations of good practice applicable to text annotation are suggested, namely:

1. The original text should be easily *recoverable* by taking away the annotations to it added.
2. Annotations should be facilely *extricable* from the annotated text.
3. Every annotated text must be accompanied with a thorough *documentation,* including, among others, the **annotation scheme** –the particular and precise guidelines used to annotate a text–, how (manually and/or automatically), by whom the text was annotated and the quality of the annotation (e.g. an accuracy rate).
4. The corpus annotation is *not infallible*: any act of annotation is also an act of interpretation.
5. Annotation schemes should be based as far as possible on *consensual*, widely agreed and theory-neutral principles.
6. *No* annotation scheme should claim authority as an absolute *standard.*

It is obvious that the inclusion of recommendation (4) in an annotation scheme requires only inserting a few lines in its documentation manual; currently, recommendations (1), (2) and (3) –to some extent– are easily fulfilled with the use of HTML, XML or similar mark-up languages[1]. Recommendation (5) can be accomplished through the use of broadly known ontologies or by the definition of some kind of standard, but this latter would prevent recommendation (6) from being fulfilled. Considering that research funding authorities are highly encouraging the unification and standardisation of annotation schemes (through the EU EAGLES initiative, for example (EAGLES, 1996a), (EAGLES, 1996b), (EAGLES, 1996c)) we must come to the conclusion that recommendation (6) must be at least relaxed. As stated in (EAGLES, 1996b), as for a standard "there is no absolute normative prescription of annotation practices, but at most a set of recommendations (criteria) from which the annotator may justify departures or extensions for particular purposes".

In fact, one of the results of the EAGLES project work is the **Corpus Encoding Standard** (CES) (CES, 1999), which include some general criteria, which should be considered when elaborating an annotation scheme. These criteria are:

1. *Adequate coverage*: Most linguistic features and properties of texts must be susceptible of annotation but it is desirable that no unnecessary elements are included in the annotation scheme.
2. *Consistency*: An annotation scheme should be built around consistent principles to determine what kind of objects are tags, what kind of objects are attributes, what kind of object(s) appear as tag content, etc.
3. *Recoverability*: The annotation scheme must enable recovering the source text from its annotated version (analogous to recommendation (1)).
4. *Validatability*: The validation of a text annotation must be possible, understanding validation as the process by which software checks that the mark-up in a document conforms to the structural specifications given in a SGML, HTML or XML DTD[2].
5. *Capturability*: The annotation scheme should accommodate the various levels of analysis of the text and it should also be *refinable*, by providing tags at various levels of specificity together with a *taxonomy* identifying the hierarchical relations among them.
6. *Processability*: An annotation scheme must be designed taking into account (computer) processing considerations and needs.

[1] These recommendations were made before the family of mark-up languages such as SGML, HTML and XML was fully developed.

[2] The CES was developed following the TEI recommendations and, thus, presupposes the use of SGML as encoding language. Since HTML and XML are also TEI-conformant, but were less extended or even unknown when the TEI recommendations where stated, have been included here for the sake of generalisation.

7. *Extensibility*: It is essential that systematic means for extension of the annotation scheme be developed, to ensure that extensions are made in a controlled and predictable way.
8. *Compactness*: In order to reduce the number of characters added to an annotated text.
9. *Readability*: Annotated texts must be intelligible.

Two criteria out of the nine above stated are considered secondary: *compactness* (8) and *readability* (9), since most texts nowadays can be viewed and processed with appropriate software, which can reduce the impact of handling non-compact or non-easily-readable annotated texts (EAGLES, 1996b). Criterion (5) introduces a new concept: the layered levels of linguistic analysis, which generate their own different annotation types, to be presented in the next section.

2.2. Levels of Linguistic Annotation.

In (Leech, 1997a), a list of the different levels of linguistic annotation can be found. As Leech states, no corpus includes all of them, but only two or, at most, three of them. Some of them were only in their first state of conception at the time of writing his paper. A smaller but more realistic list of annotation levels (lemma, morphosyntactic, syntactic, semantic and discourse) included in (EAGLES, 1996b) is introduced in the next subsections.

2.2.1. Lemma Annotation.

Lemma annotation (lemmatisation) accompanies every word-token in a text with its **lemma**, that is, the head word form that one would look up if one were looking for the word in a dictionary. In English, lemma annotation may be considered redundant but, in more highly-inflected languages, such as Spanish, the ratio of word-forms per lemma makes lemma annotation a very valuable contribution to information extraction (Leech, 1997a).

2.2.2. Morphosyntactic Annotation.

This is one of the most extended types of annotation in *Corpus Linguistics*, together with the syntactic annotation. *Morphosyntactic annotation*, *part of speech annotation*, *POS tagging* or *grammatical tagging* is the annotation of the grammatical class (e.g. noun, verb, etc.) of each word-token in a text[3], together with (possibly) the annotation of its morphological analysis. As claimed in (McEnery & Wilson, 2001), POS information forms an essential foundation for further forms of analysis such as syntactic parsing and semantic field annotation. Even though a computer can carry out this task currently with a high degree of accuracy without manual intervention, it must not be thought of as trivial. Disambiguation of homographs, identification of word idiomatic sequences and compounds or separation of contracted forms are some of the different irregularities an annotator must face at this level. This is due to the fact that a one-to-one correspondence between orthographic words and morphosyntactic words cannot be established (Leech, 1997b). Solutions for these problems can be found in (McEnery & Wilson, 2001), (Leech, 1997b) and, for Spanish, in (Pino & Santalla, 1996).

2.2.3. Syntactic Annotation.

Once the morphosyntactic categories in a text have been identified, the *syntactic annotation* adds the annotation of the higher-level syntactic relationships between these categories, determined e.g. by means of a phrase-structure or dependency parse. Different parsing schemes are employed by different annotators; according to (McEnery & Wilson, 2001), these schemes differ in:

§ The number of constituent types they employ (typically, the number of tags in the POS tagset).
§ The way in which constituents are permitted to combine with one another.
§ The grammar followed to parse and annotate the text.

2.2.4. Semantic Annotation.

As asserted in (McEnery & Wilson, 2001), 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. 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[4].

Although some preliminary recommendations on lexical semantic encoding have already been posited (EAGLES, 1999), no EAGLES semantic corpus annotation standard has yet been published; nevertheless, for the second type of semantic annotation alluded, a set of reference criteria has been proposed by Schmidt and mentioned in (Wilson & Thomas, 1997) for choosing or devising a corpus semantic field[5] annotation system. 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

[3] In other words, a POS tagging system holds the answer to the questions: a) How to divide the text into individual word tokens (words) b) How to choose a tagset (= a set of word categories to be applied to the word tokens) c) How to choose which tag is to be applied to which word (token).

[4] See, for example, the controversies within the SENSEVAL initiative meetings – (Kilgarriff, 1998), (Kilgarriff & Rosenzweig, 2000).

[5] 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 (Wilson & Thomas, 1997).

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[6].

2.2.5. Discourse Annotation.

This is the least frequently encountered kind of annotation (in corpora). Still, two main different kinds of approaches on annotation at this level can be found. *Stenström's approach* (McEnery & Wilson, 2001) is based on what she called *discourse tags*, derived empirically from an initial analysis of a subsample of a corpus. These included categories such as 'apologies' (e.g. *sorry, excuse me*) or 'greetings' (e.g. *hello, good evening*) and were used to mark items whose role in the discourse dealt primarily with discourse management rather than with the propositional content. This first approach has never become widely used in corpus linguistics. Conversely, the pronoun reference or *anaphoric annotation* approach considers *cohesion*[7] as a crucial factor in our understanding of the processes involved in reading, producing and comprehending discourse. A clear exponent of this approach is the UCREL discourse annotation scheme, together with many other anaphoric annotation schemes, such as De Rocha's, Gaizauskas and Humphries' and Botley's (Garside et al., 1997).

2.3. EAGLES Recommendations for (Corpora) Annotation.

For some of the above levels of annotation, a consensus about what, to what extent and how must be annotated has been achieved through the EU EAGLES initiative, which provides recommendations gathered up in a set of documents of good practices for annotation. These recommendations share some principles (EAGLES, 1996b), (EAGLES, 1996c):

1. Make use of an attribute-value formalism.
2. Do not adhere to a strict attribute-value hierarchy (in terms of monotonic inheritance).
3. Use three sublevels of constraint (obligatory, recommended and optional) in defining what is acceptable according to the guidelines.
   - § **Obligatory annotations** are required if the annotation scheme for that level is to be conformant with EAGLES standards.
   - § **Recommended annotations** are not required, but should not be omitted. The standard requirement for these recommended attributes and values is that, if they occur in a particular language, then it is advisable that the tagset of that particular language should encode them.
   - § **Optional annotations** are not required nor recommended, but are specific to a (set of) language(s) or a language engineering application.

Let us now see which attributes and values are considered as obligatory, recommended and optional in EAGLES for morphosyntactic (EAGLES, 1996b) and syntactic (EAGLES, 1996c) annotation recommendations, the only ones made public up to now.

2.3.1. Morphosyntactic Level.

- § Only one attribute is considered *obligatory*: that of the major word categories, or parts of speech (N–noun–, V –verb–, AJ –adjective–, etc.).
- § Attributes such as: type –common/proper–, gender, number or case are *recommended* for nouns, as well as person, gender, number, tense, voice, etc. for verbs, and degree, gender, number and case for adjectives.
- § *Optional* attributes and values, or *special extensions*, as they are called in this document, are subdivided into:

[6] Once again the SENSEVAL initiatives must be mentioned: they reveal the demand for semantic standardization in the field of word sense disambiguation (Kilgarriff, 1998), (Kilgarriff & Rosenzweig, 2000).

[7] **Cohesion** (Halliday & Hasan, 1976) is the vehicle by which elements in texts are interconnected through the use of pronouns, repetition, etc..

» *Optional generic attributes and values:* For instance, countability–countable/mass– for nouns; aspect–perfective/imperfective–, separability–non-separable/separable–, etc. for verbs.
» *Optional language-specific attributes and values*: For instance, definiteness –definite/ indefinite/unmarked– for Danish nouns.

2.3.2. Syntactic Level.

§ It is suggested that no part of the syntactic annotation be regarded as *obligatory*, since syntactic annotations can take different forms, according to the grammar they are based on (for example, phrase structure grammar, dependency grammar or functional grammar).
§ If a phrase structure annotation is adopted (no hints are given in other cases) the following categories are *recommended*: Sentence, Clause, Noun Phrase, Verb Phrase, Adjective Phrase, Adverb Phrase and Prepositional Phrase.
§ Examples of *optional* annotations include the marking of sentence types (Question, Imperative, etc.), the functional annotation of subjects and objects and the identification of semantic subtypes of constituents such as adverbial phrases.

3. Ontologies and Semantic Web Annotations.

AI researchers have found in *ontologies* (Gruber, 1993), (Studer et al., 1998) 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 (Niremburg & Raskin, 2001) 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 (Luke et al., 2000); the $(KA)^2$ initiative (Benjamins et al., 1999); PlanetOnto (Motta et al., 1999) and the Semantic Community Web Portals project (Staab et al., 2000). Semantic annotation tools have also been developed so far: COHSE (COHSE, 2002), MnM (Vargas-Vera et al., 2001), OntoMat-Annotizer (OntoMat, 2002), SHOE Knowledge Annotator (SHOE, 2002) and AeroDAML (AeroDAML, 2002).

[8] Ontological Semantics (Niremburg & Raskin, 2001) is a theory of meaning in natural language and an approach to natural language processing (NLP) which 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. Integration of Paradigms: an Example.

As we have already mentioned, the goal of this paper is to present the complementarity 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 existent 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.
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 (MBT, 2002)) and *ConstrAnnot* (Constraint Grammar - CONEXOR tagset (Conexor, 2002)). 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.

```
<!-- 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).

In the semantic level (see Figure 3) some components of lower level annotations are annotated with semantic references to the concepts, attributes and relationships determined by our (domain) ontology, implemented in the language DAML+OIL. Further elements susceptible of semantic annotation are being sought and research is being done towards their determination by the linguist team in our project. The pragmatic counterpart of OntoTag has not yet been tackled at this phase of the project and, thus, this level is not included in the example.

5. Advantages of the Integrated Model.

As shown in the previous example from the previous section, 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: first, at the semantic level; second, at the rest of levels. Finally, particular subsections are dedicated to re-usability and multi-functionality.

5.1. At the Semantic Level.

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.

5.1.1. 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 the 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.

5.1.2. 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 (subsection 2.2.4) "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 (UNSPSC, 2002) or RosettaNet (RosettaNet, 2002) 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 (Mann & Thomson, 1988), seems a fairly promising linguistic enhancement.

So far, we have seen how ontologies can fit in the semantic annotation of texts; let us see in the next subsections how linguistic annotations in all of its levels can improve the potential of Semantic Web Pages.

5.2. Meaning Is Not Only within Semantics.

As stated in (Pulman, 1995), all linguistic levels interact closely in order to determine the meaning of a whole sentence, utterance or expression. On the one hand, even though the basic constituents of an expression[9] will ultimately be the meanings of words, an expression meaning will be characterised not only by its word meanings, but also by the manner in which they are put together. Since the modes of constituent combination are largely determined by the syntactic structure of the language, we will need to capture the piece of meaning given by every syntactic rule applied to generate the expression being analysed, that is, the semantic operation combining the meanings of the (parse) children to produce the meaning of the father. Hence, *we need the parse of an expression to help determine its meaning*. (Dik, 1989), (Aguado & Pareja-Lora, 2000) and (Vargas-Vera et al., 2001) reinforced the importance of mixing the syntactic and semantic. On the other hand, Pulman also pointed out the need for more integration between sentence or utterance level semantics and theories of text or dialogue structure, including aspects such as dialogue or text settings, or on the goals of speakers. Thus, *some kind of explicit or implicit pragmatic analysis has to be done, to help determine the meaning of the expressions in a text.* So, we come to the conclusion that it would be very useful for the Semantic Web community to have some model of annotation that allows not only the semantic level to be annotated and made explicit, but also allows the other levels to contribute to the machine-readability of web pages by their inclusion and explicit annotation in Semantic Web pages.

[9] Much of the information contained in a web page is given in a sub-sentential form (mainly nominal phrases). Thus, the term expression is preferred henceforth.

5.2.1. Meaning and Lemma Annotation.

Lemmatisation may be a valuable contribution, for example, to facilitate information extraction for highly-inflected languages, such as Spanish or German (Kietz et al., 2000). This is particularly true when ontologies are considered: lemmatisation annotation paves the way for an ontology-based (semi-)automatic semantic annotation.

5.2.2. Meaning and Morphosyntactic Annotation.

Many ontology-based information extraction projects make use of some kind of morpho-syntactic analysis ((Vargas-Vera et al., 2001), (Kietz et al., 2000)) as a preliminary phase towards semantic processing. Then, we must consider POS tagging as a kind of 'base camp' annotation, a first step towards more difficult levels of annotation such as those of syntax and semantics. As stated in subsection 2.2.2, some nominal groups and phrases and other idiomatic word sequences or phraseology (e.g.: "llega a nuestras pantallas", "El Señor de los Anillos"[10]) should be identified and marked as a lexical unit and annotated consistently. A smart way of achieving this goal for Spanish can be found in (Pino & Santalla, 1996).

5.2.3. Meaning and Syntactic Annotation.

Once again we must mention (Vargas-Vera et al., 2001) and (Kietz et al., 2000), since the projects there described make use of some kind of syntactic analysis when processing documents. Two kinds of syntactic annotations are considered to be very useful from a semantic point of view:

1. EAGLES optional annotations such as sentence type marking (Question, Imperative, etc.), subject and object functional annotation or constituent (i.e. adverbial phrases) semantic subtype identification.
2. Particular syntactic language phenomena, such as separable verb identification and marking for German.

5.2.4. Meaning and Discourse Annotation.

Since this level of annotation is to be tackled in further stages of our project, we can only remind the potential usefulness of an anaphoric annotation scheme in order to bring the cohesion out of the document processed.

5.3. Reusability.

As stated above, the need for (shallow) parsing in web page semantic processing is found in (Vargas-Vera et al., 2001) and also in (Kietz et al., 2000): most information extraction systems (as well as other NLP applications) use some form of shallow parsing[11] to recognise syntactic constructs or, in other words, to syntactically identify some fragments of the sentences. Thus, the process of semantically analysing a web page gets complicated and its speed reduced. Although the process of creation and edition of a page might seem then overwhelming, we must not forget that some tools are freely available for these (research) purposes. In this way, tools are reused, together with the results they render which are included as web page annotations (see example in section 4).

[10] (A film) is premiered, "The Lord of the Rings": Both examples have been extracted from our corpus in the entertainment domain.

[11] Without generating a complete parse tree for each sentence. Such partial parsing has the advantages of greater speed and robustness.

5.4. Multi-Functionality.

Even though much of the benefits mentioned hitherto apply to information extraction systems, these are not exclusive to this kind of NLP applications. Since the proposed annotation model adds overt linguistic information to any kind of document, it then can be used for a wide range of purposes that require a semantic analysis or processing (i.e. machine-aided translation, information retrieval, etc.).

6. Conclusions.

We have seen that, even though AI researchers are devoting many efforts to finding an optimal model for the semantic annotation of web pages, the decades of work and the results obtained in the field of *Corpus Linguistics* on corpus annotation have been, somehow, neglected, especially in levels different from the semantic. We have seen also that these other linguistic levels carry some semantic information, which can help a computer understand Semantic Web pages. This paper has introduced the different linguistic levels a document can be annotated at and 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–conjugating semantic annotation models from AI and the annotations proposed for every linguistic level from Corpus Linguistics.

The integration of these two approaches (Corpus Linguistics and AI) in the different levels of annotation aforementioned entails many advantages for language engineering and AI applications. First of all, language resources will be more reusable: many of the projects involving the use of semantically annotated (web) documents must also parse to some extent the information and, prior to that, must determine the grammatical category associated to every word in the document. Introducing the annotation of these two levels into the document, hence re-using one of the tools already developed for this purpose, prevents this whole process of document text tokenisation and parsing or chunking from being unnecessarily repeated each time the document is processed (reusing the annotation). Since parsing, for example, is a high time-consuming task, we can have an additional advantage, that is, reducing our overall Semantic Web page processing time. The second main advantage is that the meaning of a page with explicit semantic annotation can be reinforced by the meaning contribution provided by all of the

linguistic levels; semantic analysis can also benefit from the invaluable work done so far on the development of ontologies as conceptual and consensual models.

However, the main disadvantage lies in the limitations imposed by current technologies: the process of obtaining automatically compact, readable and verifiable pages is quite a hard task to be fully specified and delimited, but the work being done in our laboratory is trying to bring some light upon it.

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 Óscar Corcho, Socorro Bernardos and Mariano Fernández for their help with the ontological aspects of this paper.

References.

Aguado, G., Pareja-Lora, A. (2000) A competition model for the generation of complementation patterns in machine translation. *International Journal of Translation*, Vol. 12, N° 1-2, Jan-Dec 2000. Bahri Publications. New Delhi, INDIA.

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

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.

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.

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

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

CES (1999) http://www.cs.vassar.edu/CES/

COHSE (2002) http://cohse.semanticweb.org/

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

Dik, S.C. (1989) *The Theory of Functional Grammar.* Dordrecht: Foris Publications.

EAGLES (1996a) *EAGLES: Text Corpora Working Group Reading Guide.* EAGLES Document EAG-TCWG-FR-2.

EAGLES (1996b) *EAGLES: Recommendations for the Morphosyntactic Annotation of Corpora.* EAGLES Document EAG--TCWG—MAC/R.

EAGLES (1996c) *EAGLES: Recommendations for the Syntactic Annotation of Corpora.* EAGLES Document EAG--TCWG—SASG/1.8.

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

Garside R., Fligelstone, S., Botley, S. (1997) Discourse Annotation: Anaphoric Relations in Corpora. In Garside R., Leech, G., McEnery, A. M. (eds.) *Corpus Annotation: Linguistic Information from Computer Text Corpora.* Longman, London. 1997.

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

Halliday, M. and Hasan, R. (1976) *Cohesion in English.* London: Longman.

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

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

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

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

Kietz, J-U., Maedche, A., Volz, R. (2000) A Method for Semi-Automatic Ontology Acquisition from a Corporate Intranet. In Proceedings of the EKAW'00 Workshop on Ontologies and Text. Juan-Les-Pines, France. October, 2000.

Kilgarriff, A. (1998) SENSEVAL: An Exercise in Evaluating Word Sense Disambiguation Programs. In Proceedings of LREC, 581–588. Granada, Spain. ftp://ftp.itri.bton.ac.uk/reports/ITRI-98-09.ps.gz

Kilgarriff, A. & Rosenzweig, J. (2000) English SENSEVAL: Report and Results. In Proceedings of LREC. Athens, Greece. ftp://ftp.itri.bton.ac.uk/reports/ITRI-00-25.ps.gz

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

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.

Leech, G. (1997b) Grammatical tagging. In Garside R., Leech, G., McEnery, A. M. (eds.) *Corpus Annotation: Linguistic Information from Computer Text Corpora*. London: Longman.

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

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

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

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

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.

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

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

Pino, M. & Santalla, P. (1996) http://www.cica.es/sepln96/sepln96.html

Pulman, S. G. (1995) http://cslu.cse.ogi.edu/HLTsurvey/ch3node7.html#SECTION35

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

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

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.

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

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

Vargas-Vera, M., Motta, E., Domingue, J., Shum, S. B., Lanzoni, M. (2000) 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.

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.