RDF(S)/XML LINGUISTIC ANNOTATION OF SEMANTIC WEB PAGES

Guadalupe Aguado de Cea
DLACT[1]
Fac. Informática, UPM
Madrid, Spain, 28660
lupe@fi.upm.es

Inmaculada Álvarez-de-Mon
DLACT
E.U.I.T. Informáticos, UPM
Madrid, Spain, 28031
ialvarez@euitt.upm.es

Antonio Pareja-Lora
DSIP[2]
Fac. Informática, UCM
Madrid, Spain, 28040
apareja@sip.ucm.es

Rosario Plaza-Arteche
DLACT
Fac. Informática, UPM
Madrid, Spain, 28660
rplaza@fi.upm.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 (RDF(S)/XML) is being developed to fit the needs of both communities; the present paper focuses on its semantic level.

INTRODUCTION

All of us are by now accustomed 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. Due to the astonishing growth of Internet use, new technologies emerged and, with them, machine-aided web page generation appeared. Up to that point, the structure and the edition of these pages fitted only human needs – and this, only to some extent. 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.

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, the above mentioned tasks – relevant information access, extraction and interpretation – cannot be wholly performed by computers yet. Hence, the goal of enabling computers to understand the meaning (the semantics) of written texts and web pages to make it explicit to computers is gaining a growing relevance. That is the main pillar sustaining the development of what we understand by *Semantic Web:* "the conceptual structuring of the web in an explicit machine-readable way" (Berners-Lee et al., 1999). In this context, the *semantic annotation of texts* makes meaning explicit, and has become a key topic. Thus, great efforts are being devoted to the design and application of models and formalisms for the semantic annotation of web pages to make these documents more machine-readable.

Following the guidelines of the Semantic Web initiative, much research has already been carried out by AI researchers on the semantic annotation of web pages (Luke et al., 2000), (Benjamins et al., 1999), (Motta et al., 1999), (Staab et al., 2000). However, these researchers have neglected, somehow, the decades of work and 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 also add some semantic information

[1] Dept. of Languages Applied to Science and Technology.
[2] Dept. of Computer Systems and Programming.

and help a computer understand a text or, in our case, web pages.

This paper will show the results of our research on how linguistic annotation can help computers understand the text contained in a document – a Semantic Web document, for example. Special efforts are 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. As stated in this paper, far from being irreconcilable, they are more than close and may be considered complementary.

This paper is organised as follows: firstly, an introduction to the state of the art in text semantic annotation in corpus linguistics is presented (section 1). Secondly, in section 2, some brief notes on the use of ontologies in semantic annotation is sketched. Thirdly, in section 3, 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 is analysed afterwards – section 4 – and some conclusions are stated – section 5 –, followed by the acknowledgments section and, finally, the references.

1. 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, 1997). Following McEnery & Wilson (2001), **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, 1997). 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. In Leech (1997), a list of the different (possible) **levels of linguistic annotation** can be found. As Leech himself states, for the time being, no corpus includes all of them, but only two or, at most, three of them. Some of them were only in their first stage of conception at the time of writing his paper. A smaller but more realistic list of annotation levels is included in EAGLES (1996a) namely: lemma, morpho-syntactic, syntactic, semantic and discourse annotation. Standard recommendations on morpho-syntactic and syntactic annotation of corpora can be found in (EAGLES, 1996a) and (EAGLES, 1996b). A complementary list of general criteria that should be considered when elaborating an annotation scheme can be found in one of the results of the EAGLES project work, the **Corpus Encoding Standard** (CES, 2000) which are being taken into account in the elaboration of our model (Aguado de Cea, 2002). With respect to the previous and well-known standardization initiative, TEI[3], all these works mentioned are TEI-compliant. Thus, for the sake of brevity, we will focus on semantic annotation henceforth.

As asserted in McEnery & Wilson (2001), two broad **types of semantic annotation** may be identified:

A. The *marking of semantic relationships between items in the text* (for example, the agents or patients of particular actions). This type of annotation has scarcely begun to be applied.
B. The *marking of semantic features of words in a text*, essentially the annotation of word senses in one form or another. This trend has quite a longer history but 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 enunciated, a set of reference criteria has been proposed by Schmidt and

[3] http://etext.virginia.edu/TEI.html

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

mentioned in Wilson & Thomas (1997) for choosing or devising a corpus semantic field[5] annotation system. These criteria can be summarized as follows[6]:

1. It should make sense in linguistic or psycholinguistic terms.
2. It should be able to account exhaustively for the vocabulary in the corpus, not just for a part of it.
3. It should be sufficiently flexible.
4. It should operate at an appropriate level of granularity (or delicacy of detail).
5. It should, where appropriate, possess a hierarchical structure.
6. It should conform to a standard, if one exists[7].

2. ONTOLOGIES AND SEMANTIC WEB ANNOTATIONS.

AI researchers have found in *ontologies* (Gruber, 1993), (Guarino et al., 1995), (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 concepts 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. The influential WordNet and EuroWordNet (Fellbaum, 2001) ontologies should be mentioned as valuable resources for this purpose. Many systems and projects have been developed towards this aim hitherto: SHOE (Luke et al., 2000) proposes HTML page semantic annotation with a Horn clause-based language also called SHOE; the (KA)2 initiative (Benjamins et al., 1999) seeks to annotate HTML documents with ontological information, taking Knowledge Acquisition Community ontologies as a basis; PlanetOnto (Motta et al., 1999) aims at automatically annotating the HTML news pages of an organisation by means of the information obtained from an event-ontology based knowledge base; finally, within the Semantic Community Web Portals project (Staab et al., 2000) an ontology-based architecture for editing and maintaining web portals in an easier way is being developed. Besides, a number of 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).

3. INTEGRATION OF PARADIGMS: AN EXAMPLE

The model here shown, *OntoTag*, is developed within *ContentWeb*, a Ministry funded project, which aims at creating 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. Besides, a prototype in the entertainment domain will be developed. *ContentWeb* objectives can be found in (Aguado de Cea, 2002).

Within the elaboration of *OntoTag*, a first exploration phase has been performed. A short example of this first phase is presented next. It has been implemented in RDF(S), but an XML version was also developed and the possibility of using any other language has *a priori* not been discarded. In the annotation example given below, two different morpho-syntactic tools were applied: Conexor (Conexor, 2002) and MBT (MBT, 2002). Some other tools are being evaluated for further use and the XML and RDF(S) annotation tools and wrappers are being designed at the moment.

---

[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).

[6] For a more detailed explanation, see (Aguado de Cea, 2002).

[7] 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).

[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&amp;gt;</constr:synfunction>
</morphAnnot:ConstrAnnot>
```

Figure 1: Morpho-Syntactic Annotation Excerpt.

3.1. RDF(S) EXAMPLE DESCRIPTION

In Figure 1, Figure 2and Figure 3, we can see the annotation of the following Spanish sentence in the first three levels "*Tras cinco años de espera y después de muchas habladurías, llega a nuestras pantallas la película más esperada de los últimos tiempos."*[9]

In the morpho-syntactic level (Figure 1) every word or lexical token is given a different Uniform Resource Identifier (URI henceforth) and three possible categorisations are included, according to the three different tagsets and systems we want to evaluate. Each tagset has been assigned a different class in the morphAnnot namespace: *TradAnnot* (CRATER tagset), *MBTAnnot* (MBT tagset) and *ConstrAnnot* (Constraint Grammar – CONEXOR FDG tagset). For the sake of space saving, 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">*). Again for the sake of space saving, just 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 annotated with semantic references to the concepts, attributes and relationships determined by our (domain) ontology, implemented in DAML+OIL.

```
<!-- 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 Excerpt.

[9] After five years of expectation and gossiping, here comes the most expected film for the time being.

```
<!-- 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 Excerpt.

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.

3.2. THE XML DATA MODEL

In our **XML data model**, every token from the raw text is labelled with a <*Word*> tag and a RDF URI specified by the attribute *rdf:ID*. Immediately after, nested, a <*surface_form*> tag will be inserted, introducing the token as it appeared in the source text; then come the morpho-syntactic, syntactic and semantic annotations for this token.

The tagset associated to our **morpho-syntactic XML data model** (namespace *pos*) includes the union of those three others defined for CRATER[10] (a Spanish POS tagset, TEI and EAGLES conformant, applied also in *SonIsa*, the tagger developed in our laboratory), MBT (a web-based tagger, also EAGLES conformant) and the web-based version of CONEXOR FDG parser morpho-syntactic part. The tool that produced a particular annotation is tagged with <*Traditional*>, <*MBT*> and <*Constraint*>, respectively. Lemma information is annotated by means of an attribute *lemma*, associated to the tag of the namespace *pos*.

The **syntactic** counterpart of our **XML data model** (namespace *syn*) contains, in a TEI conformant manner, only the syntactic information given by FDG at the moment (more tags may be added as the model is refined). This syntactic information covers EAGLES syntactic layers (c) and (d): showing dependency relations and indicating functional labels. Thus, the attributes defined at this level are: *dependent_on*, which shows the token on which the present one depends (via its *rdf:ID*); *dependency*, which describes the kind of dependency between them both and *surface_syn_tag*, which denotes the surface syntactic function of the token in a Constraint Grammar approach. We are now studying the best way to cover EAGLES syntactic layers (a) and (b) – bracketing and labelling of segments – from a Constraint Grammar perspective, not developed in the EAGLES syntactic guidelines aforementioned.

[10] http://arxiv.org/ps/cmp-lg/9406023

The **semantic** counterpart of our **XML data model** (namespace *sem*) is ontology-based and defined by means of the tags given in the DAML+OIL implementation of our domain ontologies.

4. ADVANTAGES OF THE INTEGRATED MODEL

As shown in the previous section example, 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, not restricted to the semantic level, below analysed. A particular subsection is dedicated to multi-functionality

4.1. 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.

4.1.1. Regarding ontologies from a linguistic point of view

Taking a closer view to sections 1 and 2, and comparing the proposals from both Corpus Linguistics and AI, we find out 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 former is related to the capacity of an ontology to grow horizontally (in breadth) and the latter to the capacity of an ontology to grow vertically (in depth or in specification). Hence, the end user can decide the level of specificity needed. 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 models and formalisms or, as criteria (6) requires, at least they lay a framework of properties and axioms (principles) and major categories that can be modified to some extent to fit individual needs. 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; thus, an ontology-based approach should also make sense in linguistic terms.

4.1.2. 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 statement in section 1 that says, "there is no universal agreement in semantics about which features of words should be annotated" or on that other statement in Schmidt's criterion 1, in the same section, that says, "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, 88), 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.

4.2. MULTI-FUNCTIONALITY

The need for (shallow) parsing in 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. A chunker[12] called Marmot is included in the annotation process presented in the

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

[12] A chunker is a natural language (pre)processing tool that separates and segments sentences into its subconstituents, i.e. noun, verb and prepositional phrases, etc.

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

5. 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. This paper shows 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.

The integration of these two approaches (Corpus Linguistics and AI) 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 somehow 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: obtaining automatically compact, readable and verifiable pages is a task hard to be fully specified and delimited, but the work being done in our laboratory tries 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

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

Aguado de Cea, G., Álvarez de Mon, I., Gómez-Pérez, A., Pareja-Lora, A., Plaza-Arteche, R. (2002) *A Semantic Web Page Linguistic Annotation Model*. "AAAI 2002 Workshop: Semantic Web Meets Language Resources". Edmonton, Alberta, Canada. *(To appear)*

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, pp. 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*. Harper. San Francisco.

Bray, T., Paoli, J., Sperberg, C. (1998) http://www.w3.org/TR/REC-xml.

Brickley, D., Guha, R.V. (2000) http://www.w3.org/TR/PR-rdf-schema.

CES (2000) http://www.cs.vassar.edu/CES/

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

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

EAGLES (1996a) ftp://ftp.ilc.pi.cnr.it/pub/eagles/corpora/annotate.ps.gz

EAGLES (1996b) ftp://ftp.ilc.pi.cnr.it/pub/eagles/corpora/sasg1.ps.gz

EAGLES (1999) http://www.ilc.pi.cnr.it/EAGLES/EAGLESLE.PDF

Fellbaum, C., Palmer, M., Trang Dang, H., Delfs, L., Wolff, S. (2001) *Manual and Automatic Semantic Annotation with WordNet*. In "Proceedings of the NAACL Workshop on WordNet and Other Lexical Resources: Applications, Customizations". Carnegie Mellon University, Pittsburg, PA.

Gruber, R. (1993) *A Translation Approach To Portable Ontology Specification.* Knowledge Acquisition, 5, pp. 199–220.

Guarino, N., Giaretta, P. (1995) *Ontologies and Knowledge Bases: Towards a Terminological Clarification.* In "Towards Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing", N. Mars, ed., IOS Press, Amsterdam, pp. 25–32.

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", Springer-Verlag, Berlin, Germany, pp. 1–16.

Horrocks, I., Van Harmelen, F. (2001) http://www.daml.org/2000/12/reference.html

Karp, R., Chaudhri, V., Thomere, J. (1999) 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.

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

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

Lassila, O., Swick, R. (1999) http://www.w3.org/TR/PR-rdf-syntax

Leech, G. (1997) *Introducing corpus annotation.* In "Corpus Annotation: Linguistic Information from Computer Text Corpora", R. Garside, G. Leech & A. M. McEnery, ed., Longman, London.

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

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

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

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

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) http://crl.nmsu.edu/Staff.pages/Technical/sergei/book/index-book.html

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

RosettaNet (2002) 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), pp 161-197.

UNSPSC (2002) http://www.unspsc.org/

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.

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