Lexical Semantics in the Age of the Semantic Web

Paul Buitelaar

DFKI GmbH, Language Technology Department
Stuhlsatzenhausweg 3, D-66123 Saarbruecken, Germany
paulb@dfki.de

Abstract

Lexical semantics is the study of word meaning. The semantic web is a vision of what the web could be if it would foremost consist of knowledge (structured data) rather than text or other unstructured data as it is today. This talk is about the future of word meaning if the semantic web becomes a reality. First, I will therefore briefly clarify what the semantic web vision consists of, followed by a sketch of lexical semantics. Finally, I will speculate on how the inherent semantic standardization process of the semantic web could have a dramatic influence on the study and use of word meaning.

1. Introduction

Lexical semantics is the study of word meaning. The semantic web is a vision of what the web could be if it would foremost consist of knowledge (structured data) rather than text or other unstructured data as it is today. This talk is about the future of word meaning if the semantic web becomes a reality. First, I will therefore briefly clarify what the semantic web vision consists of, followed by a sketch of lexical semantics. Finally, I will speculate on how the inherent semantic standardization process of the semantic web could have a dramatic influence on the study and use of word meaning.

2. The Semantic Web

2.1. Vision

In (Berners-Lee et al., 2001) Tim Berners-Lee and his co-authors sketched a vision on the future of the world wide web, in which all knowledge is encoded in a formal way in order to let intelligent agents provide services to their human 'masters' in an autonomous way.

As illustrated in Figure 1, this entails the definition of formal, web-based *ontologies* to express the knowledge that is understood by humans as well as agents, and *knowledge markup* of (textual, multimedia) documents and databases using these ontologies. Knowledge markup is an elaboration of so-called *metadata* as currently defined and in use for a restricted set of applications, e.g. the Dublin Core set of bibliographical metadata such as 'title, 'author', etc. (http://dublincore.org/). It is to be expected that over the next decades the knowledge structures of many more such applications will be formally encoded in web-based ontologies. Specifically in the context of e-business this will become apparent, as companies (or rather integrated sections of industry) will need a common and explicit understanding of their products and services in order to allow for an automatic commercial exchange by artificial agents.

2.2. Implementation

The definition of web-based knowledge representation languages is currently an active field of study, which has led to a number of proposals and emerging standards. Foremost among these are RDF Schema (http://www.w3.org/TR/rdf-schema/) and DAML+OIL (http://www.daml.org/2001/03/daml+oil-index), the latter of which is defined on top of the other. Besides these, also XML Schema (http://www.w3.org/XML/Schema) and Topic Maps (http://www.topicmaps.org/xtm/1.0/) are sometimes seen as a knowledge representation language.

In Figure 2 an overview is given of some important aspects of the XML/RDF family of knowledge markup languages (overview based on (Gil and Ratnaker, 2001)). From a syntactic point of view, RDF is written in XML, whereas DAML+OIL is written in RDF. On the semantic side, ontologies written in XML Schema, RDF Schema or DAML+OIL are all based on the notion of a namespace, which defines the interpretation context of any XML, RDF or DAML+OIL expression.

For instance, defining the following XML statement to be in the 'JOBS' namespace ensures that the job of John Smith as a systems-analyst is interpreted exactly as defined in this particular ontology.

```
<xmnls:jobs="http://www.jobs.org/daml+oil-jobs#">

<jobs:systems-analyst>John Smith</jobs:systems-analyst>, a senior systems analyst with IBM, concluded that…
```

In this way, a semantic web agent will be able to identify John Smith as a systems-analyst and look up additional knowledge on this concept in the daml+oil-jobs ontology, which it can access in a distributed fashion at the indicated namespace address.

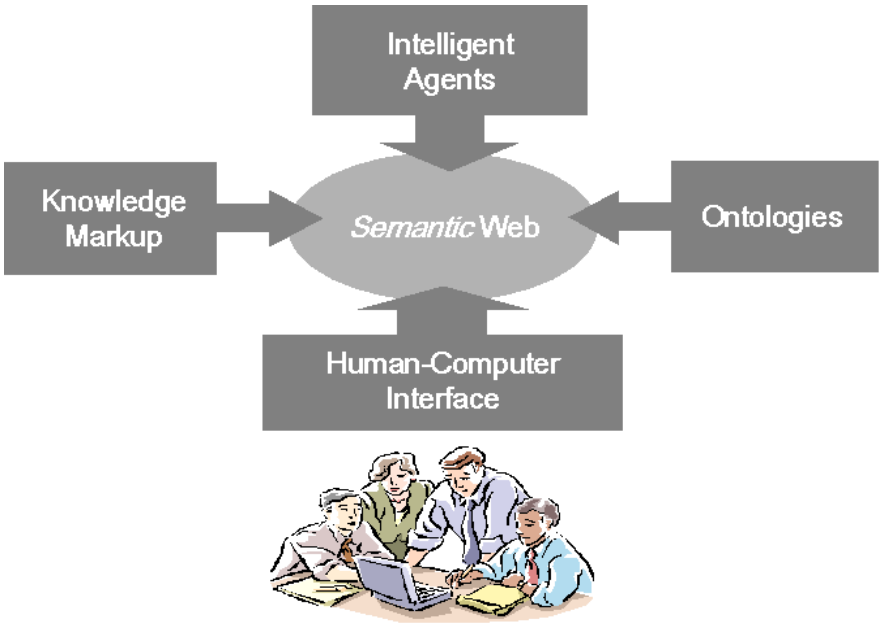

Figure 1: The Semantic Web Vision.

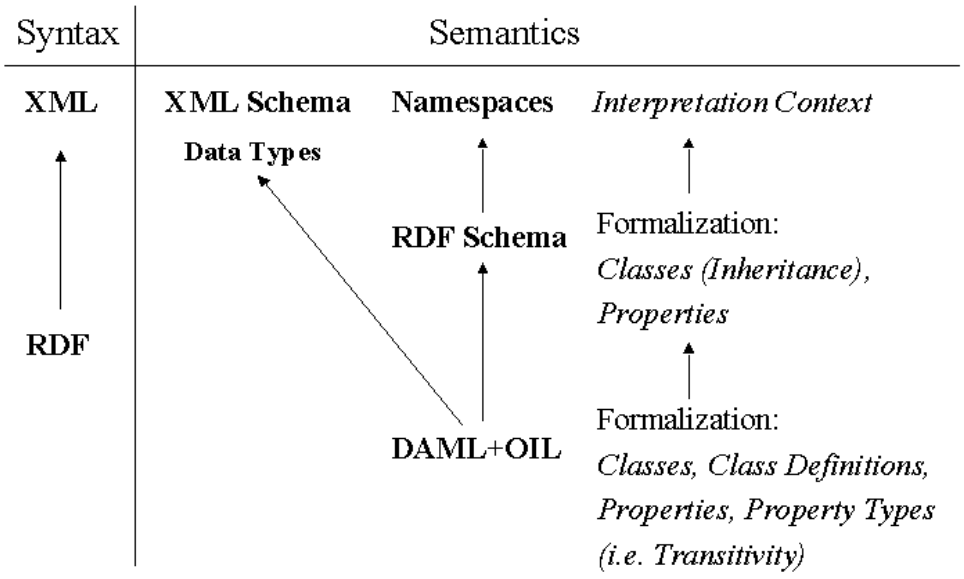

Figure 2: XML/RDF Based Knowledge Markup Languages.

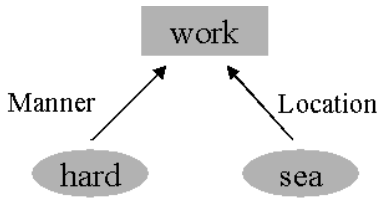

Figure 3: Dependency Structure of the Phrase "hard work at sea".

3. Lexical Semantics: A Sketch

In order to determine the meaning of a word we may look at its context. For instance, word combinations like `hard` and `work` will occur together more often than they individually occur with other words. Such combinations are called *collocations*, which express a simple level of lexical semantic information. A more detailed account of word meaning will be found by analysing the *dependency structure* of those phrases and sentences in which a particular word occurs. For instance, the phrase `hard work at sea` has a structure depicted in Fig. 3.

We can use this analysis to encode some aspects of the lexical semantics of the word work:

work [modifiers
[manner: hard,
location : sea]]

However, what we are missing in this representation is the notion of class, expressing a generalization over a group of words with identical or similar meaning. We can construct such classes by checking for the possibility of substitution. For instance, in the example at hand we can substitute the following words with others that have a similar meaning:

hard work at sea
nice job at sea
nice job on land

We can use this information to encode further, class-based aspects of the lexical semantics of the word work:

work [class : work, job,…
modifiers
[manner: [class : hard, nice,…]
location: [class : sea, land,…]]]

Often, however, we can also substitute context words with others that have a slightly different meaning. For instance, we can substitute some of the words in the context of work also as follows:

beautiful work on paper
beautiful painting on paper
colourful painting on canvas

On the basis of these examples we can now introduce a further class for the word work with corresponding lexical semantic structure:

work [[class : work, job,…
modifiers
[manner: [class : hard, nice,…]
location : [class : sea, land,…]]]
` [class : work, painting,…
modifiers
[manner:[class : beautiful, colourful,…]
medium: [class : paper, canvas,…]]]]

Obviously, this particular interpretation of the word work is connected to its use in the art world. Therefore, in order to identify the validity of a particular interpretation in the context of a corresponding domain, we may introduce also a domain indication in the lexical semantic structure:

work [[class : work, job,…
domain : general
modifiers
[manner: [class : hard, nice,…]
location : [class : sea, land,…]]]
[class : work, painting,…
domain : art_world
modifiers
[manner: [class : beautiful, colourful]
medium : [class : paper, canvas,…]]]]

4. Lexical Semantics on the Semantic Web

4.1. Example

On the semantic web, lexical semantics will be encoded in ontologies that are written in languages such as RDF Schema, DAML+OIL, or Topic Maps. For instance, the lexical semantic structure of work as defined in the previous section could be represented in DAML+OIL as follows:

```
<rdf:RDF
  xmnls:rdf  = "http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmnls:rdfs = "http://www.w3.org/2000/01/rdf-schema#"
  xmnls:xsd  = "http://www.daml.org/2000/10/XMLSchema#"
  xmnls:daml = "http://www.daml.org/2001/03/daml+oil#"
  xmnls:art  = "http://www.art-world.org/art-world#"
>

<daml:Ontology rdf:about="Concepts in the Art World">
  <daml:imports
rdf:resources="http://www.daml.org/2001/03/daml+oil#">
</daml:Ontology>

<daml:Class rdf:ID="art-world.01">
  <rdfs:label>art-world.01</rdfs:label>
  <rdfs:subClassOf      rdf:resource="http://www.art-world.org/art-
world.00#"/>
</daml:Class>

<art-world.01 rdf:ID="work"/>
<art-world.01 rdf:ID="painting"/>

<daml:Class rdf:ID="art-world.02"/>

<art-world.02 rdf:ID="beautiful"/>
<art-world.02 rdf:ID="colourful"/>

<daml:Class rdf:ID="art-world.03"/>

<art-world.03 rdf:ID="paper"/>
<art-world.03 rdf:ID="canvas"/>

<daml:ObjectProperty rdf:ID="manner">
  <rdfs:range rdf:resource="#art-world.02"/>
  <rdfs:domain rdf:resource="#art-world.01"/>
</daml:ObjectProperty >

<daml:ObjectProperty rdf:ID="medium">
  <rdfs:range rdf:resource="#art-world.03"/>
  <rdfs:domain rdf:resource="#art-world.01"/>
</daml:ObjectProperty >

</rdf:RDF>
```

This fragment of the 'art-world' ontology defines three classes that are identified by abstract ids (art-world.01 - art-world.03) and two properties (manner, medium) of the class art-world.01 (i.e. work, painting,…).

4.2. Emerging Semantic Standards and Lexical Semantics: Some Speculation

The example presented above shows how communities with a shared interest, such as companies or non-commercial organisations that are active in a particular area, would be able to define concepts that are common to their activities. If in addition also explicit links are made to corresponding lexical items (individual words, but also more complex terms), standards will most likely emerge that stipulate how such communities should use concepts and corresponding language in their organisations and in interaction with intelligent agents on the semantic web. Obviously, such semantic standards will then also influence in a more general way how language is viewed and used. In fact, if we speculate further on the importance of such standardization, an image emerges in which lexical meaning in particular areas will be more and more determined by the most widely used ontologies in those areas. For instance, in the example at hand, an influential 'art-world' ontology could be defined by a large organisation such as the Getty institute, which already compiles a comprehensive thesaurus on art, architecture and related topics (http://www.getty.edu/research/tools/vocabulary/aat/about.html). A formalized, semantic web-based version of this resource could have as an ultimate consequence that anybody who wants to publish anything on art would need to refer to this ontology in order to be widely understood by semantic web users, be they humans or artificial agents.

5. Conclusions

This paper described the influence of developments around the semantic web on the study and use of lexical semantics. Exemplified by a fragment of an 'art world' ontology it is argued that the semantic web will lead to the emergence of (lexical) semantic standards that will become central to communication between humans and intelligent agents when using information available on the semantic web.

Acknowledgements

This research has in part been supported by EC grant IST-2000-29243 for the OntoWeb project.

6. References

Berners-Lee, T., Hendler, J. and Lassila O. (2001). *The Semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities.* Scientific American. May, 2001.

http://www.sciam.com/print_version.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21

http://dublincore.org/

http://www.w3.org/TR/rdf-schema/

http://www.daml.org/2001/03/daml+oil-index

http://www.w3.org/XML/Schema

http://www.topicmaps.org/xtm/1.0/

Gil, Y. and Ratnaker, V. (2001). *A Comparison of (Semantic) Markup Languages.* In: Proceedings of AAAI 2001. http://trellis.semanticweb.org/expect/web/semanticweb/comparison.html

http://www.getty.edu/research/tools/vocabulary/aat/about.html