Linguistic Annotation for the Semantic Web

Paul Buitelaar and Thierry Declerck
DFKI GmbH, Language Technology Department
Stuhlsatzenhausweg 3, D-66123 Saarbruecken, Germany

Abstract. Establishing the semantic web on a large scale implies the widespread annotation of web documents with ontology-based knowledge markup. For this purpose, tools have been developed that allow for semi-automatic annotation of web documents with ontology-based metadata. However, given that a large number of web documents consist either fully or at least partially of free text, language technology tools will be needed to support this authoring process by providing an automatic analysis of the semantic structure of textual documents. In this way, free text documents will become available as semi-structured documents, from which meaningful units can be extracted automatically (*information extraction*) and organized through clustering or classification (*text mining*). Obviously, this is of importance for both knowledge markup and ontology development, i.e. the dynamic adaptation of ontologies to evolving applications and domains. In this paper we present the following linguistic analysis steps that underlie both of these: *morphological analysis, part-of-speech tagging, chunking, dependency structure analysis, semantic tagging*. Examples for each are given in the context of two projects that use linguistic and semantic annotation for the purpose of cross-lingual information retrieval and content-based multimedia access.

1 Introduction

Establishing the semantic web on a large scale implies the widespread annotation of web documents with ontology-based knowledge markup. For this purpose, tools have been developed that allow for semi-automatic annotation of web documents with ontology-based metadata. However, given that a large number of web documents consist either fully or at least partially of free text, language technology tools will be needed to support this authoring process by providing an automatic analysis of the semantic structure of textual documents. In this way, free text documents will become available as semi-structured documents, from which meaningful units can be extracted automatically (*information extraction*) and organized through clustering or classification (*text mining*). Obviously, this is of importance for both knowledge markup and ontology development, i.e. the dynamic adaptation of ontologies to evolving applications and domains.

Information extraction and text mining are handled in more detail in other chapters of this volume. Here we will focus on the following linguistic analysis steps that underlie both of these: *morphological analysis, part-of-speech tagging, chunking, dependency structure analysis, semantic tagging*. Examples for each are given in the context of two projects that use linguistic and semantic annotation for the purpose of cross-lingual information retrieval and content-based multimedia access.

2 Morphological Analysis

Morphological analysis is concerned with the inflectional, derivational, and compounding processes in word formation in order to determine properties such as stem and inflectional information. Together with part-of-speech (PoS) information this process delivers the morpho-syntactic properties of a word. As a crucial pre-processing step, morphological analysis is used in virtually all fields of natural language processing and in applications such as information retrieval[1]. Some well-known systems are `PC-KIMMO` [1], `GERTWOL` [2], Morphix [3], Mmorph [4], ChaSen [5], the Xerox `MLTT` system [6], and `MULTEXT` [7].

Morphological analysis gives information on the stem of a word, possible parts-of-speech (substantive, adjective, verb, etc.), its inflectional properties (information on *gender: male, female, neuter; number: plural, singular; case: nominative, accusative, dative, etc.*) and possible compound analysis (specifically for languages such as German and Dutch). For example, for the German word Häusern (houses) the following morphological information should be analysed:

```
[PoS=N NUM=PL CASE=DAT GEN=NEUT STEM=HAUS]
```

The stem of the word is Haus, with PoS 'noun' (N). The inflectional properties are `NUM` (number) with value plural, `CASE` with value dativ and `GEN` (gender) with value neuter. Similarly, for the compound word Wintergarten (*winter garden*), a morphological analysis system should return the available morpho-syntactic information for each of its compound words: Winter and Garten.

Morphological analysis is based on the existence of an available lexicon for the language under consideration. Since words are very often highly ambiguous with respect to PoS and inflection, some disambiguation steps are needed. Consider for example the German word Gewinn (*profit*), which is associated with:

a verb reading - imperative singular of gewinnen (*to win*):

```
[POS=V FORM=IMP NUM=SG]
```

a noun reading - singular nominative, dative or accusative of Gewinn (*profit*):

```
[POS=N NUM=SG CASE=(Nom, Dat, Acc)]
```

Morphological disambiguation interacts with PoS tagging (see next section) as well as with chunking (see section 4), which is putting words together in one fragment and thus provides an indirect morphological and PoS disambiguation.

3 Part-of-Speech Tagging

Part-of-Speech (PoS) tagging is the process of determining the correct syntactic class (a part-of-speech, e.g. noun, verb, etc.) for a particular word given its current context. For instance, the word `works` in the following sentences will be either a verb or a noun:

```
He  works  the whole day for nothing.
His  works  have all been sold abroad.
```

[1]For example, [8] shows that adding morphological information (lemma) improves accuracy of a German document retrieval task up to 6%.

As illustrated by this example, PoS tagging involves disambiguation between multiple part-of-speech tags, next to guessing of the correct part-of-speech tag for unknown words on the basis of context information. Currently available tools for PoS tagging are based either on rule-based or stochastic methods to disambiguate and to tag unknown words (this overview based on [9]).

3.1 Rule-Based

Rule-based approaches use hand crafted or automatically extracted rules that use contextual information to assign tags to unknown or ambiguous words (see for instance: [10] [11] [12]). For example, such a rule could describe the fact that if a word is preceded by a determiner and followed by a noun, is should be tagged as an adjective. In addition, many rule-based systems use morphological information to aid in the disambiguation process. For instance, a morphological rule could state that a word, which is preceded by a verb and ends on *-ing* should be tagged as a verb.

3.2 Stochastic

Stochastic PoS taggers are based on statistical models, incorporating frequency or probability (see for instance: [13] [14] [15]). A simple stochastic tagger disambiguates words solely on the probability that a word occurs with a particular tag in a given training set. In other words, the most frequent tag in the training set will be the one assigned to an ambiguous instance of that word. A more advanced alternative to this is to calculate the probability of a given sequence of tags (a so-called *n-gram*), i.e. the probability that a tag occurs with the *n* previous tags. The most common algorithm for implementing an n-gram approach is *Viterbi*, a breadth-first search algorithm [16].

4 Chunking

The concept of chunks has been introduced originally in relation with so-called *performance structures* that reflect the intuitive subdivision of sentences as uttered by a speaker. Such structures, which have been experimentally verified, can sometimes be different from purely linguistically motivated constituent analysis of sentences that reflect the *competence* of a speaker [17]. Based on this observation, [18] defines chunks as the non-recursive parts of core phrases, such as nominal, prepositional, adjectival and adverbial phrases, and verb groups.

4.1 Chunk Parsing

Chunk parsing is an important step towards making natural language processing robust, since the goal of chunk parsing is not to deliver a full analysis of sentences, but to extract just the linguistic fragments that can be surely identified as a reflection of the linguistic performance. This parsing method is said to be robust, since it delivers always some linguistic information, whereas full parsers would fail to deliver any (even partial) linguistic information if the whole utterance cannot be completely analysed in accordance with some competence model of the particular language.

4.2 Partial Parsing

The concept of partial parsing is closely related to chunk parsing. Basic chunks are the results of a first analysis that detects linguistic fragments in a very accurate way, since it is based on secure local information. On the base of this analysis, rules for the combination of partial results are defined and this process can go through further cycles, called *cascades* [19]. This ensures a higher accuracy of the analysis, since problems are only then handled when it can be reasonably assumed that enough linguistic information has been generated during previous cascades. However, even if this strategy fails to produce an analysis for the whole sentence, the partial linguistic information gained so far will still be useful for many applications, such as information extraction and text mining.

4.3 Named Entities

Related to chunking is the recognition of so-called *named entities* (names of institutions and companies, date expressions, etc.). The extraction of named entities is mostly based on a strategy that combines look up in gazetteers (lists of companies, cities, etc.) with the definition of regular expression patterns. Named entity recognition can be included as part of the linguistic chunking procedure. So for example, the following sentence fragment:

```
...the secretary-general of the United Nations, Kofi Annan,...
```

will be annotated as a nominal phrase, including two named entities: `United Nations` with named entity class: organization, and `Kofi Annan` with named entity class: person[2].

5 Annotation in MuchMore: PoS, Morphological Analysis, Chunks

In this section we present an example of linguistic annotation as used in the MuchMore project on concept-based, cross-lingual information retrieval [20]. The MuchMore annotation format integrates multiple levels of linguistic analysis in a multi-layered XML-based DTD, which organizes each level as a separate track with options of reference between them via indices [21]. Linguistic annotation in MuchMore is based on *ShProT*, a shallow processing tool that consists of a tokenizer [22], the TnT part-of-speech tagger [23], a morphological analyser based on Mmorph and Chunkie [24] for chunk parsing. In addition, MuchMore also covers semantic tagging of terms and relations using EuroWordNet [25] and UMLS (Unified Medical Language System) [26] as primary semantic resources (see section 8).

5.1 Part-of-Speech

Part-of-Speech tagging is performed by TnT, which is an HMM-based part-of-speech tagger trained on general language corpora (the NEGRA corpus for German [27], the SUSANNE corpus for English [28]). In order to perform in an optimal way, TnT needs to be adapted to a specific domain. Two approaches may be considered: 1. Retrain TnT on an annotated domain-specific corpus, or 2. Update the underlying TnT lexicon.

As part-of-speech annotated medical corpora are difficult to obtain, we decided in the context of MuchMore to extend the existing TnT lexicon with information from a medical

[2] See also [56] [57] [59] for more details on named entity recognition in the context of information extraction.

lexicon. Because of a similar syntax for general language and the medical language of the MuchMore corpus of scientific abstracts, we obtained good results without retraining.

5.2 Morphological Analysis

Morphological analysis is based on a full-form lexicon generated by Mmorph. Each token is looked up for a matching entry that will provide its morphological information. If no valid word form has been matched, the token is analysed as a potential compound. Initial decompounding experiments produced poor results. However, after adapting the existing Mmorph lexicon to the medical domain results rather improved. Adaptation proceeded according to the following two steps. First, the Mmorph lexicon was updated with additional morphological information from medical lexicons for both German and English. This enabled us to avoid incorrect decompositions like:

```
Zoonoses   ->   zoo + nose
Epicillin  ->   epic + ill + in
Endoral    ->   end + oral
```

Secondly, general-language word forms that function as prefixes in the medical domain (e.g. `auto`) were removed from the Mmorph lexicon, avoiding incorrect decompositions like:

```
Autoimmune          ->   auto + immune
Postinflammatory    ->   post + inflammatory
Radiogram           ->   radio + gram
```

5.3 Chunks

Chunk analysis is performed by Chunkie, which is an HMM-based partial parser that is capable of recognizing not only the boundaries, but also the internal structure of simple as well as complex phrases. On the basis of the PoS and morphological information, Chunkie is able to determine noun phrases (NP), adjectival phrases (AP) and prepositional phrases (PP). Similar with TnT, the performance of Chunkie can be improved by adaptation to a specific domain. For this purpose however, a domain-specific tree bank should be available.

5.4 Example

Linguistic annotation of PoS, morphology and chunking may be illustrated with an analysis of the following sentence:

```
Balint syndrom is a combination of symptoms including simultanagnosia,
a disorder of spatial and object-based attention, disturbed spatial
perception and representation, and optic ataxia resulting from bilateral
parieto-occipital lesions.
```

In the MuchMore annotation format, each sentence contains a <text> block that holds the tokens as XML content, and both lemma and part-of-speech information as XML attributes:

```
<text>
   <token id="w1" pos="NN">Balint</token>
   <token id="w2" pos="NN">syndrom</token>
   <token id="w3" pos="VBZ" lemma="be">is</token>
   <token id="w4" pos="DT" lemma="a">a</token>
   <token id="w5" pos="NN" lemma="combination">combination</token>
```

```
    <token id="w6" pos="IN" lemma="of">of</token>
    <token id="w7" pos="NNS" lemma="symptom">symptoms</token>
    ...
    <token id="w20" pos="JJ" lemma="spatial">spatial</token>
    <token id="w21" pos="NN" lemma="perception">perception</token>
    <token id="w22" pos="CC" lemma="and">and</token>
    <token id="w23" pos="NN" lemma="representation">
     representation</token>
    ...
</text>
```

Each phrase is annotated by use of indices over tokens. In the current example an NP is found that covers the tokens w1,w2 (`Balint syndrom`) and a more complex NP that covers the tokens w20-w23 (`spatial perception and representation`):

```
<chunks>
    <chunk id="c1" from="w1" to="w2" type="NP"/>
    <chunk id="c7" from="w20" to="w23" type="NP"/>
</chunks>
```

6 Dependency Structure

In information extraction, which is concerned with the detection of complex relations or events[3], there is a need to analyse the internal dependency structure of sentences or chunks. Dependency structure may be illustrated by the following simple example:

```
computer terminal switch
```

This noun phrase consists of three nouns that have a semantic dependency to each other: *switch* depends on `terminal`, whereas `terminal` depends on `computer`. At the same time, either `terminal switch` depends on `computer`, or `switch` depends on `computer terminal`. In other words, we are either dealing with a 'switch' for a 'computer terminal', or a 'terminal switch' for a 'computer'. Identifying the appropriate dependency structure is a linguistic analysis task that takes into account the information supplied by the shallow processing tasks mentioned before, as well as the semantic class of individual words and phrases (i.e. *semantic tagging*) that will be discussed in more detail in section 8 and 9.

A *dependency structure* consists of two or more linguistic units that immediately dominate each other in a syntax tree. Therefore, shallow processing (i.e. *chunking*) is insufficient since this typically does not consider such structures. There are two main types of dependencies that are relevant for the kind of applications under consideration. On the one hand, so-called *grammatical functions* (like subject and direct object) for each of the linguistic chunks in the sentence and which allows to identify the actors involved in certain events. So for example in the following sentence, the syntactic subject is the constituent `the shot by Christian Ziege`:

```
The shot by Christian Ziege goes over the goal.
```

However, in order to detect the actor of the event 'goal scene', it is also necessary to analyse its internal dependency structure. In linguistic analysis, for this we use the terms *head*, *complements* and *modifiers*, where the head is the dominating node in the syntax tree of a phrase, complements are necessary qualifiers thereof, and modifiers are optional qualifiers.

[3] See for example [56] [57] [59] [60].

In the example above, the prepositional phrase `by Christian Ziege` (containing the named entity `Christian Ziege`) depends on (and *modifies*) the noun `shot`, whereas the nominal phrase `the shot` depends on (and *complements*) the verb `goes`. Only through the detection of this particular dependency structure, a system would be able to extract the semantic and domain specific information that a player was involved in a 'goal scene'[4].

7 Annotation in MUMIS: PoS, Morphological Analysis, Chunks, Named Entities, Grammatical Functions and Dependency Structure

In this section we present linguistic annotation of PoS, morphology, chunks, named entities, grammatical functions and dependency structure as performed in the MUMIS project on content-based multimedia access [58]. MUMIS uses an integrated set of linguistic tools - SCHUG: Shallow and Chunk based Unification Grammar, which implements a rule-based system of cascades [29]. This system will not be described in detail here (for details see the chapter on *Content-based Indexing and Searching of Multimedia Documents* in this volume). The application defined by the MUMIS scenario implies annotation of named entities, grammatical functions and head-modifier structure, in addition to the shallow processing information also covered by MuchMore. SCHUG has adopted the annotation schema developed in this project, which enables easier performance comparison, and also allows for a smooth integration of the various annotation layers provided by these two systems. Therefore, it is possible to include dependency structure information in this annotation format and also to add a new annotation layer that provides information on grammatical functions.

7.1 Morphological Analysis and Part-of-Speech

For morpho-syntactic information, SCHUG integrates parts of the system described in [22] and maps the results to the MuchMore format (see section 5.4 above), as shown in the following example:

```
Industrie, Handel und Dienstleistungen werden in der ersten Liste
aufgefuehrt, wobei die in Klammern gesetzten Zahlen auf die
Mutterfirmen hinweisen.
(Industry, trade and services are shown in the first list, where
the numbers in brackets point to the parent company.)

<text>
   <token id="w1" pos="NN" lemma="Industrie">Industrie</token>
   <token id="w2" pos="PUNCT">,</token>
   <token id="w3" pos="NN" lemma="Handel">Handel</token>
   <token id="w4" pos="CONJ">und</token>
   <token id="w5" pos="NN" lemma1="Dienst" lemma2="Leistung">
    Dienstleistungen </token>
   ...
</text>
```

[4]See for more details on dependency structure, for instance the introduction to parsing strategies in [62]. More elaborated discussions on relevant syntactic theories and on head grammars are given in [63] [64] [65] [66].

7.2 Chunking and Dependency Structure Analysis

SCHUG implements a modular strategy for the recognition of domain specific named entities. For MUMIS in particular, the task is to detect soccer relevant named entities (e.g. `player, team, trainer, referee, time code, specific-event`). This information will be encoded on the chunk annotation level, with an index pointing to the distinct tokens that correspond to each individual named entity.

The chunking procedure of SCHUG consists of a rule-based sequence of cascades, which produces a richer linguistic representation than the MuchMore tools. So, for example, also (complex) verb groups are annotated (chunks are put into square brackets):

```
[NP Industrie, Handel und Dienstleistungen] [VG werden] [PP in der ersten
Liste] [VG aufgefuehrt], wobei [NP die in Klammern gesetzten Zahlen]
[PP auf die Mutterfirmen] [VG hinweisen].
```

The MuchMore annotation format allows for a straightforward mapping of this information in the chunk layer as shown below. Note that now also information on heads, modifiers and complements are represented. In the case of the first NP, which is*coordinated*, all nouns are considered to be heads. In the case of a PP, the head is always the preposition and its complement is always an NP. The internal structure of the complement NP is not given here.

```
<chunks>
    <chunk id="c1" from="w1" to="w5" type="NP" head="w1,w3,w5"/>
    <chunk id="c2" from="w6" to="w6" type="VG"/>
    <chunk id="c3" from="w7" to="w10" type="PP" head="w7"
     complement="w8,w9,w10"/>
    <chunk id="c4" from="w11" to="w1" type="VG"/>
    ....
</chunks>
```

Next, also grammatical functions are annotated. In order to detect these accurately, an analysis of the *clauses* of a sentence is required. Clauses are the subparts of a sentence that correspond to a (possibly complex) semantic unit, each of which contains a main verb with its complements (grammatical functions) and possibly other chunks (modifiers). For the same example this would produce:

```
<clauses>
  <clause id="cl1" from="c1" to="c4" pred_struct="c2 c4" GF_Subj="c1"/>
  <clause id="cl2" from="c6" to="c9" pred_struct="c9" GF_Subj="c6"/>
</clauses>
```

The annotation identifies two clauses, each with a span of several chunks (pointed to by an index), with information on its predicative structure (pred_struct) and grammatical function (e.g. GF_Subj). Predicative structure can be complex as with the first clause, where the predicate corresponds to a discontinuous verb group with a an auxiliary verb - `werden` (to be) and a main verb - `aufführen` (perform).

8 Semantic Tagging

In the semantic web context, web documents are marked up with metadata, using manual annotation with web-based knowledge representation languages such as RDF and DAML+OIL for describing the content of the document. Some ongoing projects along these lines are SHOE [30], COHSE [31], and OntoAnnotate [32], all of which aim at motivating people to richly annotate electronic documents in order to turn them into a machine-understandable format, and at developing and spreading annotation-aware applications such as content-based information presentation and retrieval.

From a somewhat different angle, automatic semantic annotation has developed within language technology in recent years in connection with more integrated tasks like information extraction. Natural language applications, such as information extraction and machine translation, require a certain level of semantic analysis. An important part of this process is semantic tagging: the annotation of each content word with a semantic category. Semantic categories are assigned on the basis of a semantic lexicon like WordNet for English [33] or EuroWordNet, which links words between many European languages through a common inter-lingua of concepts.

Especially from a domain specific point of view, these separate developments now seem to converge on a common goal of relating textual units in documents to information organized in structured ways. In the following sections we will take a closer look at some of the semantic resources that are available and their use in semantic tagging. Special emphasis is given also to the important subtask of sense disambiguation, which is needed if a word or term corresponds to more than one possible semantic class. Finally, semantic tagging of terms (and of relations between terms) is illustrated by an example from the MuchMore annotation.

8.1 Semantic Resources

Semantic knowledge is captured in resources like dictionaries, thesauri, and semantic networks, all of which express, either implicitly or explicitly, an ontology of the world in general or of more specific domains, such as medicine. They can be roughly distinguished into the following three groups:

- Thesauri: Semantic resources that group together similar words or terms according to a standard set of relations, including broader term, narrower term, sibling, etc.
- Semantic Lexicons: Semantic resources that group together words (or more complex lexical items) according to lexical semantic relations like synonymy, hyponymy, meronymy, and antonymy
- Semantic Networks: Semantic resources that group together objects denoted by natural language expressions (terms) according to a set of relations that originate in the nature of the domain of application.

8.1.1 Thesauri

Roget is a thesaurus of English words and phrases. It groups words in synonym categories or concepts. Besides synonyms also antonyms are covered. A sample categorization (for the concept Feeling) is:

```
I.   Words Expressing Abstract Relations
II.  Words Relating To Space
III. Words Relating To Matter
IV.  Words Relating To The Intellectual Faculties;
     Formation and Communication of Ideas
V.   Words Relating To The Voluntary Powers;
     Individual And Intersocial Volition
VI.  Words Relating To The Sentiment and Moral Powers

I. AFFECTIONS IN GENERAL
   Affections.
      Feeling.
        warmth, glow, unction, gusto, vehemence;
        fervor, fervency;
        heartiness, cordiality;
        earnestness, eagerness;
        empressment, gush, ardor, zeal, passion, ...
```

MeSH (Medical Subject Headings) is a thesaurus for indexing articles and books in the medical domain, which may then be used for searching MeSH-indexed databases [34]. MeSH provides for each term a number of term variants that refer to the same concept. It currently includes a vocabulary of over 250,000 terms. The following is a sample entry for the term `gene` library (MH is the term itself, ENTRY are term variants):

```
MH      =   Gene Library
ENTRY   =   Bank, Gene
ENTRY   =   Banks, Gene
ENTRY   =   DNA Libraries
ENTRY   =   Gene Banks
ENTRY   =   Gene Libraries
ENTRY   =   Libraries, DNA
ENTRY   =   Libraries, Gene
ENTRY   =   Library, DNA
ENTRY   =   Library, Gene
```

8.1.2 Semantic Lexicons

WordNet has primarily been designed as a computational account of the human capacity of linguistic categorization. It therefore covers a rather extensive set of semantic classes (called *synsets*), currently over 90.000. Synsets are collections of synonyms, grouping together lexical items according to meaning similarity. For instance, a board and a plank are similar lexical items and can thus be grouped together in the synset: {board, plank}. At the same time, however, board also refers to a group of people, which may be represented by the synset {board, committee}. So, in fact synsets are actually not made up of lexical items, but rather of lexical meanings (i.e. senses). Observe, that synsets define lexical meaning in an implicit way, contrary to using explicit definitions.

Synsets range from the very specific to the very general. Very specific synsets typically cover only a small number of lexical items, while very general ones tend to cover many. The following example for ’tree’ illustrates how in WordNet the hyponymy relation is used. The word ’tree’ has two meanings that roughly correspond to the classes of plants and that of diagrams, each with their own hierarchy of classes that are included in more general superclasses:

```
09396070 tree 0
 09395329 woody_plant 0 ligneous_plant 0
  09378438 vascular_plant 0 tracheophyte 0
   00008864 plant 0 flora 0 plant_life 0
    00002086 life_form 0 organism 0 being 0 living_thing 0
     00001740 entity 0 something 0

10025462 tree 0 tree_diagram 0
 09987563 plane_figure 0 two-dimensional_figure 0
  09987377 figure 0
   00015185 shape 0 form 0
    00018604 attribute 0
     00013018 abstraction 0
```

EuroWordNet is a multilingual semantic lexicon for several European languages and is structured in similar ways to WordNet. Each language specific (Euro)WordNet is linked to all others through the Inter-Lingual-Index (ILI), which is based on WordNet1.5. Via this index the languages are interconnected, so that it is possible to move from a word in one language to similar words in any of the other languages in the EuroWordNet semantic lexicon.

8.1.3 Semantic Networks

UMLS is one of the most extensive semantic resources available. It is based in part on the MeSH thesaurus and is specific to the medical domain. UMLS integrates linguistic, terminological and semantic information in three corresponding parts: the Specialist Lexicon, the Metathesaurus and the Semantic Network.

The Metathesaurus is a multilingual thesaurus that groups term variants together that correspond to the same concept, for instance the following term variants in several languages for the concept `C0019682` (`HIV`):

```
C0019682   ENG   HIV
C0019682   ENG   HTLV-III
C0019682   ENG   Human Immunodeficiency Virus
C0019682   ENG   Virus, Human Immunodeficiency
C0019682   FRE   VIRUS IMMUNODEFICIENCE HUMAINE
C0019682   GER   HIV
C0019682   GER   Humanes T-Zell-lymphotropes Virus Typ III
```

The Semantic Network organises all concepts in the Metathesaurus into 134 semantic types and 54 relations between semantic types. Relations between semantic types are represented in the form of triplets, with two semantic types linked by one or more relations:

```
Pharmacologic Substance   affects       Pathologic Function
Pharmacologic Substance   causes        Pathologic Function
Pharmacologic Substance   complicates   Pathologic Function
Pharmacologic Substance   diagnoses     Pathologic Function
Pharmacologic Substance   prevents      Pathologic Function
Pharmacologic Substance   treats        Pathologic Function
```

CYC is a semantic network of over 1,000,000 manually defined rules that cover a large part of common sense knowledge about the world [35]. For example, CYC knows that trees are usually outdoors, or that people who died stop buying things. Each concept in this semantic network is defined as a constant, which can represent a collection (e.g. the set of all people), an individual object (e.g. a particular person), a word (e.g. the English word `person`), a quantifier (e.g. `there exist`), or a relation (e.g. a predicate, function, slot, attribute). Consider for instance the entry for the predicate #$mother:

```
#$mother :
   (#$mother ANIM FEM)
      isa: #$FamilyRelationSlot #$BinaryPredicate
```

This says that the predicate #$mother takes two arguments, the first of which must be an element of the collection #$Animal, and the second of which must be an element of the collection #$FemaleAnimal.
Further semantic networks used in semantic tagging are: **Mikrokosmos** [36] and **Sensus** [37].

9 Sense Disambiguation

Words mostly have more than one interpretation, or *sense*. If natural language were completely unambiguous, there would be a one-to-one relationship between words and senses. In fact, things are much more complicated, because for most words not even a fixed number of senses can be given. Therefore, only in certain circumstances and depending on what we mean exactly with sense, can we give restricted solutions to the problem of *Word Sense Disambiguation* (WSD.)

9.1 Methods

WSD involves two parts, a semantic lexicon that associates words with sets of possible semantic classes (i.e. *senses*) and a method of associating (annotating, tagging) occurrences of these words with one or more of its senses. The systems and algorithms that have been developed for this cover the full spectrum of methods developed in natural language processing, artificial intelligence and more recently also of machine learning. For our purposes here, we may group these as follows: *Knowledge-based*, *Hybrid*, and *Empirical*.

In knowledge-based approaches, the construction of the tag set (the senses used and their association with word types in the semantic lexicon) and the tagging (disambiguation between possible senses and association of the preferred sense with a given word token) are both supervised. These approaches use small, but deep, handcrafted lexicons to analyse a small number of examples in a non-robust way, that is, the systems can handle only certain input (see for instance: [38] [39] [40] [41]). All of these rely on pre-coded, domain-specific knowledge, which is the heaviest cost factor in work on WSD, yet indispensable (*knowledge acquisition bottleneck*). Typically, handcrafted rules are constructed, according to given examples. There is no automatic training of the system.

In hybrid approaches, the construction of the tag set is supervised, but the training for the tagging can be either supervised or unsupervised. In the first case, corpora that have been manually annotated are used for training. In the second case, corpora that have not been annotated are used for training. These approaches combine hand-crafted knowledge bases with empirical data derived from large corpora. These approaches use large scale, more shallow, hand-crafted lexicons (WordNet, Roget, lexical database versions of standard English dictionaries like LDOCE, OALD, etc.) to analyse text in a robust way, that is, the systems can handle free, naturally occurring text. Most of these systems became possible thanks to technological advances, using large-scale, machine-readable dictionaries [42], thesauri such as Roget's [43], and computational dictionaries like WordNet [44], [45].

In empirical approaches, the construction of the tag set and the training for the tagging are both unsupervised. These approaches use no external knowledge base at all, but instead derive the tag set itself from the corpus as well. This might be called *self-organizing* WSD. It seeks to do without the pre-defined set of alternative senses, inferring them instead by working, as it were, in the opposite direction: A corpus is used to classify words based solely on patterns of occurrence. The resulting clusters are presumed to represent senses. The two stages that this process consists of are (1) clustering the occurrences of a word into a number of categories and (2) assigning a sense to each category. This idea was first discussed in [46]. It can dispense with the sense-labeling step if the results are only used machine-internally. To stress this subtle but important difference, the method is called *Word Sense Discrimination* in [47]. However, a notable problem with such methods is the close dependence of the resulting classification on the training corpus and the choice of clustering granularity.

9.2 Evaluation

A problem with the various methods proposed for WSD is the lack of a standardized evaluation metric. Publications often focus on only a few words. In a special issue on WSD of the journal for *Computational Linguistics* two out of four articles concentrate exclusively on the three words `line`, `serve`, and `hard`, while many other papers use some other small set of words [48]. In hand-tagged corpora, disagreement between human judges invariably introduces considerable noise [49] [50]. Moreover, it is difficult to draw comparisons across domains. These problems motivated the SENSEVAL [51] and SENSEVAL2 [52] competitions, aimed at developing a standardized evaluation metric for WSD systems.

10 Annotation in MuchMore: Terms and Relations

In addition to the annotation of corpora with shallow processing information as discussed in section 2.4, also semantic information is annotated in the context of the MuchMore project. This includes semantic tagging with EuroWordNet synsets as well as semantic tagging with UMLS concepts and relations as specified in the Metathesaurus and Semantic Network.

10.1 UMLS

A major objective of the MuchMore project is to explore techniques for enhancing cross-lingual information retrieval (CLIR) through automatic semantic annotation of domain-specific terms and relations. For this purpose, the publicly available medical language resource UMLS is used. At the level of terms, the following semantic information is used in annotation:

- Concept Unique Identifier (CUI): maps a term to a concept in the Metathesaurus
- Type Unique Identifier (TUI): maps a concept to one or more semantic types in the Semantic Network
- Medical Subject Headings (MeSH id): maps a CUI to one or more MeSH codes

- Preferred Term: a term that is marked as preferred for a given set of terms and a corresponding concept

Semantic relations are currently annotated between semantic types (TUIs) that co-occur within a sentence. This means that we can only annotate relations between items that were previously identified as terms. The semrel element thus refers to the level of UMLS terms by specifying the pair of terms and the type of relation found. Due to the generic nature of semantic types, the number of possible semantic relations specified between them in UMLS can be considerable. However, through term disambiguation and relevance-based selection of relations it is possible to prune them.

10.2 EuroWordNet

In addition to UMLS, terms are annotated with EuroWordNet to compare domain-specific and general language use. We annotate both single- and multi-word EWN terms, whereby each possible sense of a term represents a separate XML element `sense` with the attribute `offset` (EWN code of the sense). For the purpose of cross-lingual information retrieval we limit the annotation with EWN to senses for nouns only.

10.3 Example

For the example sentence of section 5.4, the words w20-w21 (spatial perception) point to the concept `Space Perception` in the UMLS Metathesaurus, which corresponds to the CUI code C0037744 and TUI code T041 (`Mental Process`) and to two MeSH codes. Word w26 (optic) triggered the concept Optics with corresponding CUI, TUI and MeSH codes.

```
<umlsterm id="t7" from="w20" to="w21">
   <concept id="t7.1" cui="C0037744" preferred="Space Perception"
    tui="T041">
       <msh code="F2.463.593.778"/>
       <msh code="F2.463.593.932.869"/>
   </concept>
</umlsterm>
<umlsterm id="t8" from="w26" to="w26">
    <concept id="t8.1" cui="C0029144" preferred="Optics" tui="T090">
       <msh code="H1.671.606"/>
    </concept>
</umlsterm>
```

The UMLS Semantic Network further defines that `Space Perception` is an issue in `Optics` (expressed by the relation: issue_in), which is coded as follows (note that attributes t7.1 and t8.1 point to the UMLS concepts introduced in the example above):

```
<semrel id="r7" term1="t7.1" term2="t8.1" reltype="issue_in"/>
```

In EuroWordNet annotation, word w21 (perception) has the following senses:

```
<ewnterm id="e2" from="w7" to="w7">
   <sense offset="0487490"/>
   <sense offset="3955418"/>
   <sense offset="4002483"/>
</ewnterm>
```

corresponding to the following synsets:

```
0487490 perceiving, perception, sensing
3955418 perception
4002483 percept, perception, perceptual experience
```

We are currently working on a word sense disambiguation module to cut down on ambiguities concerning EuroWordNet senses, based on unsupervised training as described in [53] [54]. Evaluation of the disambiguation module is undertaken as part of the CLIR evaluation task (comparing disambiguated and non-disambiguated versions of the annotated document collection), as well as separately by using a manually tagged lexical sample corpus [55].

11 Annotation in MUMIS: Entities, Relations and Events

The primary goal of the MUMIS project is to generate formal annotation of multimedia documents that allows for content-based indexing of soccer videos. The project provides highly structured domain-specific annotation of relevant entities, relations and events that can be extracted from transcribed audio and video broadcasts of specific soccer games and corresponding on-line textual documents. The linguistic annotation of these documents as provided by the SCHUG tools are associated with domain-specific information that is encoded in a ontology on the soccer domain. The ontology represents in a hierarchical fashion the main events, relations and entities relevant to the domain (see also [29] and the chapter on multimedia annotation in this volume).

One of the main text types that MUMIS is dealing with is the *on-line ticker* (short descriptions of interesting soccer events, each event indicated by a time-code), for example:

```
7. Ein Freistoss von Christian Ziege aus 25 Metern geht ueber das Tor.
(A 25-meter free-kick by Christian Ziege goes over the goal.)
```

In this example, various entities, relations and events that are relevant to the soccer domain occur (e.g. the artefact: `Tor`, person-player: `Ziege` and two events: the first one expressed by a nominalization (corresponding to an np chunk: `free-kick: Freistoss`) and the second one by a verbal phrase (`goal-scene: geht ueber das Tor`). Some relations are not explicitly mentioned, but can still be inferred by the MUMIS system. For example, the team for which `Ziege` is playing can be inferred from the ontological information that a player is part of a team and the instance of this particular team can be extracted from additional texts. So information not present in the text directly can be added by additional information extraction and reasoning (see also [29]).

```
<chunks>
   <chunk id="c1" from="w1" to="w2" type="NP" head="w1"/>
   <chunk id="c2" from="w3" to="w7" type="NP" head="w4"/>
   <chunk id="c3" from="w8" to="w10" type="PP" head="w8"
    complement="w9 w10"/>
   <chunk id="c4" from="w11" to="w11" type="VG"/>
   <chunk id="c5" from="w12" to="w14" type="PP" head="w12"
    complement="w13 w14"/>
</chunks>
```

```
<clauses>
   <clause id="cl1" from="c1" to="c5" pred_struct="c3"
    GF_Subj="c2"/>
</clauses>

<events>
   <event id="e1" clause="cl1" event-name="free-kick">
      <relations>
        <relation id="r1" reltype="player" player-name="Christian
         Ziege" in-team="Germany"/>
        <relation id="r2" reltype="location" loc-from="25-meter"/>
        <relation id="r3" reltype="time" time-val="07:00"/>
      </relations>
   </event>
    <event id="e2" clause="cl1" event-name="goal-scene-fail">
      <relations>
        <relation id="r1" reltype="player" player-name="Christian
         Ziege" in-team="Germany"/>
        <relation id="r2" reltype="location" loc-from="25-meter"
         loc-to="goal-area"/>
        <relation id="r3" reltype="time" time-val="07:00"/>
      </relations>
   </event>
</events>
```

12 Conclusions: From Language to Semantics to Information Structure

In this paper we showed how to connect language with knowledge as encoded in ontologies through an integration of linguistic and semantic annotation. This touches an important point of discussion in linguistic semantics, namely if it should cover knowledge of language or knowledge of the world. In a way, it should cover both, as what is expressed in language tells us about the worlds built up in our memory. It is therefore not so much a question about what constitutes the meaning of a text, but rather what constitutes a particular world about which a particular text is saying something. This, of course, has also always been a central question in much of philosophy of language, as only through language and most importantly, through categorization in language, we can talk about the world, or the worlds that we imagine. Especially also in the context of emerging semantic web technology this is an important observation, as the information structure that is built up by such an entity (i.e. a domain or organization specific semantic web) reflects the world of a particular organization or group of organizations. A particular semantic web then functions as the collective memory of an organization and the corresponding information structure is the image of the world that it has access to.

13 Acknowledgements

This research has in part been supported by EC/NSF grant IST-1999-11438 for the MuchMore project, EC grant IST-1999-10651 for the MUMIS project.
Many thanks to Spela Vintar, Diana Raileanu and Bogdan Sacaleanu (MuchMore), Mihaela Hutanu and Claudia Crispi (MUMIS).

References

[1] Koskenniemi K. Two-level morphology: a general computational model for word-form recognition and production. Publication No. 11. Helsinki: University of Helsinki Department of General Linguistics. 1983.

[2] http://www.lingsoft.fi/doc/gertwol/

[3] Finkler W. and Neumann G. Morphix: A Fast Realization of a Classification-Based Approach to Morphology. Proceedings of the 4th Austrian Artificial Intelligence Conference. 1988.

[4] Petitpierre, D. and Russell, G. MMORPH - The Multext Morphology Program. Multext deliverable report for the task 2.3.1, ISSCO, University of Geneva. 1995.

[5] Matsumoto Y., Kitauchi A., Yamashita T., Hirano Y., Matsuda H., Asahara M. Japanese Morphological Analysis System ChaSen, version 2.0, Manual 2nd edition. 1999. http://chasen.aist-nara.ac.jp/

[6] http://www.xrce.xerox.com/research/mltt/fsnlp/morph.de.html

[7] http://www.issco.unige.ch/projects/MULTEXT.html

[8] Volk M., Vintar S., Buitelaar P., Raileanu D., Sacaleanu B. Semantic Annotation for Concept-Based Cross-Language Medical Information Retrieval. To appear in the International Journal of Medical Informatics.

[9] http://www.georgetown.edu/cball/ling361/tagging_overview.html

[10] Brill E. A simple rule-based part of speech tagger. Proceedings of the Third Annual Conference on Applied Natural Language Processing, ACL. 1992.

[11] Brill E. Unsupervised learning of disambiguation rules for part of speech tagging. Proceedings of the third ACL Workshop on Very Large Corpora. 1995.

[12] Tapanainen, P., Voutilainen, A. Tagging accurately: don't guess if you don't know. Technical Report, Xerox Corporation. 1994. http://www.ling.helsinki.fi/ tapanain/cg/index.html

[13] Cutting D., Kupiec J., Pedersen J., Sibun P. A Practical Part-of-Speech Tagger. In Proceedings of the 3rd conference on Applied Natural Language Processing (ANLP). 1992. ftp://parcftp.xerox.com/pub/tagger/

[14] Schmid H. Probabilistic Part-of-Speech Tagging Using Decision Trees. In International Conference on New Methods in Language Processing. Manchester. 1994. (http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html)

[15] Brants, T. TnT - A Statistical Part-of-Speech Tagger. In: Proceedings of 6th ANLP Conference, Seattle, WA. 2000.

[16] Brill E., Marcus M. Tagging an unfamiliar text with minimal human supervision. ARPA Technical Report. 1993. (ftp://ftp.cs.jhu.edu/pub/brill/Programs/UNSUP_TAGGER_V0.8.tar.gz)

[17] Chomsky N. Aspects of the Theroy of Syntax. The MIT Press, Cambridge, MA, 1965.

[18] Abney S. Chunks and Dependencies: Bringing Processing Evidence to Bear on Syntax. In: Computational Linguistics and the Foundations of Linguistic Theory. CSLI. 1995.

[19] Abney S. Partial Parsing via Finite-State Cascades. Journal of Natural Language Engineering, 2(4): 337-344. 1996.

[20] http://muchmore.dfki.de

[21] Vintar S., Buitelaar P., Ripplinger B., Sacaleanu B., Raileanu D., Prescher D. An Efficient and Flexible Format for Linguistic and Semantic Annotation In: Proceedings of LREC2002 , Las Palmas, Canary Islands - Spain, May 29-31, 2002.

[22] Piskorski J., G. Neumann. An Intelligent Text Extraction and Navigation System. Proceedings of the 6th International Conference on Computer-Assisted Information Retrieval (RIAO). 2000.

[23] http://www.coli.uni-sb.de/ thorsten/tnt/

[24] Skut W. and Brants T. A Maximum Entropy partial parser for unrestricted text. In: Proceedings of the 6th ACL Workshop on Very Large Corpora (WVLC), Montreal. 1998.

[25] Vossen, P. 1997. EuroWordNet: a multilingual database for information retrieval. In: Proceedings of the DELOS workshop on Cross-language Information Retrieval, March 5-7, 1997.

[26] http://umls.nlm.nih.gov

[27] http://www.coli.uni-sb.de/sfb378/negra-corpus/

[28] http://www.cogs.susx.ac.uk/users/geoffs/Rsue.html

[29] Declerck T. A set of tools for integrating linguistic and non-linguistic information. Proceedings of SAAKM 2002, ECAI 2002, Lyon.

[30] Heflin J., Hendler J., and Luke S. SHOE: A Knowledge Representation Language for Internet Applications. Technical Report CS-TR-4078. Department of Computer Science, University of Maryland, 1999.

[31] Bechhofer S. and Goble C. Towards Annotation using DAML+OIL Communications of the ACM, 2000.

[32] Staab S., Maedche A., Handschuh S. An Annotation Framework for the Semantic Web. In The First International Workshop on Multimedia Annotation, Tokyo, Japan, 2001.

[33] Miller, G.A. WordNet: A Lexical Database for English. Communications of the ACM 11. 1995.

[34] http://www.nlm.nih.gov/mesh/meshhome.html

[35] http://www.cyc.com

[36] Kavi Mahesh and Sergei Nirenburg. A situated ontology for practical NLP. In Proceedings of IJCAI-95 Workshop on Basic Ontological Issues in Knowledge Sharing. 1995.

[37] Knight, K., Luk . Building a Large Knowledge Base for Machine Translation. Proceedings of the American Association of Artificial Intelligence Conference AAAI-94. Seattle, WA. 1994.

[38] Small, S.L. Word Expert Parsing: A Theory of Distributed Word-based Natural Language Understanding. Ph.D. thesis, The University of Maryland, Baltimore, MD. 1980.

[39] Small, S.L. Parsing as cooperative distributed inference. In King, M. (ed.): Parsing Natural Language. Academic Press, London. 1983.

[40] Hirst, G. Semantic Interpretation and the Resolution of Ambiguity. Cambridge University Press. 1988.

[41] Adriaens, G., and S.L. Small. Word expert revisited in a cognitive science perspective. In Small, S., G.W. Cottrell, and M.K. Tanenhaus (eds.): Lexical Ambiguity Resolution: Perspectives from Psycholinguistics, Neuropsychology, and Artificial Intelligence. Morgan Kaufmann, San Mateo, CA, pages 13-43. 1988.

[42] Lesk, M.E. Automated sense disambiguation using machine-readable dictionaries: How to tell a pine cone from an ice cone. In Proceedings of the SIGDOC Conference. 1986.

[43] Yarowsky, D. Word-sense disambiguation using statistical models of Roget's categories. In Proceedings of COLING-92, Nantes, France. 1992.

[44] Resnik, P. Using information content to evaluate semantic similarity in a taxonomy. In Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI). 1995.

[45] Ng, H.T., and H.B. Lee. Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach. In Proceedings of ACL96. 1996.

[46] Schtze, H. Context space. In Goldman, R., P. Norvig, E. Charniak, and B. Gale (eds.): Working Notes of the AAAI Fall Symposium on Probabilistic Approaches to Natural Language, AAAI Press, Menlo Park, CA, pages 113-120. 1992.

[47] Schtze, H. Automatic word sense discrimination. Computational Linguistics, 24(1):97-123. 1998.

[48] Ide N., Veronis J. Introduction to the special issue on word sense disambiguation: The state of the art. Computational Linguistics, 24(1):1–40. 1998.

[49] Kilgariff, A. 1998. Gold standard datasets for evaluating word sense disambiguation programs. Computer Speech and Language 12(4), Special Issue on Evaluation.

[50] Vronis, J. 1998. A study of polysemy judgements and inter-annotator agreement. In Programme and advanced papers of the Senseval workshop, Herstmonceux Castle (England), pages 2-4.

[51] Kilgariff, A., and M. Palmer. 2000. Introduction to the special issue on SENSEVAL. Computers and the Humanities 34(1/2):1-13.

[52] http://www.sle.sharp.co.uk/senseval2/

[53] Buitelaar P., Alexandersson J., Jaeger T., Lesch S., Pfleger N., Raileanu D., von den Berg T., Klckner K., Neis H., Schlarb H. An Unsupervised Semantic Tagger Applied to German. In: Proceedings of Recent Advances in NLP (RANLP) , Tzigov Chark, Bulgaria. 2001.

[54] Buitelaar P., Sacaleanu B. Ranking and Selecting Synsets by Domain Relevance. In: Proceedings of WordNet and Other Lexical Resources: Applications, Extensions and Customizations. NAACL 2001 Workshop, Carnegie Mellon University, Pittsburgh. 2001.

[55] Raileanu D., Buitelaar P., Bay J., Vintar S. An Evaluation Corpus for Sense Disambiguation in the Medical Domain. In: Proceedings of LREC2002, Las Palmas, Canary Islands. 2002.

[56] Appelt D.E. An Introduction to Information. AI Communications, 12. 1999.

[57] Cunningham H. Information Extraction: A user Guide. Research Report CS-9907,Department of Computer Science, University of Sheffield. 1999.

[58] Declerck T., Wittenburg P., Cunningham H. The Automatic Generation of Formal Annotations in a Multimedia Indexing and Searching Environment. Proceedings of the Workshop on Human Language Technology and Knowledge Management, ACL-2001.

[59] MUC-7: Seventh Message Understanding Conference. http://www.muc.saic.com/, SAIC Information Extraction. 1998.

[60] Neumann G., Backofen R., Baur J., Becker M., Braun C. An Information Extraction Core System for Real World German Text Processing. Proceedings of the 5th Conference on Applied Natural Language Processing, ANLP-97, 209-216. 1997.

[61] Lappin, S., Shih H-H. A generalized Algorithm for Ellipsis Resolution. Proceedings of the 16th International Conference on Computational Linguistics, COLING-96. 1996.

[62] Hellwig P. NATURAL LANGUAGE PARSERS A ”Course in Cooking. COLING-ACL ’98– Pre-Conference Tutorial. 1998.

[63] Balari S. Information-Based Linguistics and Head-Driven Phrase Structure. In Miguel Filgueiras and Lus Damas and Nelma Moreira and Ana Paula Toms, editor(s), Natural Language Processing. 55-101. Berlin: Springer-Verlag. 1991.

[64] Borsley R. D. Modern Phrase Structure Grammar. Blackwell textbooks in linguistics, number 11. : Blackwell Publishers. 1996.

[65] Borsley R. D. Heads in HPSG. In Greville Corbett and N. Fraser and S. McGlashan, editor(s), Heads in Grammatical Theory. Forthcoming.

[66] Pollard C., Sag. Head-Driven Phrase Structure Grammar. Chicago: University of Chicago Press. 1994.