Project ref. no.	*IST-1999-11438*
Project acronym	**MUCHMORE**
Project full title	Multilingual Concept Hierarchies for Medical Information Organization and Retrieval

Security (distribution level)	*Public*
Contractual date of delivery	*Month 12 (June 2001)*
Actual date of delivery	*Month 17 (November 2001)*
Deliverable number	*D4.1*
Deliverable title	*MUCHMORE Annotation Format*
Type	*PU*
Status & version	*Final*
Number of pages	*17*
WP contributing to the deliverable	*WP4.1*
WP / Task responsible	*DFKI*
Author(s)	*Špela Vintar, Paul Buitelaar, Bogdan Sacaleanu, Diana Raileanu, Detlef Prescher (DFKI); Bärbel Ripplinger (EIT); Ralf Brown (CMU); Jörg Bay, Oktavian Weiser (ZInfo); Eric Gaussier, Hervé Dejean (XRCE); Dominic Widdows (CSLI)*
EC Project Officer	*Yves Paternoster*
Keywords	*Annotation Format, XML, DTD, Morphosyntactic Annotation, Semantic Annotation, Sentence Alignment, UMLS, EuroWordNet*
Abstract (for dissemination)	*This report describes the XML-based annotation format (DTD) that was developed according to the aims and needs of the MUCHMORE project. In order to exploit multiple layers of information, including morphological, syntactic and semantic annotation, it was necessary to develop an encoding format that would offer efficient access to information next to flexibility in adapting it to specific tasks.*

1 Introduction

This report describes the XML-based annotation format (DTD) that was developed according to the aims and needs of the MUCHMORE project. In order to exploit multiple layers of information, including morphological, syntactic and semantic annotation, it was necessary to develop an encoding format that would offer efficient access to information next to flexibility in adapting it to specific tasks.

The following section outlines corpus selection and pre-processing steps. The next section describes the different annotation levels and the resources used for linguistic and semantic annotation, while the last section presents a detailed description of the DTD, in both its structure and elements. Finally, Appendix I gives the DTD itself, while Appendix II presents a detailed example.

2 Corpus Selection and Preparation

The corpus used in the development of the annotation scheme is a parallel corpus of English-German scientific medical abstracts obtained from the Springer Link web site[1]. The corpus consists approximately of 1 million tokens for each language. Abstracts are from 41 medical journals (e.g. Der Nervenarzt, Der Radiologe, etc), each of which constitutes a relatively homogeneous medical sub-domain (e.g. Neurology, Radiology, etc.).

Corpus preparation included removing HTML-tags, removing English segments from German abstracts and vice versa, deleting names of authors, addresses, etc., removing or converting symbols and other non-ASCII elements and producing a clean, plain text version of each abstract, consisting of title (if available), text and keywords (if available).

3 Annotation Levels and Resources

The corpus is annotated with several layers, involving a number of steps in linguistic and semantic analysis. In the context of the MUCHMORE objectives, semantic annotation of medical terms and semantic relations between such terms is primary among these. Nevertheless, a number of additional levels of annotation were deemed to be necessary in order to facilitate various intermediate objectives, e.g. (bilingual) term extraction, relation extraction and sense disambiguation. Additional annotation includes part-of-speech tagging, lemmatizing (morphological analysis), chunking and grammatical relation tagging. Finally, to compare domain specific and more general sense distinctions, terms are also annotated with a multilingual lexical semantic resource (EuroWordNet).

[1] http://link.springer.de/

3.1 Morphosyntactic Annotation

3.1.1 Morphosyntactic Annotation -- DFKI

For morphosyntactic annotation, DFKI uses ShProT, a shallow processing tool that consists of four integrated components: the SPPC tokenizer (..., 199), TnT (Brants, 2000) for part-of-speech tagging, Mmorph (based on Petitpierre and Russell, 1995) for morphological analysis and Chunkie (Skut and Brants, 1998) for phrase recognition.

3.1.1.1 Tokenizing

The first step in processing is to segment the input text into individual sentences, words and other syntactic elements, e.g. hyphenated compounds (*side-effects, short-term, follow-up*, etc.) or abbreviations (*aquos., emulsific., Ungt.*, etc.). It is important that tokenizing is done fairly well, because all other processing steps depend on it. Therefore, the tokenizer needs to be adapted to the medical domain in some respects, for instance in handling abbreviations. To this end, a list of German medical abbreviations is available from ZInfo.

3.1.1.2 Part-of-Speech Tagging

TnT is an HMM-based part-of-speech tagger trained on general language corpora (the NEGRA[2] corpus for German, the SUSANNE[3] corpus for English) . In order to perform in an optimal way, we adapted it to the medical domain. Two approaches were considered: 1. retraining the tagger on an annotated domain-specific corpus; or 2. an update of its underlying lexicon. As part-of-speech annotated medical corpora are difficult to obtain, we decided to adapt the lexicon with information from the UMLS English and German Specialist Lexicons (German lexicon constructed by and available through ZInfo). Because of a similar syntax for general language and the medical language used in our corpus of scientific abstracts, we obtained good results without retraining.

3.1.1.3 Morphological Analysis

The morphological analyser is based on a dumped, full-form Mmorph lexicon. Initial experiments with lemmatisation (including compound analysis) produced poor results on our corpus, particularly in the analysis of medical compounds in German. We therefore decided to update the existing lexicon with new morphological information from the UMLS Specialist Lexicon (for English) and some morphological resources provided by ZInfo (for German).

3.1.1.4 Chunking

The UMLS Metathesaurus provides an extensive inventory of technical terms for the medical domain, but in addition we also extract novel terms from our corpus. As the technical terms within a domain are mainly phrases, such as NPs and PPs, the recognition of these syntactic structures, also called chunks, is a necessary task. The tool that DFKI uses in this process, Chunkie, is an HMM-based partial parser that goes beyond simple bracketing and is capable of recognizing not only the boundaries, but also the internal

[2] The NEGRA corpus is available under http://www.coli.uni-sb.de/sfb378/negra-corpus/
[3] The SUSANNE corpus is available under http://www.cogs.susx.ac.uk/users/geoffs/Rsue.html

structure of simple as well as complex NPs, PPs and APs. Similar with TnT, the performance of Chunkie would improve by adaptation to the medical domain. The only possible approach to this is by retraining on a domain-specific treebank. However, as we are not aware of any existing medical treebank, we decided to use Chunkie as is, that is, trained on general language data.

3.1.2 Morphosyntactic Annotation -- XRCE

The morphosyntactic annotation done by XRCE is based on existing tools (Schiller, 1996). The different operations are: part-of-speech tagging, lemmatization, word decomposition and NP extraction. Lemmatization and word decomposition being two operations improving the quality of the bilingual alignment, they needed to be adapted to the corpus. Using corpus-based heuristics, the coverage of the lemmatization and of the (German) word decomposition has been significantly increased, precision being equal. The NP extraction was also tuned in order to mark-up only NPs relevant from a terminological point-of-view.

3.2 Grammatical Relations

In addition to morphosyntactic annotation, and in preparation of semantic relation extraction and annotation, the corpus will be automatically annotated with grammatical relations, such as verb-argument and adjective-noun relations. Here is an example:

> *Untersucht wurden 30 Patienten, die sich einer elektiven aortokoronaren Bypassoperation unterziehen mussten.*

This German sentence has two main verbs, *Untersuch* and *unterziehen*, where the verb *Untersucht* occurs in a passive construction with one single argument, the subject *30 Patienten*, whereas the verb *unterziehen* occurs in an active construction with three arguments: the subject *30 Patienten*, an object *einer elektiven aortokoronaren Bypassoperation* and an indirect object *sich*. Using the notational definitions PRED1=first predicate, PRED2=second predicate, ACT=active, PAS=passive, SUBJ=subject, OBJ=object, IOBJ=indirect object, the relevant information can be annotated as follows:

> *Untersucht <PRED1:PAS> wurden 30 Patienten <PRED1:SUBJ> <PRED2:SUBJ>, die sich <PRED2:SUBJ> einer elektiven aortokoronaren Bypassoperation <PRED2:IOBJ> unterziehen <PRED2:ACT> mussten.*

Note that the gathered information about grammatical relations were simply added to the lexical heads of the affected verbal and nominal chunks (between brackets: "<" and ">"), yielding the following four lexicalized grammatical relations:

> *"Untersucht" PAS.SUBJ:SUBJ "Patienten"*
> *"unterziehen" ACT.SUBJ*OBJ*IOBJ:SUBJ "Patienten"*
> *"unterziehen" ACT.SUBJ*OBJ*IOBJ:OBJ "sich"*
> *"unterziehen" ACT.SUBJ*OBJ*IOBJ:IOBJ "Bypassoperation"*

The displayed four relations consist of three elements each. The first element is the lexical head of a verb phrase/chunk as occurring in the given sentence, the second

element is its subcat frame, e.g. PAS.SUBJ or ACT.SUBJ*OBJ*IOBJ together with a subcat slot filled by the third element, in general the lexical head of a noun phrase/chunk.

Parsing with higher-order syntactic frameworks (HPSG, LFG, etc.) generally provides analyses with functional information (SUBJ, OBJ, etc.). Moreover, stochastic versions of these parsers might be used on free text to get unambiguous output, i.e. unambiguous functional annotations. Unfortunately, stochastic unification-based parsers are currently not available for our purposes. Therefore, experiments with a (relatively shallow) stochastic approach for identification of lexicalized grammatical relations have been set up at DFKI. Note, that the above displayed format of grammatical relations has been used in this section for illustrative purposes. It can be automatically mapped to the agreed-on format, as defined in Section 4, and especially displayed by the example in Appendix II.

3.3 Semantic Annotation

A major objective of the MUCHMORE project is to explore techniques for enhancing cross-lingual information retrieval through automatic semantic annotation of domain-specific terms and relations. For this purpose, primarily the publicly available medical resource UMLS[4] (Unified Medical Language System) is used besides other available semantic resources in the medical domain. In addition, terms are annotated also with EuroWordNet (Vossen, 1997) to compare domain-specific and general language use.

3.3.1 Semantic Annotation -- DFKI

3.3.1.1 UMLS Terms

UMLS organizes linguistic, terminological and semantic information in three parts: Specialist Lexicon, Metathesaurus and Semantic Network. At the level of terms, the following semantic information is used in annotation:

- **Concept Unique Identifier (CUI):** a code that represents a concept in the Metathesaurus to which terms are mapped

- **Type Unique Identifier (TUI):** a code that represents a semantic type in the Semantic Network; one or more semantic types are mapped to a concept

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

The CUI's and the preferred terms can be taken from the MRCON database of the UMLS Metathesaurus, while the mapping of CUI's to TUI's can be found in the MRSTY database. In the process of identifying terms in the corpus, we decided to filter and normalise the MRCON database according to the following criteria:

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

- Only terms from MeSH2001 (Medical Subject Headings) were used, because this is the only source within the MetaThesaurus that is available for both English and German.
- Term variants with commas were reversed to normal word order (e.g. *Virus, Human Immunodeficiency -> Human Immunodeficiency Virus*).
- Terms longer than five words were filtered out.

The filtered MRCON database was also PoS-tagged and lemmatised to facilitate term matching. The current version of term tagging includes lookup of uni-, bi- and trigrams as well as the identification of part-of-compound terms on the morphological level (specifically in the case of German compounds).

There are two possible kinds of ambiguity at the level of terms: a single term may be assigned several CUI's, and a single CUI may be mapped to several TUI's. Thus, for example, the term *Thrombocytopenia* can be assigned two concept codes, C0040034 and C0740405, and the term *Type I Collagen* with C0041455 can have the semantic type T116 or T123. The disambiguation process (see deliverables for WP5) will reduce these to a single sense (i.e. CUI and/or TUI).

3.3.1.2 UMLS and Novel Semantic Relations

Semantic relations are annotated between semantic types (TUI's) that co-occur within one sentence, using the SRSTRE1 database of the UMLS Semantic Network. The Semantic Network further organises all concepts in the Metathesaurus into 134 semantic types and 54 relations between semantic types. Both, the semantic types and the relations are given textual definitions in the SRDEF database, e.g.:

> RL|T151|**affects**|R3.1|Produces a direct effect on. Implied is the altering or influencing of an existing condition, state, situation, or entity. This includes has a role in, alters, influences, predisposes, catalyzes, stimulates, regulates, depresses, impedes, enhances, contributes to, leads to, and modifies.||||AF|affected_by|

The relations between semantic types are represented in the form of triplets (T121|T147|T046), whereby two semantic types may be linked by several relation types:

> 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|

Annotation of semantic relations is based on prior identification of UMLS concepts and their semantic types, whereupon all possible pairs of semantic types occurring within one sentence are looked up and annotated within the `semrels` layer as references to the `terms` layer.

The most common relations found in the Springer corpus include *associated_with, interacts_with, result_of, affects, location_of, issue_in, degree_of, produces, uses, occurs_in* etc. Also, some relations are more typical of a particular sub-domain than others. For example, *location_of* occurs more frequently in all of the anatomy-based medical sub-domains (sub-corpora: Der Chirurg, Der Pathologe, etc.) but less in psychology or ethics (sub-corpora: Der Psychologe, Ethik in der Medizin, etc.).

Due to the generic nature of semantic types and the ambiguity mentioned before, the number of possible semantic relations derived from two co-occurring terms can be considerable, while their actual relevance for the retrieval task remains questionable. Therefore, further work is underway to integrate disambiguation (WP5) and to identify novel semantic relations (WP7.2), based also on grammatical relation annotation.

3.3.1.3 EuroWordNet Terms

In addition to annotation with UMLS, terms are annotated also with EuroWordNet (Vossen, 1997) to compare domain-specific and general language use. EuroWordNet is a multilingual database for several European languages and is structured in similar ways to the Princeton WordNet (Miller, 1995). Each language specific (Euro)WordNet is linked to all of the others through the so-called 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 database. For our current purposes we use only the German and English parts of EuroWordNet.

All information in (Euro)WordNet is centered around so-called synsets, which are sets of (near-) synonyms. The different senses of a term are therefore simply all the synsets that contain it. As with UMLS terms, disambiguation (see deliverables for WP5) will reduce these to a single sense. A term can be simple (*man*) or complex (*rock_and_roll*). A synset is identified by a unique identifier, called offset. Because meanings between languages cannot be exactly mapped one-on-one, there may be more than one synset within a language that is mapped on the same concept in the ILI. In order to distinguish between these, every synset was given a unique identifier (ID)[5], as shown by the following example:

	Offset - ID	Synset
German	*3824895-1*	*Fingergelenk*
	3824895-2	*Fingerknochen*
	3824895-3	*Knöchel*
English	*3824895*	*knuckle, knuckle joint, metacarpophalangeal joint*

[5] In our case only for German, as the English synsets correspond to the ILI directly.

3.3.2 Semantic Annotation -- Zinfo

3.3.2.1 Xmed

Xmed is a tool for automatic classification (diagnosis and procedures) of medical text. The text is analyzed by TRANSOFT, a thesaurus-based translation-system, which rebuilds every sentence into a normalized format. Every term is replaced by its preferred term (from the thesaurus) with additional information like ICD-10 codes or UMLS-Metathesaurus concepts.

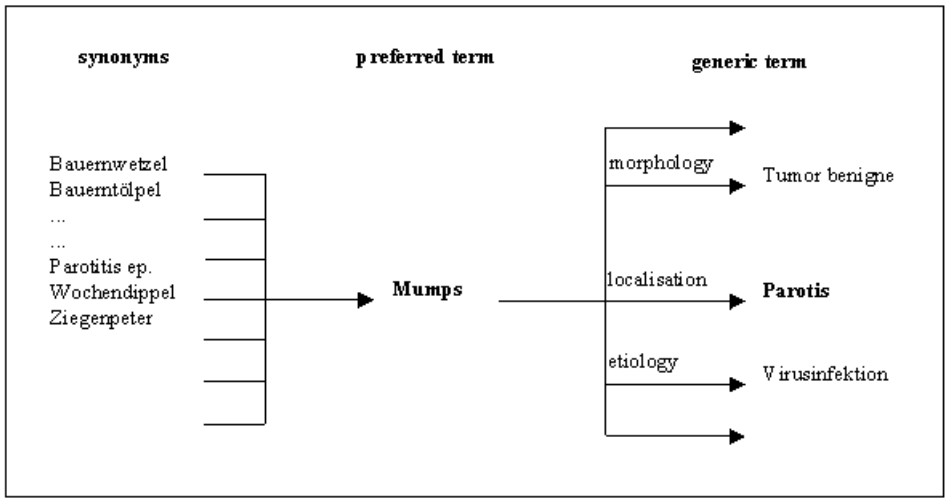


Figure 1: Example "Mumps"

3.3.2.2 AGK-thesaurus

The thesaurus used by Xmed is based on the multi-axial AGK-thesaurus with more then 100.000 terms, which is developed since 1968, first by a GMDS-workgroup and later by W. Giere at ZInfo. AGK is the abbreviation for "**A**rbeits**g**ruppe **K**lartextdokumentation der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie e.V." (GMDS - German Association of Medical Informatics, Biometrics, and Epidemiology). The multi-axial (e.g. etiology, morphology, topography, function) thesaurus contains also semantic relations like "caused" or "temporal". It is translated into English and mapped to UMLS-Metathesaurus concepts.

3.3.2.3 IDT

Based on the ICD-10 (International Statistical Classification of Diseases and Related Health Problems) the German ICD-10 Diagnoses Thesaurus (IDT) links synonyms, spelling variants and permutations of multi-word terms to the appropriate ICD-10 entry and its code. It is translated into English and mapped to UMLS-Metathesaurus concepts.

AGK and IDT are two bilingual thesauri that are not part of the UMLS-Metathesaurus and could therefore be used to enlarge the German part of this resource as used in MUCHMORE. It should be possible to find more terms also through linked UMLS-Metathesaurus concepts.

3.4 Sentence Alignment

The texts are sentence-aligned using the XRCE alignment tool (Eisele, ...). The alignment method is based on the assumption that the given texts represent the same contents in essentially the same order, and it considers deviations in this order only on the level of small groups of sentences. Based on this assumption, the method tries to find the best monotonic alignment, and consists of the following steps:

- monolingual pre-processing (tokenization, lemmatisation, case normalisation, paragraph segmentation)
- score different alignments based on different features (sentence length, and, optionally, cognates and translations derived from existing resources)
- search for the best path through the space of potential alignments

The tool produces stand-off correspondence tables of aligned segments, which are integrated in the XML-encoded corpus by references to corresponding segments in the parallel text.

4 Annotation Format

The annotation task involves combining multiple levels of linguistic and semantic information that are interrelated in various ways. Our aim was to design an annotation format that would encompass all of these layers and adequately represent the relationships between them, while at the same time remaining logical and readable, efficient for indexing and flexible for future additions and adjustments. For this reason we decided on a format that is conceptually related to stand-off annotation (e.g. McKelvie et al., 1997; Thompson and McKelvie, 1997) by including various levels of information into separate annotation layers, however keeping them within the same document. Below we describe the DTD in more detail.

4.1 Overall Structure

The document consists of the body, which is further divided into a title (optional), any number of sentences and a group of keywords (also optional). The main division of layers occurs on the sentence level, with elements for chunks, grammatical relations, UMLS terms, EuroWordNet terms, UMLS semantic relations and for the text itself (tokens). The elements `chunk`, `const`, `term` and `ewnterm` refer to the `text` level through indices on tokens, while the `semrel` element refers to the `terms` level through indices on the pair of terms between which the relation was identified.

New annotation levels can be added simply by referring to existing indices (tokens, terms). Similarly, for project tasks that do not require all information, levels can be removed through simple reformatting without corrupting the document’s consistency.

4.2 Elements and Attributes

4.2.1 Document

The `document` element consists of one `title`, any number of `sentences` and one `keywords` element. It has the attributes `id`, `type`, `lang` and `corresp`. The `id` is the name of the original text file which also includes the name of the sub-corpus (e.g. *Arthroskopie.00130003.ger*). The only document type at present is abstract, however new types may be added when the corpus is expanded. The `lang` attribute specifies the language of the document and the `corresp` gives the id of its parallel document.

4.2.2 Title

The `title` of the document contains `terms`, `ewnterms`, `chunks`, `gramrels`, `semrels` and `text`, all of which may occur only once. The attribute `id` records the document id and has the suffix 0 (e.g. *Arthroskopie.00120003.ger.0*). The attribute `corresp` gives the `id` of the `title` element in the parallel abstract.

4.2.3 Sentence

Similarly to `title`, the `sentence` element contains `terms`, `ewnterms`, `chunks`, `gramrels`, `semrels` and `text`, once each, and as attributes each sentence has an `id` beginning with the letter *s* and the sentence number within the abstract (e.g. *s1*). The attribute `corresp` gives the id of the corresponding sentence in the parallel abstract.

4.2.4 Keywords

The `keywords` element contains any number of `keyword` elements and has no attributes. The `keyword` element contains the string of the keyword (`PCDATA`).

4.2.5 Text

The `text` element contains any number of `token` elements. The `token` element contains the token string (`PCDATA`), further information is encoded in the following attributes: `id` of the form sentence number, dot, prefix w and token number (e.g. *s1.w1*); `pos` giving the part of speech; `lemma` containing the base form of a simple token (if known) and six numbered attributes for lemmas if the token was analysed as a compound (`lemma1`, `lemma2` etc.). If the token is a punctuation mark, the `pos` attribute is set to `PUNCT`.

4.2.6 Chunks

The `chunks` element contains any number of `chunk` elements. The `chunk` is an empty element containing the attributes `id`, which gives the sentence number and the chunk number beginning with letter c (e.g. *s1.c1*); `from` which refers to the starting token of the chunk and `to` which refers to the final token of the chunk. The attribute `type` specifies the type of phrase (NP, VP, AP, CONJP).

4.2.7 GramRels

The `gramrels` element contains any number of `gramrel` elements. The `gramrel` element records an instance of a grammatical relation within the sentence and contains the attributes `id` of the form sentence number and gramrel number beginning with letter g (e.g. s1.g1), `tokenid` referring to the tokens belonging to the relation, `gramtype` specifying the grammatical role of the constituent (SUBJ, OBJ, IOBJ, ACT, PAS, PART, REFL) and `prob` given the probability with which the relation was identified.

4.2.8 Terms

The `terms` element contains any number of `term` elements. The `term` element is an empty element containing no string but the attributes `id`, which records the sentence number and the term number within the sentence beginning with the letter t (e.g. *s1.t1*); `tokenid` which refers to one or more tokens representing the term; `lemmaref` referring to the lemma number if the term was identified on the subtoken level; `type` specifying whether the term on the subtoken level was found as the modifyer (`Mod`) or the head (`Head`) of the compound; `preferred` giving the preferred variant of the term; `cui` giving the Concept Unique Identifyer and `tui` giving the Type Unique Identifyer.

Example:

```
<term id="s1.t4" tokenid="s1.w7 s1.w8"
preferred="Platelet Count" cui="C0032181" tui="T059"
/>
```

4.2.9 SemRels

The `semrels` element contains any number of `semrel` elements. Each `semrel` element records an instance of a semantic relation between two terms within a sentence. Its attributes are `id` built of the sentence number and the number of the relation prefixed with r (e.g. *s1.r1*), `relterms` giving the pair of term id's between which the relation was identified, and `reltype` specifying one of the 51 types of relations defined in the UMLS Semantic Network.

4.2.10 EwnTerms

The element `ewnterms` contains any number of `ewnterm` elements. The `ewnterm` element is empty and contains the following attributes: `id` giving the sentence number and the number of `ewnterm` beginning with the prefix e (e.g. *s1.e1*); `tokenid` referring to the token(s) in question and `sense` listing the code of the disambiguated sense from EuroWordnet.

References

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

McKelvie D., Brew C. and Thompson H. 1997. *Using SGML as a Basis for Data-Intensive NLP*. In Proceedings of ANLP97, Washington, DC.

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

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

Schiller A. 1996. *Multilingual Finite-State Noun Phrase Extraction* In: Proceedings of a Workshop on Extended Finite State Models of Language, Budapest, Hungary, ECAI'96.

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

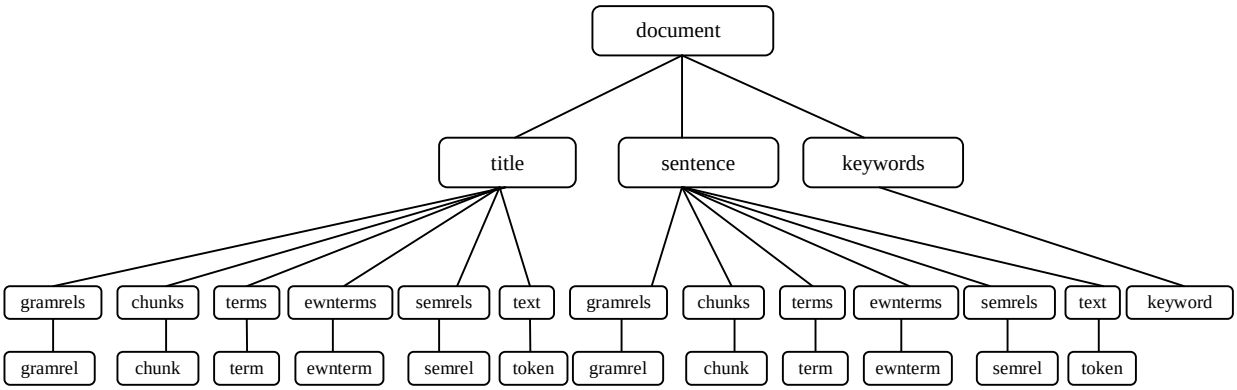
Thompson H. and McKelvie D. 1997. *Hyperlink semantics for standoff markup of read-only documents.* In Proceedings of SGML Europe 97, Barcelona.

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, Zurich.

Appendix I

```
<?xml version="1.0"?>
<!ELEMENT document (title?, sentence+, keywords?)>
<!ELEMENT title (chunks, gramrels, terms, ewnterms, semrels, text)>
<!ELEMENT sentence (chunks, gramrels, terms, ewnterms, semrels, text)>
<!ELEMENT keywords (keyword)*>
<!ELEMENT keyword (#PCDATA)*>
<!ELEMENT chunks (chunk)*>
<!ELEMENT chunk EMPTY>
<!ELEMENT gramrels (gramrel)*>
<!ELEMENT gramrel EMPTY>
<!ELEMENT terms (term)*>
<!ELEMENT term EMPTY>
<!ELEMENT ewnterms (ewnterm)*>
<!ELEMENT ewnterm EMPTY>
<!ELEMENT semrels (semrel)*>
<!ELEMENT semrel EMPTY>
<!ELEMENT text (token)*>
<!ELEMENT token (#PCDATA)*>

<!ATTLIST document id ID #REQUIRED
                   type CDATA #REQUIRED
                   lang (eng|ger) #REQUIRED
                   corresp CDATA #IMPLIED>

<!ATTLIST title id ID #REQUIRED
                corresp CDATA #IMPLIED>

<!ATTLIST sentence id ID #REQUIRED
                   corresp CDATA #IMPLIED>

<!ATTLIST chunk id ID #REQUIRED
                from IDREF #REQUIRED
                to IDREF #REQUIRED
                type (NP|PP|ADJP|CONJP)>

<!ATTLIST gramrel id ID #REQUIRED
                  tokenid IDREFS #REQUIRED
                  gramtype (SUBJ|OBJ|IOBJ|ACT|PAS|PART) #IMPLIED
                  prob CDATA #IMPLIED>

<!ATTLIST term id ID #REQUIRED
               tokenid IDREFS #REQUIRED
               lemmaref CDATA #IMPLIED
               type (Head|Mod) #IMPLIED
               preferred CDATA #IMPLIED
               cui CDATA #IMPLIED
               tui CDATA #IMPLIED>

<!ATTLIST ewnterm id ID #REQUIRED
                  tokenid IDREFS #REQUIRED
                  sense CDATA #IMPLIED>

<!ATTLIST semrel id ID #REQUIRED
                 relterms IDREFS #REQUIRED
```

```
                reltype CDATA #IMPLIED>

<!ATTLIST token id ID #REQUIRED
               pos CDATA #IMPLIED
               lemma CDATA #IMPLIED
               lemma1 CDATA #IMPLIED
               lemma2 CDATA #IMPLIED
               lemma3 CDATA #IMPLIED
               lemma4 CDATA #IMPLIED
               lemma5 CDATA #IMPLIED
               lemma6 CDATA #IMPLIED>
```



Appendix II

```
<?xml version="1.0" encoding="ISO-8859-1" ?>
<document id="DerHautarzt.80490581.eng" type="abstract" lang="eng">

 <title id="DerHautarzt.80490581.eng.s0" corresp=" DerHautarzt.80490581.ger.s0">

  <ewnterms>
   <ewnterm id="s0.e1" tokenid="s0.w7" sense="00769899" />
  </ewnterms>

  <terms>
   <term id="s0.t1" tokenid="s0.w2" preferred="Exanthema" cui="C0015230" tui="T184" />
   <term id="s0.t2" tokenid="s0.w7" preferred="Women" cui="C0043209" tui="T098" />
  </terms>

  <semrels>
   <semrel id="s0.r1" relterms="s0.t1 s0.t2" reltype="associated_with" />
  </semrels>

  <gramrels>
  </gramrels>

  <chunks>
   <chunk id="s0.c1" from="s0.w1" to="s0.w2" type="NP" />
   <chunk id="s0.c2" from="s0.w3" to="s0.w7" type="PP" />
   <chunk id="s0.c3" from="s0.w8" to="s0.w9" type="PP" />
  </chunks>

  <text>
   <token id="s0.w1" pos="ADJA" lemma="papular">Papular</token>
   <token id="s0.w2" pos="NN" lemma="exanthema">exanthema</token>
   <token id="s0.w3" pos="IN" lemma="in">in</token>
   <token id="s0.w4" pos="DT" lemma="a">an</token>
   <token id="s0.w5" pos="JJ" lemma1="HIV" lemma2="infect">HIV-infected</token>
   <token id="s0.w6" pos="JJ" lemma="african">African</token>
   <token id="s0.w7" pos="NN" lemma="woman">woman</token>
   <token id="s0.w8" pos="CC" lemma="with">with</token>
   <token id="s0.w9" pos="NN">histoplasmosis</token>
  </text>

 </title>

 <keywords>
  <keyword>HIV infection</keyword>
  <keyword>Histoplasmosis</keyword>
  <keyword>AIDS-defining illness</keyword>
 </keywords>

 <sentence id="s1" corresp="s1">

  <ewnterms>
   <ewnterm id="s1.e1" tokenid="s1.w5" sense="00769899" />
   <ewnterm id="s1.e2" tokenid="s1.w13" sense="00004767" />
   <ewnterm id="s1.e3" tokenid="s1.w15" sense="00239789" />
  </ewnterms>

  <terms>
   <term id="s1.t1" tokenid="s1.w5" preferred="Women" cui="C0043209" tui="T098" />
   <term id="s1.t2" tokenid="s1.w7" preferred="Fevers" cui="C0015967" tui="T184" />
   <term id="s1.t3" tokenid="s1.w9" preferred="Weights" cui="C0043100" tui="T081" />
   <term id="s1.t4" tokenid="s1.w15" preferred="Arms" cui="C0003792" tui="T029" />
   <term id="s1.t5" tokenid="s1.w9 s1.w10" preferred="Weight Loss" cui="C0043096"
    tui="T184" />
  </terms>

  <semrels>
   <semrel id="s1.r1" relterms="s1.t3 s1.t4" reltype="measurement_of" />
  </semrels>

  <gramrels>
```

```
   <gramrel id="s1.g1" tokenid="s1.w6 s1.w6" gramtype="ACT" prob="0.750" />
   <gramrel id="s1.g2" tokenid="s1.w5 s1.w6" gramtype="SUBJ" prob="0.017" />
   <gramrel id="s1.g3" tokenid="s1.w7 s1.w6" gramtype="OBJ" prob="0.056" />
   <gramrel id="s1.g3" tokenid="s1.w10 s1.w6" gramtype="OBJ" prob="0.106" />
  </gramrels>

  <chunks>
   <chunk id="s1.c1" from="s1.w1" to="s1.w5" type="NP" />
   <chunk id="s1.c2" from="s1.w9" to="s1.w10" type="NP" />
   <chunk id="s1.c3" from="s1.w11" to="s1.w13" type="PP" />
  </chunks>

  <text>
   <token id="s1.w1" pos="DT" lemma="a">A</token>
   <token id="s1.w2" pos="JJ">34-year-old</token>
   <token id="s1.w3" pos="VBN" lemma1="HIV" lemma2="infect">HIV-infected</token>
   <token id="s1.w4" pos="JJ" lemma="african">African</token>
   <token id="s1.w5" pos="NN" lemma="woman">woman</token>
   <token id="s1.w6" pos="VBN" lemma="develop">developed</token>
   <token id="s1.w7" pos="NN" lemma="fever">fever</token>
   <token id="s1.w8" pos="CC" lemma="and">and</token>
   <token id="s1.w9" pos="NN" lemma="weight">weight</token>
   <token id="s1.w10" pos="NN" lemma="loss">loss</token>
   <token id="s1.w11" pos="IN" lemma="on">on</token>
   <token id="s1.w12" pos="PRP" lemma="her">her</token>
   <token id="s1.w13" pos="NN" lemma="trunk">trunk</token>
   <token id="s1.w14" pos="CC" lemma="and">and</token>
   <token id="s1.w15" pos="NN" lemma="arm">arms</token>
   <token id="s1.w16" pos="punct">.</token>
  </text>

 </sentence>

</document>
```