Towards a Swedish Medical Treebank

Dimitrios Kokkinakis
Department of Swedish Language, Språkdata
Göteborg University

E-mail: svedk@svenska.gu.se

1. Introduction

Treebanks are linguistically annotated corpora with some previously established scheme of grammatical analysis. In any domain and in the (bio-) medical field in particular, such resources constitute a fundamental piece of knowledge for empirically-based, data-driven language processing, human language technologies and linguistic research and have attracted an increased interest during recent years. The interest for treebanks in biomedicine is guided by the fact that information extraction and (bio-) text mining research is shifting focus from the extraction and annotation of named entities to the extraction and annotation of relations and interactions between entities. This is usually associated by the extraction of verbal – alias predicate-argument – structures (*cf.* Kulick et al. [1], Tateisi et al. [2]). Semantic relations (e.g. between entities) and role extraction and labelling (e.g. agent, object) constitute a considerable challenge for automatic tools, although recent evaluation competitions such as the PASBio (Wattarujeekrit et al., [3]) and the BioCreAtIvE (Hirschman et al., [4]) revealed that some systems could present significant progress in performance in this area.

In this paper, we present our current activities towards the compilation and the multi-layered annotation of a domain-dependent corpus for Swedish in the area of medicine. The focus of the paper is based on the description of the constituent structure and functionally oriented annotation of the corpus. Moreover, the annotation scheme adopted, which incorporates three main layers of linguistic processing, lexical analysis, shallow semantic analysis and syntactic processing, will be exemplified. For the syntactic analysis we use a cascaded finite-state parser, aware of the shallow semantic annotations produced. The result of this analysis, including syntactic parsing and shallow semantic analysis, is transformed into the TIGER-XML interchange format ([5]). Our goal is to produce a large, rich in annotations, medical treebank suitable for both corpus-based grammar learning systems, for semantic relation extraction and for linguistic exploration of theoretical nature.

Motivation for this work is given in Section 2. Background work in the area of biomedical syntactic analysis and treebanking is presented in Section 3. Section 4 gives a brief description of the corpus used in this work, while Section 5 deals with the pre-processing steps applied into a sample of the corpus. Section 6 presents evaluation results based on this sample, while Section 7 summarizes the paper and proposes directions for future work.

Hajič, J. and Nivre, J. (eds.): Proceedings of the TLT 2006, pp. 199–210.

2. Motivation

Our motivation for processing a Swedish medical corpus initiated by the need to support lexical acquisition, terminology management and population of medical termbases. Corpus data is a valuable source for aiding e.g. the production of laymen dictionaries which we believe that in the long term will increase the accessibility of medical literature. Moreover, we are interested in supporting and improving information extraction and natural language processing (NLP) in the biomedical domain in general, particularly extraction of relations between terms. A research area that has not yet attracted much attention in Sweden, as opposed to NLP research in general discourse.

3. Background

While medium- to large-scale treebanks exist for English and other languages the situation for Swedish is surprisingly poor not only for genre-specific but also for general language corpora. Exceptions to this are the pioneering work of *Talbanken* ([6]) and *SynTag* ([7]) and the recent conversion of *Talbanken* to modern formats ([8]). Nevertheless, there are current activities in Sweden and in Scandinavia as a whole, through the *Nordic Treebank Network* (Nivre et al. [9]), aiming at the promotion of research related to treebanks. Activities that should have a positive impact during the coming years.

Considering now the approach we apply for parsing, this is based on finite-state cascades (Abney, [10]). Sequential finite-state transducers for the extraction of syntactic relations and for dealing with complex syntactic phenomena such as coordination and non-standard word order have been appeared in the literature a few times in the past. One of the earliest approaches to a deeper structural annotation similar to the approach that is described in this paper is given by Aït-Mokhtar & Chanod [11] for French, while a similar approach is applied by Müller [12] for German. Particularly in the biomedical field, there have been a number of different approaches to the annotation and extraction of various syntactic phenomena. Yakushiji et al., [13], applied a full parser for the extraction of argument structures (74% success) from biomedical papers. Pustejovsky et al. [14], for the extraction of "inhibit-relations" based on a similar process as Leroy et al. [15], for the extraction of various types of relations between entity noun phrases, e.g. proteins. A HPSG parser to identify predicate argument structures by inducing rules from a training corpus applied on a test corpus achieving 33% f-score is given by Yakushiji et al. [16]. Rinaldi et al. [17] used a probabilistic dependency parser on the GENIA corpus (Kim et al., [18]). The output is a hierarchical structure of syntactic relations achieving 90% precision and 86,2% recall on the identification of *subject* and 94,1% precision and 94,9% recall of *objects*. Moreover, Lease & Charniak [19] presented various adaptation techniques (domain specific part-of-speech, dictionary collocations and named entities) of a Penn Treebank-trained parser to the biomedical literature in order to overcome the need for a domain dependent treebank.

Their results showed improvements of the parsing accuracy considering the combination part-of-speech/named entities from 81.5% to 82.9%**.**

During recent years, evaluation competitions such as the *PASBio*, (Wattarujeekrit et al. [3]), aiming at the identification of predicate-argument structures have emerged. Despite some criticism (Cohen et al. [20]), PASBio is considered a viable formalism for building shallow semantic representations, suggesting a set of propositions and argument structures for biomedical verbs. The difficulties however for research related to treebanking in a specific genre corpora should not be underestimated. A discussion on the potential difficulties of work in the intersection between biomedicine and treebanking is given in Tateisi et al. [21]. Finally, the work by Chou et al. [22] is relevant in this context. Chou et al. developed the *BioProp*, a biomedical proposition bank, where predicate argument structures and semantic roles are annotated similarly to the *PropBank* (Palmer et al., [23]).

4. The MEDLEX Corpus

To the best of our knowledge there haven't been, until recently, any efforts for collecting Swedish medical corpora apart from the one described by Kokkinakis, [24]. Even for more widely-spoken languages, except probably for English, there only a few biomedical annotated resources known to the scientific community (e.g. Wermter & Hahn [25], for German), a fact that might have an implication for the design and implementation of a whole range of more effective biomedical applications for languages other than English.

Even though the situation for English is far better compared to other languages, there are still issues that need to be tackled. In a survey conducted by Cohen et al. [20], six English corpora (data sets) were examined w.r.t. structural and linguistic characteristics, and only one of these, GENIA, was found suitable for evaluating the performance on basic pre-processing tasks.

In our work we use parts of the MEDLEX-corpus, Kokkinakis [24]. MEDLEX consists of a variety of text-documents related to various medical text genres. The whole collection comprises 15 million tokens and includes: *teaching material, official documents, scientific articles from med. journals, conference abstracts, consumer health care documents, descriptions of diseases* etc. Out of this corpus, a subcorpus of 50 articles was selected. All articles come from the weekly edition of *Läkartidningen*, the official magazine of the Swedish Medical Association; section *Nya Rön*, (New Findings). Table 1 shows the characteristics of the corpus sample.

# tokens/#average length per article	17.230/344	
# discontinuous structures	21	see 5.1
# multi-words/auxiliaries/premodif.	207/390/101	see 5.1
# MeSH annotations	1078 – 1174 tokens	see 5.2
# named entities # named entities (time/measure)	522 –731 tokens 293 – 698 tokens	see 5.2 see 5.2
# medical	137 – 175 tokens	see 5.2

Table 1. Characteristics of the corpus sample

5. Corpus Processing

In order to capture a number of difficult linguistic problems at an earlier stage prior to parsing, and thus reduce ambiguity at the various levels of the linguistic processing, we decided to put emphasis on a number of pre-processing steps of the texts to be analyzed and which run sequentially.

5.1 Lexical Analysis - Layer 1

Initially, the corpus sample is processed with a module that recognizes and restores discontinuous structures which involve some sort of ellipsis as shown in figure (1a). In order to perform this step, it was necessary to recognize and segment compound forms, which in Swedish are written as one orthographic unit. As soon as the segmentation is performed (in 1b the segmentation point(s) are marked with '||'), the restoration of such structures becomes a trivial task using simple pattern matching (1c). These types of structures are common in the MEDLEX corpus as a whole, and their restoration aids the part-of-speech tagger to increase its performance.

<table>
<tr><td>alfa-, beta- och gammaglobulin
bakterie- eller svampinfektioner
tråd- och nålelektroder
stroke- och hjärtinfarktregister
(1a)</td><td>alfa-, beta- och gamma||globulin
bakterie- eller svamp||infektioner
tråd- och nål||elektroder
stroke- och hjärt||infarkt|| register
(1b)</td></tr>
<tr><td colspan="2">alfa||globulin, beta||globulin och gamma||<i>globulin</i>
bakterie||infektioner eller svamp||<i>infektioner</i>
tråd||elektroder och nål||<i>elektroder</i>
stroke||register och hjärt||infarkt||<i>register</i>[1]
(1c)</td></tr>
</table>

Figure 1. Restoration of elliptical constructions

The corpus is then annotated with part-of-speech using the TnT tagger (Brants, [26]) and the Swedish MULTEXT tagset (http://spraakbanken.gu.se/parole/tags.phtml). The tagger is not trained on texts from the medical domain, but its lexicon has been enhanced with medical terminology, roughly 10,000 new entries. The part-of-speech annotated texts pass through a NLP pipeline that performs a number of modification and annotation tasks, including lemmatization, by modifying the morphosyntactic features of the tags, considering a number of language-specific phenomena[2].

- *Multi-word expressions* and *conjoined compounds*; e.g. complex prepositions, `i stället för` (instead of); complex adverbials, `hur som helst` (anyhow), complex pronouns/determiners, `den här` (this one);
- *Modal, temporal auxiliary verbs and phrasal verbs;* auxiliary verbs are not marked by the part-of-speech tagger and thus, their recognition is an important step for proper verbal grouping;

[1] In case of >1 segmentation points, the rightmost segmentation is considered for the restoration.

[2] The recognition of multiword units has a positive effect in the improvement of the parsing results, at least in Swedish, (*cf.* Nivre & Nilsson [27]).

- *Various types of appositive nouns;* particularly common nouns taking different types of numerical information as arguments, e.g. kapitel 3 (chapter 3) and koagulationsfaktor VIII (coagulation factor VIII). The more “traditional” types of appositions such as *professor* John Krieger are not taken in consideration since they are properly treated by grammar rules that manage the named entities (Section 4.2)
- *Premodifying measure/quantity words;* e.g. miljon (million)

An example of part-of-speech annotation is given in Figure 2, including one of the examples from (1a) in context. Before, (2a), and after, (2b), the pre/postprocessing steps. Changes are designated with a ‘*’.

“For surface muscles can this be done with the help of skin electrodes while deep muscles require intramuscular thread or nail electrodes”			
För	SPS	För	SPS
ytligt	RGPS	ytligt	RGPS
liggande	AP000N0S	liggande	AP000N0S
muskulatur	NCUSN@IS	muskulatur	NCUSN@IS
kan	V@IPAS	kan	**VAIPAS*
detta	PF@NS0@S	detta	PF@NS0@S
göras	V@N0SS	göras	V@N0SS
med	SPS	med	**SPS-M*
hjälp	NCUSN@IS	hjälp	**NCUSN@IS-M*
av	SPS	av	**SPS-M*
hudelektroder	NCUPN@IS	hudelektroder	NCUPN@IS
medan	CSS	medan	CSS
djupare	RGCS	djupare	RGCS
liggande	AP000N0S	liggande	AP000N0S
muskulatur	NCUSN@IS	muskulatur	NCUSN@IS
kräver	V@IPAS	kräver	V@IPAS
intramuskulära	AQP0PN0S	intramuskulära	AQP0PN0S
tråd-	NCU00@0C	*trådelektroder*	**NCUPN@IS*
eller	CCS	eller	CCS
nålelektroder	NCUPN@IS (2a)	nålelektroder	NCUPN@IS (2b)

Figure 2. Part-of-speech tags, before and after pre/postprocessing

5.2 Shallow Semantic Analysis - Layer 2

The second step is *named entity* and *terminology recognition* and is actually independent of the previous, since its input is raw text and can thus be run as a stand alone application. However, the results from both steps, Layer 1 and 2, are merged into a single representation format and fed into the cascaded parser (see Section 5.3). The shallow[3] semantic analysis consists of three independent processes:

- generic named entity recognition
- MeSH annotation
- medical terminology recognition, a complementary step of the previous

[3] We call this layer ”shallow” in the sense that we do not try to semantically annotate all words in the text with e.g. senses, only a subset of the vocabulary.

The generic named entity tagger (Kokkinakis [28]), can recognize and annotate eight main types of named entities; *person, location, organization, object/artifact, event, work, time* and *measure expressions.* Each main category is further subdivided into finer-grained categories, so for instance the organization category is subdivided into *financial, media-related, athletic, cultural, political, educational etc.*

The Medical Subject Headings (MeSH®) tagger is the controlled vocabulary thesaurus of the U.S. National Library of Medicine (NLM), widely used for indexing medical data. The MeSH is a hierarchical thesaurus. This means that the terms of the vocabulary are arranged in a tree structure. Every category of terms (e.g. Anatomy) has its own tree. Terms with a broad coverage of the subject are placed at the root of the tree, whereas terms with a narrower scope are placed in the branches, becoming increasingly specific for each level in the tree. The Swedish MeSH tagger is based on the Swedish translation made by staff at the Karolinska Institute Library (http://mesh.kib.ki.se/swemesh/) and covering roughly 22.325 entries. The six most important hierarchies of MeSH are used for annotation, namely: A (Anatomy), B (Organisms), C (Diseases), D (Chemicals and Drugs), E (Analytical, Diagnostic and Therapeutic Techniques and Equipment), and F (Psychiatry and Psychology).

MeSH is a valuable resource but it is rather limited in coverage considering the wealth of terminology in the medical language. Therefore, we have complemented the MeSH annotations by developing yet another module that recognizes important types of terminology, particularly *names of pharmaceutical products, drugs, symptoms* and (anatomical) *Greek and Latin terms*. Several thousand names of pharmaceutical products, particularly names of drugs, have been obtained from the http://www.fass.se, a reference book of all medicines that are approved and used in Sweden, while terminology of Greek/Latin origin, particularly anatomical terms (such as the *Encyclopedia thoracica*) have been downloaded from the Karolinska Institute, at http://www.karolinska.se.

Ann Traynor och medarbetare vid Northwestern University, Chicago, USA, har funnit att en kombination av högdos kemoterapi samt stamcellstransplantation kan framgångsrikt användas vid svår SLE.

Ann Traynor and colleagues at the […] have found that a combination of high-dosage chemotherapy and stamcell transplantation can be used with success for (the treatment of) severe SLE.

```
<ENAMEX TYPE="PRS" SBT="HUM">Ann Traynor</ENAMEX> och medarbetare vid
<ENAMEX TYPE="ORG" SBT="EDU">Northwestern University</ENAMEX>, <ENAMEX
TYPE="LOC" SBT="PPL">Chicago</ENAMEX>, <ENAMEX TYPE="LOC" SBT="PPL">
USA</ENAMEX> , har funnit att en kombination av högdos <mesh tag="
E02.186.170/E02.319.170">kemoterapi</mesh> samt <mesh tag=
"E04.936.225.687">stamcellstransplantation </mesh> kan framgångsrikt
användas vid svår <mesh tag= "C17.300.480/C20.111.590">SLE</mesh> .
```

Figure 3. Shallow semantic annotations; generic entities, *<ENAMEX…>*[4]*,* and medical MeSH-terminology, *<mesh…>*.

[4] In the ENAMEX tag, *TYPE* stands for main type (e.g. PeRSon or ORGanization) while *SBT* stands for the subtype (e.g. HUMan or EDUcational).

Figure 3 shows an example of the annotations produced by the shallow semantic processing.

5.3 Syntactic Analysis - Layer 3

The results from all the previous processes are merged into a single representation format and fed into the syntactic analysis module, which is based on the Cass-parser, *Cascaded analysis of syntactic structure* (Figure 4). Cass is a partial parser designed for use with large amounts of noisy text. Cass uses a finite-state cascade mechanism and internal transducers for inserting actions and roles into patterns, and originates from the work by Abney, [10], which states that “robustness and speed are primary design considerations”. The Swedish grammar used by the parser has been developed by Kokkinakis & Johansson Kokkinakis [29], and has been modified and adapted in such a way that it is aware of the features provided by the pre-processors, particularly the medical phrases, which are incorporated into a new phrase level (see discussion below).

```
…
<id="c.24_10">   USA       NP00N@0S      usa           LOC/PPL
<id="c.24_11">   ,         FI            ,             --
<id="c.24_12">   har       VAIPAS        ha            --
<id="c.24_13">   funnit    V@IUAS        finna         --
<id="c.24_14">   att       CSS           att           --
<id="c.24_15">   en        DI@US@S       en            --
<id="c.24_16">   kombination             NCUSN@IS      kombination   --
<id="c.24_17">   av        SPS           av            --
<id="c.24_18">   högdos    NCUSG@DS-M    högdos        --
<id="c.24_19">   kemoterapiNCUSN@IS      kemoterapi
              E02.186.170/E02.319.170
<id="c.24_20">   samt      CCS           samt          --
<id="c.24_21">   stamcellstransplantation              NCUSN@IS
              stamcellstransplantation   E04.936.225.687
…
```

Figure 4. Input to Cass, including generated index required by TIGER-XML

The phrase patterns in Cass consist of finite-state rules; in turn bundles of rules are divided into different levels depending on their internal complexity, simpler follow complex ones. The processing is deterministic that it invokes a longest-match strategy. Moreover, the scheme we have adopted is theory independent, in the sense that it does not follow an established linguistic framework, such as HPSG. The parsing involves a cascade of *two* major automata, the *phrasal* and the *clausal*. All types of clauses are divided into different levels. The division depends partly on the type of the verbal group and the word order and partly on available lexicalized complementizer or part-of-speech tags that can provide strong evidence for a particular type of clause. The phrasal automaton includes:

- *phrases which include a entity annotation;* various labels depending on the entities involved, e.g. ‘np-ORG’;
- *phrases which do not include a named-entity annotation*;
- *adjectival phrases*;
- *prepositional phrases*;

- *verbal groups/chains*, e.g. `vg[har/VAIPAS inte/RG0S behandlats/V@IUSS]` (have not been treated). The adverb `inte` becomes part of the verbal group.

The “clausal” automaton includes:

- *embedded questions with interrogative pronouns*;
- *relative clauses*;
- *adverbial* and *infinitive clauses*;
- *complement clauses*,
- *wh-questions with interrogative adverb/pronoun*;
- *yes/no questions*;
- *copula passive constructions*;
- various types of *main clauses*;
- *combinations of various types of main and subordinated clauses;*
- *constructions without a verbal predicate.*

Since Cass does not provide any visualization capabilities and since the *Nordic Treebank Network* is in favor of the TIGER-XML format, we decided to also use the same format for the annotation scheme of the parsed output. TIGER-XML, (König & Lezius [5]), a flexible graph-based architecture for storage, indexing and querying. This way the syntactically analyzed results can be visualized and easily used for the querying the partially parsed corpus, e.g. by combining lexical features, semantic annotations and phrase labels. The phrase structure annotation is represented by a tree (Figure 5). Grammatical or surface functions (e.g. SuBJect, OBJect) are shown as category labels of non-terminal nodes. Time adverbials and adverbials that specify measure and location receive appropriate labels during entity recognition.

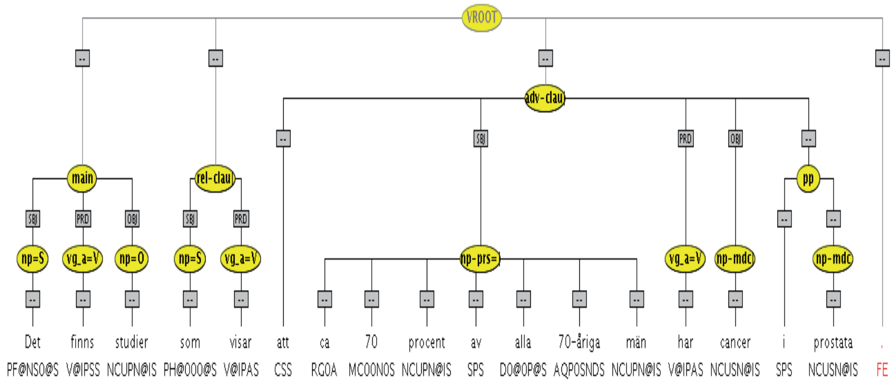

Figure 5. A TIGER-XML tree for the sentence: *There are studies showing that about 70 percent of all 70 year old men have cancer in the prostate.*

6. Towards a Swedish Medical Treebank - Evaluation

Although several errors could be found at all levels of processing, some do not seem to have impact for the purpose of grammatical relation extraction, while other play a more vital role and can explain the lower recall rates (Table 2). Errors include:

- part-of-speech errors (e.g. `s-transferas` tagged as verb instead of a noun; `associerat` tagged as verb instead of participle);
- unrecognized entities (e.g. `akut vestibulärt avbrott` [acute vestibular loss]);
- elliptic coordinations and lack of appropriate rules in the grammar. Particularly the case where scientific references were given in brackets at the end of sentences (e.g. `xxxxx [1,2].`);
- erroneous analysis of complex noun phrases (e.g. `kol-11 (11C) -märkta serotoninprecursorn 5-HTP ( 5-hydroxytryptophan )`).

Despite the errors, the terminology and entity tagging (see Section 5.2) enhances the performance of the parser in a positive direction. Phrase recognition and grouping, involving entities and terminology, depend more on the shallow semantic annotation than the part-of-speech one. Thus, for instance, the fragment "... `med rubriken » Utan vilja ingen säker vård «`" (...with the headline »...«) is annotated by the entity tagger as "... `med rubriken <ENAMEX TYPE="WRK" SBT="WAA">» Utan vilja ingen säker vård «</ENAMEX>`", here *WRK* stands for the entity category "written work and art". The parser will then group the annotated fragment as a single phrase with the label *np-WRK* irrespectively of the, erroneous or not, part-of-speech tags involved. This can be explained by the fact that the entity recognition is a reliable and accurate process (Kokkinakis [28]) and in case there are such annotations, the part-of-speech tags play a secondary role. Note though, that we have not exactly measured to which degree the shallow semantics have a positive effect for this type of part-of-speech "ignorance"

The strategy for the recognition of the syntactic functions follows the Scandinavian tradition of the topological frames/schema, which encodes word order regularities valid for a class of constituents occupying a specific position of a frame. The topographical structure of the surface strings decides which grammatical label a constituent may get. Furthermore, the subject, object and indirect objects are limited to noun phrases.

	found	*correct extracted*	*P*	*total available*	*R*
subject	#1334	#1259	94.3%	#1298	96.9%
object	#638	#564	88.4%	#608	92.7%
indirect object	#13	#2	15%	#4	50.0%

Table 2. Evaluation results of the syntactic functions

Table 2 gives the evaluation figures on the functional relations recognized by the parser. Precision was calculated as # *correct extracted relations/# total extracted relations* and recall as # *correct extracted relations/# total available*. Most of the errors in the "indirect object" case had to do with the lack of appropriate mechanism for dealing with the scientific references given in the running text.

7. Conclusions and Future Work

In this paper, we have described our efforts towards the annotation and syntactic analysis of a Swedish medical corpus sample. The annotation scheme consists of three layers, lexical analysis, shallow semantic and thesaurus lookup and syntactic analysis including the identification and annotation of grammatical functions. We believe that the two first layers have a positive impact in the performance of the parser in the form of the reliability and the quality of the evaluation results accomplished, compared to the mere use of part-of-speech tags. Although the annotation might seem "flat" at a first glance, it is rather rich and has potentials for further enhancements and improvements in order to produce a valuable labeled material. For this reason, we currently investigate the use of valency information for determining head-dependent relations between e.g. complements and predicates.

Since, manually inspected treebanks constitutes the reliable means for measuring progress in parser creation, for bootstrapping parsing systems etc., it is important that such resources keep the highest possible quality. Therefore, we have also started with the post-processing and qualitatively improvement of the material by looking at post-editing tools (Brants & Plaehn [30]). In the near future we plan to both increase the depth of the treebank by the integration of additional annotations, such as e.g. coreference chains, and its breadth by the use of additional texts from the MEDLEX corpus. More complex structures in a dependency-like fashion are under consideration (see previous discussion). At the moment, the treebank is not available, but we investigate ways to make the sample available for research.

Acknowledgements

This work was partially supported by the *Semantic mining - Semantic Interoperability and Data Mining in Biomedicine* - NoE 507505.

References

[1] Kulick S. et al. (2004). *Integrated Annotation for Biomedical Information Extraction.* Proc. of the HLT/NAACL, Boston.

[2] Tateisi Y., Yakushiji A., Ohta T. and Tsujii J. (2005). *Syntax Annotation for the GENIA Corpus*. 2nd Inter. Joint Conf. on Natural Language Processing (IJCNLP). Pp. 220-225. Korea.

[3] Wattarujeekrit T., Shah P. K. and Collier N. (2004). *PASBio: Predicate-argument Structures for Event Extraction in Molecular Biology*. BMC Bioinformatics, 5:155. OUP.

[4] Hirschman L. Yeh A., Blaschke C. and Valencia A. (2005). *Overview of BioCreAtIvE: critical assessment of information extraction for biology.* BMC Bioinformatics. 6 (Suppl 1):S1. OUP.

[5] König E. and Lezius W. (2003). *The TIGER Language - A Description Language for Syntax Graphs, Formal Definition*. Technical report. Institut für Maschinelle Sprachverarbeitung, University of Stuttgart.

[6] Einarsson J. (1976). *Talbankens skriftspråkskonkordans*. Lund

[7] Järborg J. (1986). *Manual för syntaggning*. Göteborgs universitet: Institutionen för språkvetenskaplig databehandling.

[8] Nivre J., Nilsson J. and Hall J. (2006). *Talbanken05: A Swedish Treebank with Phrase Structure and Dependency Annotation*. Proc. of the 5th Conf. on Language Resources and Evaluation (LREC2006). Genoa, Italy.

[9] Nivre J., De Smedt, K. and Volk M. (2005). Treebanking in Northern Europe: A White Paper. *Nordisk Sprogteknologi 2004: Årbog for Nordisk Sprogteknologisk Forskningsprogram 2000-2004*. København.

[10] Abney S. (1997). Part-of-Speech Tagging and Partial Parsing. *Corpus-Based Methods in Language and Speech Processing*. Young S. & Bloothooft G. (eds). Chap. 4:118-136. Kluwer AP.

[11] Aït-Mokhtar S. and Chanod J-P. (1997). *Subject and Object Dependency Extraction Using Finite-State Cascades*. Automatic Information Extraction and Building of Lexical Semantic Resources Workshop. Pp. 71-77. Spain.

[12] Müller F.H. (2004). *Annotating Grammatical Functions for German Using Finite-State Cascades*. Proc. of the 20th COLING. Pp. 268-274. Switzerland.

[13] Yakushiji A., Tateisi Y., Miyao Y. and Tsujii J. (2001). *Event Extraction from Biomedical Papers Using a Full Parser*. Pac Symp Biocomputing. Pp. 408-19

[14] Pustejovsky J., Castano J., Zhang J., Kotecki M,. Cochran B. (2002). *Robust Relational Parsing over Biomedical Literature: Extracting Inhibit Relations*. Pac Symp Biocomputing. Pp. 362-73.

[15] Leroy G., Chen H. and Martinez J.D. (2003). A Shallow Parser Based on Closed-Class Words to Capture Relations in Biomedical Text. *Journal of Biomedical Informatics*. Vol. 36:3, Pages: 145 – 158.

[16] Yakushiji A. Miyao Y., Tateisi Y. and Tsujii J. (2005). *Biomedical Information Extraction with Predicate-Argument Structure Patterns*. University of Tokyo. CREST, Japan Science and Technology Agency.

[17] Rinaldi F., Schneider G., Kaljurand K., Hess M., and Romacker M. (2006). *An Environment for Relation Mining over Richly Annotated Corpora: the case of GENIA*. 2nd International Symposium on Semantic Mining in Biomedicine. Jena, Germany.

[18] Kim J.-D. Ohta T., Tateisi Y. and Tsujii J. (2003). *GENIA corpus - a semantically annotated corpus for bio-textmining.* Bioinformatics Vol. 19 Suppl. 1. Pages i180-i182. OUP.

[19] Lease M. and Charniak E. (2005). *Parsing Biomedical Literature*. 2nd International Joint Conf. on Natural Language Processing (IJCNLP). Korea.

[20] Cohen K.B., Fox L., Ogren P.V. and Hunter L. (2005). *Empirical Data on Corpus Design and Usage in Biomedical Natural Language Processing*. Proc. of the American Medical Informatics Association (AMIA). Washington, DC, US.

[21] Tateisi Y., Ohta T. and Tsujii J. (2004). *Annotation of Predicate-argument Structure of Molecular Biology Text.* Proc. of the IJCNLP-04 Workshop on Beyond Shallow Analyses. China.

[22] Chou W.-C., Tsai R. T-H, Su, Y.-S., Ku, W., Sung, T.-Y., & Hsu, W.-L. (2006). *A Semi-Automatic Method for Annotating a Biomedical Proposition Bank.* Proc. of the ACL Workshop on Frontiers in Linguistically Annotated Corpora (LINC-2006).

[23] Palmer M., Gildea D., Kingsbury P. (2003). The Proposition Bank: An Annotated Corpus of Semantic Roles. *J. of Computational Linguistics.*Vol. 31, No. 1. Pp. 71-106.

[24] Kokkinakis D. (2006). *Collection, Encoding and Linguistic Processing of a Swedish Medical Corpus - The MEDLEX Experience*. Proc. of the 5th Languages Resources and Evalutaion (LREC). Genoa, Italy.

[25] Wermter J. and Hahn U. (2004). *An Annotated German-Language Medical Text Corpus as Language Resource*. Proc. of the 4th Conf. on Language Resources and Evaluation (LREC). Portugal.

[26] Brants T. (2000). *TnT - A Statistical Part-of-Speech Tagger*. Proc. of the Sixth Applied Natural Language Processing Conf. (ANLP)*,* Seattle, WA.

[27] Nivre, J. and Nilsson, J. (2004). *Multiword Units in Syntactic Parsing*. MEMURA 2004 - Methodologies and Evaluation of Multiword Units in Real-World Applications. Workshop at LREC. Lisbon, Portugal.

[28] Kokkinakis D. (2004). *Reducing the Effect of Name Explosion.* Proc. of the Beyond Named Entity Recognition, Semantic labelling for NLP tasks. Workshop at LREC. Lisbon, Portugal

[29] Kokkinakis D. and Johansson Kokkinakis S. (1999). *A Cascaded Finite-State Parser for Syntactic Analysis of Swedish.* Proc. of the 9th European Chapter of the Association of Computational Linguistics (EACL). Norway.

[30] Brants T. and Plaehn O. (2000). *Interactive Corpus Annotation*. Proc. of the 2nd International Conf. on Language Resources and Evaluation (LREC). Athens, Greece.