

# Developing Resources for Swedish Bio-Medical Text Mining

Dimitrios Kokkinakis

Department of Swedish Language, Språkdata

Göteborg university, Sweden

dimitrios.kokkinakis@svenska.gu.se

## Abstract

Collection and annotation of corpora in specialized fields, such as medicine, and particularly for lesser-spoken languages, than for instance English, is an important enterprise for the continuous development and growth of language technology research, for resource development and for the implementation of practical applications for these languages. In this paper, we describe our ongoing efforts to build a large Swedish medical corpus, the *MEDLEX Corpus*, how we combine generic named entity and terminology recognition for the detailed annotation of the corpus, and how these annotations are further utilized by an annotations-aware cascaded finite-state parser.

## 1 Introduction

A fundamental resource/prerequisite for empirically-based language processing and for higher levels of linguistic research, such as information extraction and text mining, is the identification and annotation of named entities and technical terminology. In this paper, we describe our ongoing efforts to build a large Swedish medical corpus, the MEDLEX Corpus, and the technologies we apply for a detailed and extensive annotation with named entities and medical terminology. For the first type of annotation, we use a generic system for Swedish named entity recognition, (Kokkinakis, 2004) while for the medical terminology, we use the Swedish translation of the MeSH thesaurus. As a complement to the MeSH annotator, we have developed yet another module that identifies *symptoms*, names of *pharmaceutical products* and *anatomical terms* of Greek/Latin origin, three categories for which MeSH lacks appropriate coverage. The result of the previous processes is further fed into an annotations-aware cascaded finite-state parser. The

parser, developed by Kokkinakis & Johansson Kokkinakis (1999), has been modified in such a way that can utilize the rich features provided by the pre-processors, which results into the effect of a slightly decreased complexity of the grammar rules. Most importantly, however, is the fact that the syntactically analyzed results can be used for querying the partially parsed corpus by combining lexical features, semantic annotations and phrase labels, since the parser's output has been converted to the TIGER-XML format (König & Lezius, 2003).

This paper starts by giving some background notes of related research in the previously outlined topics and continues with a brief description of the MEDLEX Corpus, Section 3. In Section 4 we provide the characteristics of the named entity recognizer and in Section 5 of the MeSH tagger and the normalization steps applied to the Swedish MeSH. Section 6 gives a brief description of the cascaded parser used in this study and how the results of the previous processes have been integrated in the parser and converted into the TIGER-XML format. Section 7 provides a small scale evaluation for the various components presented in this paper, based on a small sample from the weekly edition of *The Swedish Medical Association's magazine*, <http://www.lakartidningen.se>. Finally, conclusions and suggested ways to improve the various processes and directions for future research end the paper.

## 2 Background

Named entity recognition (NER) for semantic disambiguation is an important supporting technology in natural language processing (NLP) and has a great impact into progressing NLP-aware R&D activities such as text mining, information extraction and question answering. Generic NER, as used in this study, originates from the work in the Message Understanding Conferences (MUC) in the 90's, cf. Chinchor, 1997. UMLS - Unified Medical Language System - and MeSH

(a major component/subset of the previous) have been used for the annotation and indexing of various types of biomedical corpora and clinical texts, particularly for English, German and French, both in the context of monolingual (Cooper & Miller, 1998; Nadkarni *et al.*, 2001; Shin *et al.*, 2004; Friedman *et al.*, 2004 and Struble & Dharmanolla, 2004) and bi/multilingual studies (Volk *et al.*, 2002; Marko *et al.*, 2003). Considering biomedical corpora and its linguistic processing, the far more cited corpora originates from the MEDLINE database. Particularly, the GENIA corpus (Kim *et al.*, 2003) has been used in many bio-NLP related activities; e.g. Yaku-shiji *et al.* (2001), by applying a full parser for the extraction of argument structures.

### 3 A Swedish Medical Corpus

To the best of our knowledge there are currently no Swedish medical corpora (structurally and/or linguistically annotated or even un-annotated) available. 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, 2004, for German), a fact that constitutes a bottleneck for bio-NLP research and might have implications for the design and implementation of a whole range of more effective biomedical applications for lesser spoken languages. Even though, in a survey conducted by Cohen *et al.* (2005) six English corpora (data sets) were examined w.r.t. structural and linguistic characteristics, only one of these (GENIA) was found suitable for evaluating the performance of basic pre-processing tasks. The material tested by Cohen *et al.* included only abstracts and a limited range of genres; while the authors discuss that the annotation format seems to have an effect on wide-spread usage of these sets.

The MEDLEX Corpus<sup>1</sup> consists of a variety of text-documents related to various medical text subfields, and does not focus at a particular medical genre. Primarily, due to the lack of very large Swedish resources within a particular specialized area. Thus, the texts range through many sub-domains, genres and specialized topics, including pharmacology. All text samples (6 mil.

tokens) are fetched from heterogeneous web pages during the past year, and include: *teaching material, guidelines, official documents, scientific articles from medical journals, conference abstracts, consumer health care documents, descriptions of diseases, definitions from on-line dictionaries, editorial articles, patient's FAQs* etc. A large portion of the MEDLEX documents were in (X)HTML format, while there were a number of documents in PDF or in MS Word format. However, all texts have been converted to text files, tokenized and part-of-speech annotated, while a number of structural characteristics have been preserved using XML markup, particularly the source of origin, the title and date of issue of each article (where possible).

### 4 (Generic) NE Recognition

There is a whole range of named entities that can be encountered in various types of texts, and not only the “classical”, in the NER bibliography, types of named entities, i.e. *person, location* and *organization*, from which the designation “generic” originates. Following the paradigm proposed by Sekine (2004), we apply a rather fine-grained NE system for Swedish capable of recognizing eight main categories (*person, location, organisation, event, object, work & art, time* and *measure*) and nearly sixty subtype named entities, including a large set of different types of measure subgroups, such as: *pressure, frequency, weight, dosage, speed, volume* and *temperature*. The system is described in Kokkinakis (2004) and is based on a modular and scalable architecture consisting of five major components, making a separation between lexical, grammatical and algorithmic resources. The five components are:

- lists of multiword names taken from various Internet sites;
- a shallow parsing component that uses finite-state grammars, one grammar for each type of NE recognized
- a module that uses the annotations produced by the previous two components (which have a high rate in precision) in order to make decisions regarding possibly un-annotated entities. This module is inspired by the *Document Centred Approach* by Mikheev *et al.* (1999). This is a form of on-line learning from documents under processing which looks at

---

<sup>1</sup> Our motivation in collecting and annotating a Swedish medical corpus initiated by the need to support lexical acquisition and further population of term databases, during our department's involvement in the EU-funded Network of Excellence: *Semantic Interoperability and Data Mining in Biomedicine* - NoE 507505.4.

- unambiguous usages for assigning annotations in ambiguous words<sup>2</sup>
- lists of single names (approx. 100,000)
- a theory revision and refinement module making a final control on an annotated document with named-entities in order to detect and resolve possible errors and assign new annotations based on existing ones, for instance by combining various annotation fragments

The generic NER system's performance has been evaluated on Swedish electronic patient records for each named entity type separately (Kokkinakis, 2005), except the *measure* module which was only evaluated for precision<sup>3</sup>. For the evaluation, the standard metrics Precision [(Total Correct + Partially Correct) / All Produced Annotations] and Recall [Recall = (Total Correct + Partially Correct) / All Possible Annotations] were used. Partially correct means that two annotations are not completely identical but that partial credit should be given. For instance, if the system produces a partial annotation for the expression: "*National Institute of Child Health and Human Development*" as "<ENAMEX TYPE="ORG" SBT="CRP">*National Institute of Child Health*</ENAMEX> and *Human Development*" (where ORG=ORGanization and CRP=CoRPoration), instead of marking the whole string, then, the produced annotation is not 100% correct but neither 100% wrong. Therefore, such annotations received the score 0.5, half point, instead of 1. In a study reported by Kokkinakis (2005) the evaluation figures for each entity group ranged between 69,9%-100% precision and 66%-98% recall.

## 5 Swedish MeSH

The Medical Subject Headings, MeSH®, is the controlled vocabulary thesaurus of the NLM, U.S. National Library of Medicine. The original data from NLM have been supplemented with Swedish translations made by staff at the Karolinska Institute Library<sup>4</sup> based on the year 2006 MeSH.

<sup>2</sup> By "document centred" is meant that at each stage of processing the system makes decisions according to a confidence level that is specific to that processing stage, and drawing on information from other parts of the document.

<sup>3</sup> The reason has been that at the time of the evaluation there was no access to the definition of all measure-related acronyms and abbreviations used in the evaluation texts.

<sup>4</sup> For more information visit: <http://mesh.kib.ki.se/swemesh/swemesh.cfm>.

### 5.1 Term Conversion & Normalization

A number of conversion and normalization steps were applied to the original material. These steps were necessary before the actual implementation of the MeSH-annotator due to the nature of the original data. The implementation follows an almost case-independent (see later this section), finite-state approach.

The first step applied was to change the order of the head and modifier complements as well as term variants with commas, in the original material (Table 1-a). There are several hundreds of such cases in the database (for obvious terminological and lexicographic purposes, e.g. easier sorting based on head words) that had to be changed in order to be able to apply the terminological material on corpora.

The second step was to normalize all non-inflected entries into a neutral non-inflected variant, and to add inflectional morphology and morphological variants for each entry (term and modifiers) as an optional feature using regular expressions. This way the annotator could be easily applied on raw (un-stemmed) text, (Table 1-b1&b2). After some initial annotation tests on parts of the MEDLEX corpus, we made some adjustments to the implemented recognizer since we discovered and wanted to capture some frequent phenomena of misspellings, agreement errors and orthographic variants that could be observed in the annotated sample texts. Some of those errors are probably caused by the high variability in the expression of similar concepts by different authors (e.g. the following spellings of "diarrhea" could be found in MEDLEX: *diarré*, *diarre*, *diarree*, *diarée*, *diarrhea*, *diarré*, *diarreé*), and by the influence from the English language, particularly the orthographic variation (e.g. use of 'ph' instead of 'f'; use of 'th' instead of 't' and use of 'c' instead of 'k'), (Table 1-c). Some (probable) errors in the original material were also corrected, while some discrepancies were minimized and normalized (Table 1-d).

Finally, case folding was applied to all terms, except those consisting of uppercase letters, which were almost exclusively acronyms. This was necessary in order not to introduce new forms of ambiguity during testing. A 100% elimination of case information could introduce new ambiguities between homographs uppercase/low case words. For instance, between *kol/D01.268.150* (i.e. "carbon") and *KOL/C08.381.495.389* (i.e. "Chronic Obstructive Pulmonary Disease").

a) word order	Vacciner, orala <i>changed to</i> orala vacciner Cellulosa, oxiderad <i>changed to</i> oxiderad cellulosa
b1) inflection b2) inflection patterns	Sir2-liknande <b>proteiner</b> <i>changed to</i> Sir2-liknande protein(er)? mannosbindande <b>protein</b> <i>changed to</i> mannosbindande protein(er)? orala vacciner <i>changed to</i> oral(a)? vaccin(et er erna)? oxiderad cellulosa <i>changed to</i> oxiderad(e)? cellulosa(n)?
c) variability	nervus abducens <i>added</i> n. abducens; staphylococcus aureus <i>added</i> staph aureus; aorta <i>added</i> aortae; pleura <i>added</i> pleurae; escherichia coli k12, o157 <i>added</i> e. coli; dyspne <i>added</i> dyspné; +bacter <i>added</i> +bakter; +plasi <i>added</i> +plasia; +diagnos <i>added</i> +diagnosis etc.
d) (possible) errors and discrepancies	Both: “ <b>anaeroba</b> bakterier” and “gramnegativa <b>anaereoba</b> ” Both: “ <b>ärflig</b> amyloidos” and “ <b>ärflig</b> spastisk paraplegi” Both: “ <b>Barretts</b> esofagus” and “ <b>Barrets</b> metaplasii” Both: “Pyruvatkarboxylasbrists <b>jukdom</b> ” and “I-brists <b>jukdom</b> ” Use of definite/indefinite forms: “i <b>urinen</b> ” and “i <b>urin</b> ” Use of singular/plural forms: “septumdefekt” and “+septumdefekter”

Table 1. Conversion and normalization steps

For the implementation of the MeSH annotator we use the most important subtree hierarchies from MeSH, namely A (Anatomy, 3277 terms), B (Organisms, 5407), C (Diseases, 16334 terms), D (Chemicals and Drugs, 18369 terms), E (Analytical, Diagnostic and Therapeutic Techniques and Equipment, 5265 terms) and F (Psychiatry and Psychology, 1528 terms). Moreover, in order to reduce the ambiguity space of the investigated problem, we decided to only use the upper level (level 0) of the lexical hierarchy for the classification of each term<sup>5</sup>. For instance, the term *beta-Lactamases* (*Betalaktamaser*) has the label *D08.811.277.087.180* which was reduced to *D08 [Enzymes and Coenzymes]*.

## 5.2 Enhancing the Terminology Annotation - Symptoms, Pharmaceuticals and Greek/Latin Terms

Apart from the MeSH terms, medical corpora contain a lot of other types of terminology that needs special treatment. MeSH lacks for instance information on (at least) three types of such terminology: *symptoms, names of pharmaceutical*

*products, drugs, and (anatomical) Greek and Latin terms.*

Symptoms are usually realized in the Swedish texts either as periphrastic expressions (phrases containing a preposition targeting an anatomical reference), or as compounds (a single orthographic unit) and it is rather difficult to find suitable lexical resources on the Internet in order to simply apply some sort of dictionary lookup process. Therefore, we have investigated the way these expressions are constructed in the collected corpus, by initially selecting a few characteristic symptom key-words, such as *värk* i.e. ‘pain’ and short phrase fragments, such as *ont i* i.e. ‘pain in’. Then, we created regular expressions with this partial information and fragments and applied them on an analysed version of the corpus (annotated with named entities and the MeSH terminology). This way we could: (i) confirm that the patterns were relevant and accurate and, more importantly (ii) identify new symptoms in the near vicinity of the already matched ones and (iii) implement a new set of hand constructed rules using regular expressions with the data gathered by this process. In approx. 75% of the examined cases, more than one symptom was actually co-occurring with other symptoms in the same sentence, sometimes up to five symptoms. Therefore, we could rapidly compile a long list of patterns that are now used for symptom recognition with high coverage.

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 have been downloaded from the Karolinska institutet, at <http://www.karolinska.se>.

## 6 Cascaded Parsing

The results from the NER and terminology recognition are merged into a single representation format and fed into a syntactic analysis module, which is based on the Cass-parser, *Cascaded analysis of syntactic structure*.

<id=" <b>c.5_1</b> "> EU	NPOONO	ORGANIZATION
<id=" <b>c.5_2</b> "> ger	VMIPA	N/A
<id=" <b>c.5_3</b> "> 30	MCOPNO	CURRENCY
<id=" <b>c.5_4</b> "> miljoner	NCUPNI-MSR	CURRENCY
<id=" <b>c.5_5</b> "> kronor	NCUPNI	CURRENCY
<id=" <b>c.5_6</b> "> till	S	N/A
<id=" <b>c.5_7</b> "> forskning	NCUSNI	N/A
<id=" <b>c.5_8</b> "> om	S	N/A
<id=" <b>c.5_9</b> "> sjukdomen	NCUSND-MSR	mesh:C23
...		

Figure 1. Input format for the parser

<sup>5</sup> Reduced MeSH hierarchies are used among others by Rosario *et al.* (2002) in an experiment for assigning (English) noun compound relations.

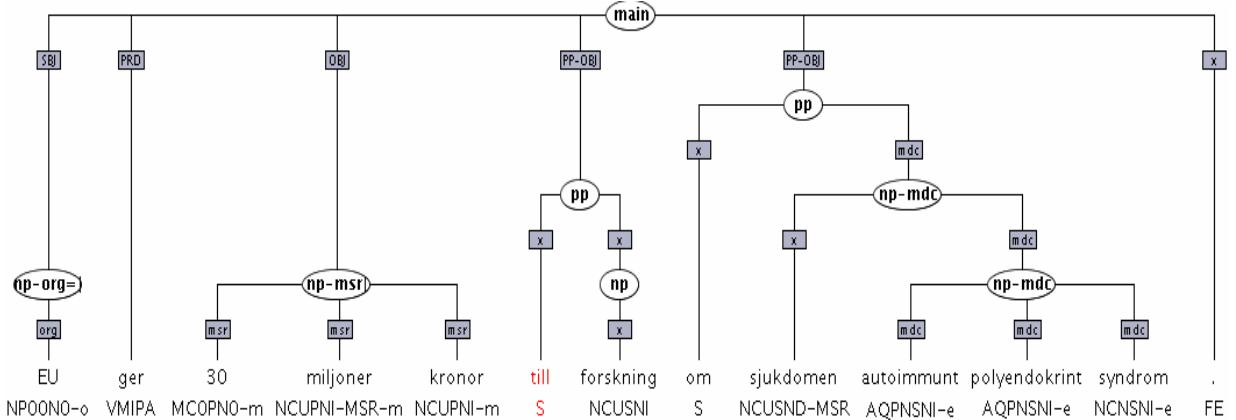


Figure 2. Annotation of the sentence: *EU gives 30 million (Swedish) krona to research on the (disease) autoimmune polyendocrine syndrome.*

Cass uses a finite-state cascade mechanism and internal transducers for inserting actions and roles into patterns, and originates from the work by Abney, (1997). The parser we use has been developed by Kokkinakis & Johansson Kokkinakis (1999), and has been modified in such a way that is now aware of the features provided by the pre-processors, which results into the effect of slightly decreased complexity of the grammar rules. Moreover, we also apply a number of pre-processing steps in order to capture a number of difficult linguistic problems at an early stage of parsing, and thus reduce ambiguity at the various levels of the linguistic processing<sup>6</sup>. Thus in all, but the final step, an input text passes a pipeline of finite-state grammars that may add or modify features to the part-of-speech annotated input; including the recognition and annotation of *multi-word expressions, conjoined compounds, phrasal verbs, various types of appositions and pre-modifying measure/quantity words*.

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. Although the evaluation provided in Section 7 is only based on the recognition of noun phrases, it is worth noting that the parsing involves a cascade of two major automata, the “phrasal” and the “clausal”. The “phrasal” includes: *phrases which include a named-entity annotation* (various labels depending on the entities involved, e.g. ‘np-ORG’); *phrases which do not include a named-entity annotation*; (‘np’); *adjectival phrases*; (‘ap’); *prepositional phrases*;

(‘pp’); *verbal groups/chains*; (‘vg’). 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 and constructions without a verbal predicate*. 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. Finally, for the annotation scheme of the parsed output we have chosen the TIGER-XML encoding format (König & Lezius, 2003), a flexible graph-based architecture for storage, indexing and querying. This way the syntactically analyzed results can be easily used for querying the partially parsed corpus by combining lexical features, semantic annotations and phrase labels.

## 7 Evaluation

We evaluated all major parts of the system on articles taken from the weekly edition of “The Swedish Medical Association’s magazine”, *Läkartidningen*, number 0550&0601/volumes 102-103, <http://www.lakartidningen.se>. The material consisted of a small number of articles, 10, a total of 8,490 tokens. The number of articles was kept small since we had to manually verify for each annotated segment, whether the entities and MeSH annotations and, particularly, the disambiguation of the MeSH annotations<sup>7</sup>, were correct

<sup>6</sup> The use of sequential finite-state transducers in a similar fashion as in our paper is described by Aït-Mokhtar & Chanod (1997) for French and Müller (2004) for German.

<sup>7</sup> The best scenario would have been to have a trained physician to evaluate the MeSH annotations, but we didn’t have that opportunity to do so at the time we conducted this study. However, this is planned in the near future due to our

or not compared to the on-line MeSH. The evaluation also included the performance of Cass on the recognition of noun phrases<sup>8</sup>, which are considered important segments for indexing.

## 7.1 Partial Disambiguation of MeSH Terms

The current implementation applies a partial and simplistic disambiguation methodology in lack of suitable training material and it is inspired by the “one sense per discourse” statement by Gale *et al.* (1992). We observed, therefore, that in many cases the unambiguous readings can help disambiguating the meaning of an ambiguous term (relationship ambiguity), this is what we also call for “contamination” principle; an unambiguous annotation “contaminates” its ambiguous neighbours, hopefully to proper disambiguation. For instance, the fragment “...*lokalanes-tetikum i inhalation (lidokain, bupivakain) kan blockera symptomen*”, i.e. “...local anesthetics during inhalation (Lidocaine, Bupivacaine) can block the symptoms”, is annotated by the MeSH tagger as “... *lokalanes-tetikum i inhalation (<mesh tag=“D02”>lidokain</mesh>, <mesh tag=“D02/D03”>bupivakain</mesh>) kan blockera symptomen*”; that is “*lidokain*” is annotated as *D02[Acetanilides]* and “*bupivakain*” as *D02[Acetanilides]* and *D03[Pipecolic Acid]*. Thus, according to our assumption that near unambiguous neighbours can disambiguate their ambiguous counterparts, the annotation of “*bupivakain*” will be reduced to *D02* which is actually the preferred meaning. At the same time the system adds a “reliability” attribute to the disambiguated annotation, which indicates the strength of the confidence for the ambiguity elimination, “<*mesh tag=“D02” indication=“VERY STRONG”> bupivakain</mesh>”.*

Thus, after a first annotation with MeSH, the system collects all annotations already identified in a document and uses the information from the already existing mark-up in order to attempt disambiguation, or even find new annotations, by applying the following algorithm, which progressively weakens the terms’ indication of strength, relative to the distance between them:

For each ambiguous annotation in *a single document at a time*:

---

department’s collaboration with the Sahlgrenska university hospital in the Semantic Mining NoE.

<sup>8</sup> The parser can also produce more than simple np chunking (see Section 6), for instance identify various types of phrases and clauses as well as syntactic functions, such as subject and object.

*if* there is  $\geq 1$  unambiguous tag(s) in the same sentence and there is an overlap between this/these and an ambiguous one, then reduce the ambiguous tag(s) and note *indication=“VERY STRONG”*

*elsif* there is  $\geq 1$  unambiguous tag(s) in the same paragraph and there is an overlap between this/these and an ambiguous one, then reduce the ambiguous tag(s) and note *indication=“STRONG”*

*elsif* there are  $>1$  unambiguous tag(s) in the same article and there is an overlap between this/these and an ambiguous one, then reduce the ambiguous tag(s) and note *indication=“MODERATE”*

The same steps as above also apply when there are more than two ambiguous annotations and there is an overlap between their tags. For instance the tag *C05/C17/C20* in the annotated segment: “[...] <*mesh tag=“C05/C17”> reumatiska sjukdomar</mesh> [...] <*mesh tag=“C05/C17/C20”>reumatoid artrit</mesh>*” i.e. “rheumatic diseases ... rheumatoid arthritis” will be reduced to *C05/C17*. Unannotated acronyms, following a MeSH annotation (ambiguous or not) between parenthesis or commas, receive the same annotation as the preceding annotated term. For instance, the acronym *ASD* in the segment: “<*mesh tag=“C14/C16”> förmaksseptumdefekt</mesh> (*ASD*)” i.e. “Atrial Septal Defect” will get the ambiguous annotation *C14/C16*.**

Complex cases, and mixture of the above, did occur, such as: “*Kronisk <mesh tag=“C08”>lungsjukdom </mesh>, som också kallas <mesh tag= “C08/C16”>bronkopulmonell dysplasi</mesh> ( BPD )*” i.e. “chronic lung disease, which is also called Bronchopulmonary Dysplasia (BSD)”. In this case, *C08/C16* will be reduced to *C08*, which will also be the tag assigned to the acronym *BSD*.

There were also cases of conflicts, when two or more different annotations can be candidates for partial or whole disambiguation of an ambiguous tag. In such cases the annotation is modified according to the one that has the most occurrences in the document.

## 7.2 Results & Discussion

For the evaluation of the named entities (including the three groups *symptoms*, *drug names* and *Greek/Latin anatomical terms*) we used the metrics precision and recall defined in Section 4. The results from the NER (Table 2) gave high figures in precision and recall which is due to the fact that the system we used has been tested dur-

ing a long period on various types of texts and that it also utilizes large lexical resources. The only category that had poor performance was the “Wrk&Art” type (e.g. names of projects, books, studies etc). This can be explained by the fact that one of the evaluation articles dealt with the comparison between international scientific studies and trials (e.g. TNT, CTT, IDEAL) using a lot of acronyms without proper introduction<sup>9</sup> with keywords, at least in the near context of the acronyms. There were no cases of event names (e.g. athletic events ‘the Olympic Games’) or Greek/Latin terms in the sample.

NE	C	P	W	M	Pr	R
Organization	29	0	3	2	.90	.93
Person	33	0	5	4	.86	.89
Location	19	0	0	0	1	1
Work&Art	13	7	0	23	.82	.38
Object	3	0	0	1	1	1
Event	0	0	0	0	-	-
Time	89	3	0	12	.98	.87
Measure	101	3	0	3	.98	.95
Symptoms	26	0	0	7	1	.78
Pharmaceut.	24	0	0	0	1	1
Greek/Latin	0	0	0	0	-	-
<b>Total</b>	<b>337</b>	<b>13</b>	<b>8</b>	<b>52</b>	<b>.97</b>	<b>.85</b>

Table 2. Evaluation results for NER (C=Correct; P=Partial, W=Wrong, M=Missed)

For the evaluation of the MeSH terms we calculated both the amount of ambiguity reduction achieved, for the complete matches, as well as the coverage of the Swedish MeSH on the sample (Table 3).

all annots. - correct annotations	601 - 594
unambiguous MeSH annotations	268
initial ambiguous MeSH annotations	105
disambiguated 1 MeSH tag left	58
disambiguated >1 MeSH tags left	17
final ambiguous MeSH annotations	30
# had 1 concept, full match, in MeSH	268 (45%)
# had 1 concept, partial match	97 (16%)
# had >1 concepts, full match, in MeSH	105 (18%)
# had >1 concepts, partial match	124 (21%)
# had no match in MeSH*	≈35

Table 3. Ambiguity reduction and MeSH coverage  
(\*subjective estimation)

The small scale experiment revealed some incompleteness of the Swedish MeSH w.r.t. applying it to the text sample. Some of the terms not recognized included: *sitosterolemi*, *kampesterol*, *lykopen*, *tunntarmspassage*. At the same time, simple steps (for instance by using orthographic

<sup>9</sup> Maybe most of these acronyms are considered as “obvious” to the target audience of the magazine.

variants and normalization, Section 5.1) have the ability to considerably increase coverage and thus aid the enhancement of the current gaps. Swedish is a compound language and thus compounding can be utilized for fast accessing to partially annotated segments that can aid the enhancement of the MeSH hierarchy (e.g. *knä<mesh tag="A02">skelettet</mesh>*, *re<mesh tag="C23">infarkt</mesh>*, *serum<mesh tag="D04/D10">kolesterol</mesh>*) by applying some suitable interface, an important research topic that requires further investigation.

A handful of simple heuristic pattern matching rules could also capture a number of unknown to the system acronyms and thus assign a MeSH label. This is an important part of the annotation of the documents, since acronyms are usually introduced once in a text and then frequently used in the same document instead of the expanded form. There were a few forms (7 occurrences) of lexical ambiguity, homography/homonymy, between terms and non-medical words (e.g. “*sena*” – ‘late’ and ‘tendon[A02]’; “*hand*” – in adverbial phrases ‘*i första hand*’ and ‘*hand[A01]*’; “*sänka*” - ‘to sink’ and ‘blood sedimentation[E01]’ and “*leder*” – ‘to lead’ and ‘*joints[A02]*’).

Finally, for the evaluation of the noun phrases we calculated the number of nps correctly and partially identified as well as the erroneously and the missed ones. There were 2,509 noun phrases marked by the parser. 2,422 (96,5%) were correctly identified, 59 (2,3%) were partially identified, while 22 (0,9%) were wrong and 6 were missed (0,2%). Most of the wrong and missed ones depend on a combination of erroneous part-of-speech annotation (e.g. long sequences of English segments, in which some words were tagged as verbs) and wrongly identified elliptical and coordinated phrases.

## 8 Conclusions

We have outlined our continuous work on gathering and linguistically processing a Swedish medical corpus. There are several issues that need to be investigated in more depth. For instance, the use of a human in the process loop, in order to inspect intermediate results. The need to conduct an evaluation on a larger scale, and possibly using the full MeSH levels, and/or doing things in another order. Maybe the MeSH results can benefit from applying parsing before annotation, and thus let the MeSH tagger only look inside np’s. For the coverage of MeSH, a trained

physician would have been the right person to mark unlabelled terminology. Some revisions and extensions of the disambiguation part are also worth further exploration. It is well-known that the polysemous words' meaning depend on the context of use, at least on non-technical corpora, a fact that might even be stronger in technical corpora, i.e. a term probably shares the same sense throughout a single document. The Swedish MeSH contains over 50,000 terms (incl. synonyms), but it still does not cover all clinically useful terminology and empirical studies can be of benefit for its content's growth.

## References

- Abney S. (1997). Part-of-Speech Tagging and Partial Parsing, *Corpus-Based Methods in Language and Speech Processing*. Young S. & Bloothoof G. (eds). Chap. 4:118-136. Kluwer AP.
- Ait-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. Vossen P. et al. (eds), pp. 71-77. Spain.
- Chinchor N. (1997). *MUC7 Named Entity Task Definition*. Technical Report, NIST.
- Cohen K.B., Fox L., Ogren P.V. and Hunter L. (2005). *Corpus Design for Biomedical Natural Language Processing*. Proceedings of the ACL-ISMB Workshop on Linking Biological Literature, Ontologies and Databases: Mining Biological Semantics. Pp 38-45. Detroit, US.
- Cooper G. and Miller R.A. (1998). *An Experiment Comparing Lexical and Statistical Methods for Extracting MeSH Terms from Clinical Free Text*. J Am Med Inform Assoc.; 5(1): 62-75.
- Friedman C., Shagina L., Lussier Y. and Hripcsak G. (2004) *Automated Encoding of Clinical Documents Based on Natural Language Processing*. J Am Med Inform Assoc. 11(5): 392-402.
- Gale W., Church K.W. and Yarowsky D. (1991). *One Sense per Discourse*. Proceedings of the DARPA Speech and Natural Language Workshop.
- 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 1180-1182
- Kokkinakis D. (2004). *Reducing the Effect of Name Explosion*. Language Resources and Evaluation Conference (LREC) Workshop: Beyond Named Entity Recognition Semantic Labelling for NLP tasks. Portugal.
- Kokkinakis D. (2005). Identification of Named Entities and Medical Terminology in Swedish Patient Records. *WSEAS Transactions on BIOLOGY and BIOMEDICINE*. Issue 3:2. Pp. 312-317.
- 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 (EAACL). Bergen, Norway.
- König E. and Lezius W. (2003). *The TIGER Language - A Description Language for Syntax Graphs, Formal Definition*. Technical report Institut für Maschinelle Sprachverarbeitung, U. of Stuttgart.
- Marko K., Daumke P., Schulz S., Hahn U. (2003). *Cross-language MeSH indexing using morpho-semantic normalization*. AMIA Annual Symp Proc.: 425-9
- Mikheev A., Moens M. and Grover C. (1999). *Named Entity Recognition without Gazetteers*. Proceedings of the EACL'99, Bergen, Norway. pp. 1-8
- Müller F.H. (2004). *Annotating Grammatical Functions for German Using Finite-Stage Cascades*. Proc. of COLING. Pp. 268-274. Switzerland.
- Nadkarni P., Chen R., Brandt C. (2001). *UMLS concept indexing for production databases: a feasibility study*. J Am Med Inform Assoc. Jan-Feb;8(1):80-91
- Rosario B., Hearst M.A. and Fillmore C. (2002). *The Descent of Hierarchy, and Selection in Relational Semantics*. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, pp. 247-254.
- Sekine S. (2004). *Definition, dictionaries and tagger for Extended Named Entity Hierarchy*. 4th Conference on Language Resources and Evaluation (LREC). Portugal
- Shin K., Han S-Y., Gelbukh A. (2004). *Balancing Manual and Automatic Indexing for Retrieval of Paper Abstracts*. Journal Title: Text, Speech and Dialogue (TSD).
- Struble C. and Dharmanolla C. (2004). *Clustering MeSH Representations of Biomedical Literature*. BioLink. Boston, MA.
- Volk M. et al. (2002). *Semantic Annotation for Concept-Based Cross-Language Medical Information Retrieval*. International Journal of Medical Informatics, Volume 67:1-3.
- Wermter J. and Hahn U. (2004). *An Annotated German-Language Medical Text Corpus as Language Resource*. 4th Conference on Language Resources and Evaluation (LREC). Portugal.
- 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.