# Collection, Encoding and Linguistic Processing of a Swedish Medical Corpus – The MEDLEX Experience

## **Dimitrios Kokkinakis**

Göteborgs Universitet

Department of Swedish Language, Språkdata

Box 200, SE-405 30 Göteborg

E-mail: dimitrios.kokkinakis@svenska.gu.se

#### **Abstract**

Corpora annotated with structural and linguistic characteristics play a major role in nearly every area of language processing. During recent years a number of corpora and large data sets became known and available to research even in specialized fields such as medicine, but still however, targeted predominantly for the English language. This paper provides a description of the collection, encoding and linguistic processing of an ever growing Swedish medical corpus, the MEDLEX Corpus. MEDLEX consists of a variety of text-documents related to various medical text genres. The MEDLEX Corpus has been structurally annotated using the Corpus Encoding Standard for XML (XCES), lemmatized and automatically annotated with part-of-speech and semantic information (extended named entities and the Medical Subject Headings, MeSH, terminology). The results from the processing stages (part-of-speech, entities and terminology) have been merged into a single representation format and syntactically analysed using a cascaded finite state parser. Finally, the parser's results are converted into a tree structure that follows the TIGER-XML coding scheme, resulting a suitable for further exploration and fairly large Treebank of Swedish medical texts.

#### 1. Introduction

A number of publicly available (bio)medical corpora and data sets have been come to light during the last years. These corpora vary considerably in size, quality, coverage, encoding and depth of linguistic and structural characteristics. The vast majority are monolingual English corpora, covering different medical genres, while a small number of text collections exist in other major languages, particularly French and German. According to Zweigenbaum et al. (2001), a corpus useful for testing and training biomedical Natural Language Processing (bio-NLP) tools must account for the variety of medical texts; diversity, (origin, genre etc.), must be obtained in addition to mere volume. Having the above statement in mind, we describe here the design, collection and development of a Swedish medical corpus, the MEDLEX Corpus. MEDLEX consists of a variety of text-documents related to various medical text genres, and it is annotated with different levels of structural and descriptive meta-information as well as linguistic information. All are automatically heterogeneous web pages from the Internet and converted to text files. The corpus consists of: teaching material, guidelines, official documents, scientific articles from medical journals, conference abstracts, consumer health care documents, descriptions of diseases, definitions from on-line dictionaries, patient's FAQs etc. 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.

This paper starts by providing some background notes on corpus collection and processing of biomedical

corpora, Section 2. Section 3 provides the characteristics and details of the content of the MEDLEX Corpus. In Section 4 we describe the various annotations added to the corpus and provide some evaluation figures for some of the processing stages. In Section 5 we discuss how the results from the previous processes have been integrated into the parser, the parsing mechanism, as well as the conversion of the parser's output into the TIGER-XML format. Finally, conclusions and suggested ways to improve the various processes, the potential use of the corpus and directions for future research end the paper.

## 2. Background

The far more cited corpora within the area of biomedicine from PubMed Central the (www.pubmedcentral.org) and the MEDLINE database. Particularly, the GENIA corpus (Kim et al., 2003), which consists of 2000 MEDLINE abstracts, has been used in many bio-NLP related activities; e.g. Yakushiji et al. (2001), by applying a full parser for the extraction of argument structures. Other corpora include the GENETAG corpus (ftp.ncbi.nlm.nih.gov/pub/tanabe) Tanabe et al, 2005; the Yapex (www.sic.se/humle/projects/prothalt) et al. (2002)and the PennBioIE (bioie.ldc.upenn.edu) Bies et al. (2005). For a survey on some of the available linguistically processed (tagged) biomedical corpora see Cohen et al. (2005). Other known, but varying in length, annotation and availability bio-medical corpora, include the large corpus of electronic articles in the cardiology domain, 85 million words, by Teufel & Elhadad (2002); the smaller, 100,000 tokens of German clinical reports by Wermter & Hahn (2004), and the MUCHMORE (muchmore.dfki.de/resources/index.htm) parallel corpus of English-German scientific medical abstracts obtained from the Springer Link web site, one million tokens for each language.

# **3.** The MEDLEX Corpus

#### 3.1 Collection

The starting point for the collection of the MEDLEX corpus was to navigate the Internet and manually inspect through links or simple keywords, using own Google searches as input, a number of potentially relevant web-sites. However, during later progressing of the corpus collection and in order to reduce the time spend for finding new sites, we found useful to apply the BootCaT toolkit by Baroni & Bernardini (2004). BootCaT implements an iterative procedure to bootstrap specialized corpora and terms from the web, requiring a list of "seed terms", typical of the domain of interest. MEDLEX does not focus at a particular subfield due to the lack of very large resources within a particular subarea for Swedish, and as such, the texts range through many sub-domains, genres and specialised topics, including pharmacology. Thus, the target has been to locate a large number of potentially interesting, though heterogeneous, medically oriented Swedish sites and extract, as automatically as possible, textual information from them (e.g. by using Lynx and the perl LWP library).

#### 3.2 Content Extraction

Most of the sampled documents were in (X)HTML format, while there were a number of documents in PDF and MS Word. However, various types of (X)HTML mark-up or other types of formatting information were removed from the original files and the whole body of texts was converted into text format (ISO Latin1). Important structural meta-information, particularly the title, the source of origin and the date of publication of each article (if available) were preserved.

The corpus includes: scientific articles from medical journals, teaching material, guidelines, official documents, conference abstracts, consumer health care documents, descriptions of diseases, definitions from on-line dictionaries/glossaries, patient's FAQs, editorial articles and encyclopaedic information.

## 3.3 Status of the MEDLEX Corpus

By the end of 2005, the MEDLEX Corpus was comprised of approximately 6 millions tokens, over 10,000 different articles. The corpus size is based on tokenized and free of formatting information text material. The MEDLEX is encoded using basic annotation featuress from the XML Corpus Encoding Standard (XCES); Ide et al. (2000). The corpus consists of text samples from a number of different text genres<sup>1</sup> and Table 1 provides some typological details

<sup>1</sup> E.g, the list of the sites of pharmaceutical companies include:

Astra Zeneca; http://www.astrazeneca.se Bristol-Myers Squibb; http://www.bms.se

Eli Lilly; http://www.lilly.se

GlaxoSmithKline; http://www.glaxowellcome.se

Ipex; http://www.ipex.se

Leo Pharma Nordic; http://www.leo.se Merck Sharp & Dohme; http://www.msd.se

OrionPharma; http://www.orion.se Pfizer; http://www.pfizer.se Recip; http://www.recip.se

Schering-Plough; http://www.clarityn.com St Jude Medical Inc; http://www.sjm.se

Wyeth; http://www.wyeth.se

of its content.

Genre	#tokens /types
journals and periodica	1,6 milj./97,000
faculties, institutes and hospitals	305,000/27,000
health-care communication companies	1,7 milj/102,000
specialised sites	890,000/61,000
conference proceedings (abstracts)	320,000/41,000
pharmaceutical companies	300,000/27,000
media (TV, daily newspapers)	600,000/57,000

Table 1: Distribution of tokens/types in MEDLEX

## 4. Structural & Linguistic Annotation

The MEDLEX corpus was linguistically processed in the lines of other similar text collections (*cf.* Teufel & Elhadad, 2002).

## 4.1 Tokenization and Part-of-Speech Tagging

MEDLEX was firstly tokenized and segmented into sentences, using regular expressions. Tokenization and sentence segmentation of the texts into individual words and sentences is an important step since all other processing steps that follow depend on it. Spelling mistakes, "spurious" quoting or other erroneous linguistic phenomena where not corrected. Since some of the texts were of very special nature, we had to make certain modifications to our generic tokenizer in order to handle special cases that were erroneously tokenised. Some additional tokenization steps included adaptations for handling particular constructions, such as "[...] ämnet NKK (4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone) i urinen [...]".

After tokenization, the corpus was annotated with part-of-speech labels using a version of Brills tagger (Brill, 1994). The morphosyntactic annotation uses the Swedish MULTEXT tagset. Medical language as any other technical language exhibits some characteristics that differentiate it from general language e.g. sentence construction rules slightly differ from the normal constructions rules, use of idiosyncratic expressions, medical jargon of many terms etc. Therefore, the tagger's lexicon was completed with several hundreds of new entries of medical jargon, such as adjectives, particularly of Greek and Latin origin, and nouns, particularly:

- names of diseases, (e.g. ending in "-it"/"lymfadenit") where the tagger annotated them as verbs (a common verb suffix in Swedish)
- common drug names and chemical substances (e.g. ending in "-das"/"aminoxidas") where the tagger once again annotated them erroneously as verbs (also a verb suffix in Swedish)
- (patho)anatomical terms (e.g. ending in "-is"/"pulmonalis")
- frequent in the corpus words (compounds) ending in "es" or "ns": "insufficiens", "inkontinens", "pares" and "fluorescens" which were erroneously annotated as genitive nouns.

After the part-of-speech tagging, the annotated instances, nouns, verbs and adjectives, were lemmatized using

finite-state machines. The input to lemmatization is tokens with their respective part-of-speech. Lemmatization provides direct access to all instances of a lexical entity via its base form.

#### **4.2** Generic Named Entities

There is a whole range of named entities (NE) that can be encountered in various types of texts, and not only the "classical", in the NE recognition bibliography, types of named entities, i.e. person, location and organization. Along the line proposed by Sekine (2004), we also 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 unambiguous usages for assigning annotations in ambiguous words<sup>2</sup>
- lists of single names (approx. 100,000)
- a theory revision and refinement module makes 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 (Kokkinakis, 2005). The evaluation figures for each entity group ranged between 69,9%-100% precision and 66%-98% recall.

During 2005 (in parallel with the implementation of a Swedish MeSH tagger, see next section), the system was extended with capabilities of recognizing and annotating medical terminology (a ninth NE category). This terminologically oriented module comprises seven subcategories, namely: "Anatomy (MDA)", "Living Organisms (MDO)" mainly names of viruses and bacteria, "Diseases (MDD)", "Chemicals, Vitamins, Enzymes and Drugs (MDC)", "Symptoms (MDS)", "Pharmaceutical Equipment/Devices (MDP)" and "Analytical, Diagnostic

and Therapeutic Techniques (MDI)". All of the resources used by the system have been either obtained from various medically-related internet sites or extracted from the MEDLEX Corpus. Some of the lexical resources we currently use, include parts of the Swedish ICD-10, the Swedish MeSH and a list of several thousands of drugs from the Swedish union of pharmaceutical industry FASS, (www.fass.se). Apart from the previous, we use morpheme lists consisting of a couple of hundreds of elements, mostly suffixes, specific to the various categories of the medical domain. For instance, in the "Diseases" module there are suffixes such as: "+opati", "+iasis", "+emi", "+itsis", "+ism" and "+uri".

## 4.3 MeSH Terminology

During the second half of 2005, we started the implementation of a tagger based on the Swedish translation of the MeSH thesaurus (hierarchical terminology), after obtaining a license by the University Library of the Karolinska Institutet. The MeSH® (Medical Subject Headings) 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 Inst. Lib. based on the year 2006 MeSH. MeSH is used for subject analysis of biomedical literature; for more information visit: mesh.kib.ki.se/swemesh/swemesh.cfm. The implementation of the tagger uses 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, since we attempt to disambiguate the MeSH annotation (see later this Section), we decided to only use the upper level (level 0) of the lexical hierarchy for the classification of each term<sup>4</sup>. For instance, the term beta-Lactamases (betalaktamaser) has the label D08.811.277.087.180 which was reduced to D08 [Enzymes and Coenzymes]. By this reduction, we assume that terms having multiple labels under the same level are very similar.

However, prior to the implementation of the MeSH tagger we applied a number of conversion and normalization steps to the original material. For instance, changing the order of the head and modifier complements for several hundreds of entries in the database (e.g. "vacc-

<sup>3</sup> Symptoms are usually realized in Swedish texts either as

hand constructed rules using regular expressions with the data

1202

<sup>&</sup>lt;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.

periphrastic expressions or compounds and it is rather difficult to find suitable lexical resources on the Internet in order to simply apply some sort of dictionary lookup. Therefore, we have investigated the way these expressions are constructed in MEDLEX, by initially selecting a few characteristic symptom key-words and short phrase fragments and then create and apply regular expressions on an analysed version of the corpus. This way we could identify new symptoms in the near vicinity of the already matched ones and iteratively implement a new set of

gathered by this process.

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

```
Han och kollegan Riitta Ylitalo och reumatologen Mikael Heimbürger , har studerat <mesh tag="A04">stämbanden</mesh> hos 14 patienter med olika typer av <mesh tag="C20">autoimmuna sjukdomar</mesh> - <mesh tag="C05/C17/C20">reumatoid artrit</mesh> , <mesh tag="C17/C20">systemisk lupus erytematosus</mesh> och <mesh tag="C07/C11/C17/C20">Sjögrens syndrom</mesh> .

Han och kollegan <ENAMEX TYPE="PRS" SBT="HUM">Riitta Ylitalo</ENAMEX> och reumatologen <ENAMEX TYPE="PRS" SBT="HUM">Mikael Heimbürger</ENAMEX> , har studerat <ENAMEX TYPE="TRM" SBT="MDA">stämbanden</ENAMEX> hos <ENAMEX TYPE="PRS" SBT="CLC">14 patienter</ENAMEX> med olika typer av autoimmuna sjukdomar - <ENAMEX TYPE="TRM" SBT="MDD">reumatoid artrit</ENAMEX> , <ENAMEX TYPE="TRM" SBT="MDD">systemisk lupus erytematosus</ENAMEX> och <ENAMEX TYPE="TRM" SBT="MDD">systemisk lupus erytematosus</ENAMEX> och <ENAMEX TYPE="TRM" SBT="MDD">sjögrens syndrom</ENAMEX> .
```

Figure 1: NER and MeSH annotations

iner, orala" changed to "orala vacciner"); normalizing all inflected/non-inflected entries into a neutral non-inflected variant with optional inflected variants (e.g. "orala vacciner" changed to "oral(a)? vaccin(et/er/erna)?"); making adjustments to the implemented recognizer in order to capture some frequent phenomena of misspellings, agreement and grammatical errors that we could observe in annotated sample texts, probably caused due to the high variability in the expression of similar concepts by different authors (e.g. "escherichia coli" added "e. coli") or by the influence or "contamination" from the English language, particularly orthographic variation (e.g. use of 'ph' instead of 'f'). Finally, case folding was applied to all terms, except those consisting of uppercase letters, particularly acronyms. This was necessary in order not to introduce new forms of ambiguity, since the complete elimination of case information could introduce new ambiguities between homographs, i.e. uppercase/low case words. For instance, "kol/D01.268.150" (i.e. "carbon") and "KOL/C08.381. 495.389" (i.e. "Chronic Obstructive Pulmonary Disease"). An example sentence illustrates the annotation provided by both the generic NER (see previous section) and the MeSH annotator, for the sentence: "He and his colleague Riitta Ylitalo and rheumatologist Mikael Heimbürger have studied the vocal cords of 14 patients with different types of auto-immune diseases - reumatoid arthritis , systemic lupus erythematosus and Sjogrens syndrome . (Figure 1).

Since the MeSH annotation is an ongoing activity and we have not any manually annotated material at our disposal at the time of writing the paper, we cannot provide any evaluation figures for its performance. However, we have experimented with the application of unsupervised disambiguation of the MeSH terms based on the principle of "one sense per discourse" statement by Gale et al. (1992). Thus, the current implementation a partial and simplistic disambiguation methodology in lack of suitable training material. We observed, therefore, that in many cases the unambiguous readings within the same document, can help disambiguating the meaning of an ambiguous term (relationship ambiguity). The results (Table 2) were obtained w.r.t. ambiguity reduction, for complete matches, on a sample of 10 articles, 8,490 tokens, from the Swedish Medical Association's magazine (www.lakartidningen.se).

For instance, the fragment "...lokalanestetikum i inhalation (lidokain, bupivakain) kan blockera symtomen", i.e "...local anesthetics during inhalation (Lidocaine, Bupivacaine) can block the symptoms", is an-

notated by the MeSH tagger as "... lokalanestetikum i inhalation (<mesh tag="D02">lidokain</mesh>, <mesh tag="D02/D03">bupivakain </mesh>) kan blockera symtomen"; that is "lidokain" is annotated as D02[Acetanilides] and "bupivakain" as D02[Acetanilides] and D03[Pipecolic Acid]. Thus, according to the principle of one sense per discourse, near unambiguous neighbours can in many cases disambiguate their ambiguous counterparts. In the previous example, the annotation of "bupivakain" will be reduced to D02 which is actually the preferred meaning. (the system also adds a "reliability" attribute to the disambiguated annotation, which indicates the strength of the confidence for the ambiguity elimination).

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

Table 2: Ambiguity reduction of MeSH terms

## **5. Cascaded Parsing of MEDLEX**

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*.

Cass applies 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 can utilize 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 reducing ambiguity at the various levels of the linguistic processing<sup>5</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*.

<sup>&</sup>lt;sup>5</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.

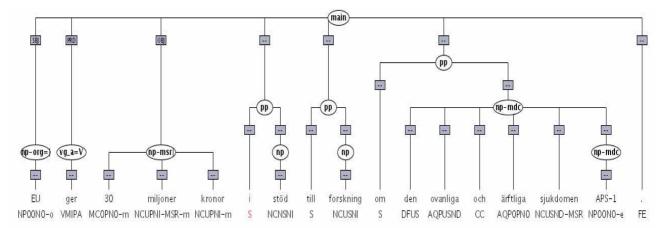


Figure 2. Annotation of the sentence: *EU gives 30 million* (*Swedish*) *krona to research on the rare and hereditary disease APS-1*." (the annotation also shows functional tags; "SBJ"[subject], "PRD" [predicate] and "OBJ" [object]).

The phrases 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 parsing involves a cascade of two major automata, the "phrasal" and the "clausal". The "phrasal" includes: phrases which include a named-entity annotation; phrases which do not include a named-entity annotation; adjectival phrases; prepositional phrases and verbal groups/chains. 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 different 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 any available lexicalized complementizer or part-of-speech tags that can provide strong evidence for a particular type of clause. For an evaluation of the parser's performance on non medical texts see Kokkinakis & Johansson Kokkinakis (1999).

Finally, for the annotation scheme of the parsed output we have chosen the TIGER-XML encoding format (König & Lezius, 2003), since Cass doesn't provide any visualization capabilities or means of correcting the syntactic/semantic analysis (see Figure 2). TIGER XML is 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 combing lexical features, semantic annotations and phrase labels. Moreover, TIGER-aware tools such as @nnotate (Plaehn & Brants, 2000) and SALSA (Erk & Pado, 2004) provide the means for modifying, correcting and enhancing both the syntactic and semantic analysis returned by Cass. At the moment the corpus is only available in the form of searchable concordances using the IMS Corpus Workbench, Figure 3; (Christ, 1994).

#### 6. Conclusions

In this paper we have described the structure and content of a Swedish medical corpus, the MEDLEX Corpus. A resource that can be useful in many areas of Swedish medical informatics and Natural Language Processing. MEDLEX comprises over 10,000 articles, over 6 million tokens, annotated in many dimensions. We currently investigate different ways to make parts or the whole material copyright-free and hence make it available for further research. The lack of resources for linguistic processing of biomedical corpora can be a major obstacle in healthcare informatics, because it prohibits access to the information treasure in biomedical corpora for lesser spospoken languages, slowing down the promotion of healthcare-related technologies.

As stated earlier, the majority of corpora in the medical domain are in English, a phenomenon easily motivated by the fact that biomedicine is a field with global interest and researchers prefer to write in a language understandable by a wide audience; Teufel and Elhadad (2002) mention that in cardiology alone there are at least 700 journals. The situation in Sweden is not that bad, although it cannot reach the volumes written in Engish. For instance, there are a handful of scientific journals that do not cover a particular subfield but rather incorporate various new findings from diverse medical subfields.

Computer understanding of the underlying language is a far from trivial task that involves several layers of knowledge intensive processing and we have described some of the possible means for corpus enhancement. There are several issues that need to be investigated in more depth and possibly in a larger scale. For instance, the use of a human in the process loop, in order to inspect intermediate results; and the use of tools such as @nnotate for correcting or enhancing the syntactic analysis. The need to conduct an evaluation on a larger material, 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 noun phrases. For the coverage of MeSH, a trained physician would have been the right person to mark unlabelled terminology; however, we could not consult such a person at the time of this study. 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 (including synonyms), but it still does not cover all clinically useful terminology and empirical studies can be of benefit for its content's growth.

Pattern ↓	Right context
et A-vitamin ger benskörhet	Höftfrakturer . Sju gånger högre risk för höftfraktur orsakad
et A-vitamin ger benskörhet	- även hos män  För mycket A-vitamin orsakar benskörhet
et A-vitamin orsakar benskörhet	och höftfrakturer , visar en ny svensk studie . Forskare
att akrylamid ger cancer	∕title> Även måttliga doser av akrylamid har en cancerframkallande effek
m akrylamid orsakar cancer	eller ej , finns dessutom andra , tyngre skäl att
va cox-2-hämmare lindrar inflammation	ungefär som övriga NSAID men ger färre allvarliga gastrointestinala (
<sup>la</sup> kortisonet dämpar eksemet	så att huden kan läka , och det är en
år pravastatin förebygga inflammation	och avstötning av langerhanska cellöar i samband med transplantation [
m testosteron stimulerar prostatatumörer	, har det länge funnits misstankar om att höga nivåer
at tramadol bedömdes förgiftningen	som allvarlig i tio procent av fallen . Drygt en
	Pattern ↓  A-vitamin ger benskörhet  A-vitamin ger benskörhet  A-vitamin orsakar benskörhet  akrylamid ger cancer  akrylamid orsakar cancer  va cox-2-hämmare lindrar inflammation  da kortisonet dämpar eksemet  år pravastatin förebygga inflammation  m testosteron stimulerar prostatatumörer  at tranadol bedömdes förgiftningen

Figure 3. Concordance results obtained by querying the MEDLEX corpus with the combination of morphosyntactic and semantic features, namely the pattern "medical substance + verb + disease".

## 7. Acknowledgements

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

## 8. References

- 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.
- 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. Vossen P. et al. (eds), pp. 71-77. Spain.
- Baroni M. and Bernardini S. (2004). *BootCaT: Bootstrapping corpora and terms from the web*. Proc. of the 4th Language Resources and Evaluation Conference (LREC). Lisbon.
- Bies A., Kulick S. and Mandel M. (2005). Parallel Entity and Treebank Annotation. ACL 2005 workshop on "Frontiers in Corpus Annotation II: Pie in the Sky". Michigan.
- Brill E. (1994). Some Advances in Transformation-Based Part of Speech Tagging. Proc. of AAAI '94. Seattle.
- Christ O. (1994). A Modular and Flexible Architecture for an Integrated Corpus Query System. COMPLEX'94. Hungary.
- Cohen K.B., Fox L., Ogren P.V. and Hunter L. (2005). *Corpus Design for Biomedical Natural Language Processing*. Proc. of the ACL-ISMB Workshop on Linking Biological Literature, Ontologies and Databases: Mining Biological Semantics. Pp 38-45. Detroit, US.
- Erk K. and Pado S. (2004). A Powerful and Versatile XML Format for Representing Role-Semantic Annotation. Proc. of the 4th Language Resources and Evaluation Conference (LREC). Lisbon.
- Franzén K. et al. (2002). *Protein Names and How to Find Them*. International Journal of Medical Informatics, Special Issue on NLP in Biomedical Applications.
- Gale W., Church K.W. and Yarowsky D. (1991). One Sense per Discourse. Proc. of the DARPA Speech and Natural Language Workshop. New York.
- Ide N., Bonhomme P. and Romary L. (2000). XCES: An XML-based Encoding Standard for Linguistic Corpora. Proc. of the 2nd International Language Resources and Evaluation Conference. Pp. 825-30. Athens, Hellas.
- Kim1 J.-D., Ohta T., Tateisi Y. and Tsujii J. (2003) GENIA Corpus - a Semantically Annotated Corpus for Bio-textmining

- BIOINFORMATICS. Vol. 19 Suppl. 1, Pp. 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. Lisbon.
- 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 S. Johansson-Kokkinakis. (1999). A Cascaded Finite-State Parser for Syntactic Analysis of Swedish. Proc. of the 9th European Chapter of the ACL (EACL). Pp. 245–248, 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, University of Stuttgart.
- Mikheev A., Moens M. and Grover C. (1999). *Named Entity recognition without gazetteers*. Proc. of the 9th European Chapter of the ACL (EACL). Pp. 1-8. Bergen, Norway.
- Müller F.H. (2004). Annotating Grammatical Functions for German Using Finite-Stage Cascades. Proceedings of the COLING. Pp. 268-274. Switzerland.
- Plaehn O. and Brants T. (2000). Annotate -- An Efficient Interactive Annotation Tool. Proceedings of the Sixth Conference on Applied Natural Language Processing (ANLP) Seattle.
- Rosario B., Hearst M.A. and Fillmore C. (2002). *The Descent of Hierarchy, and Selection in Relational Semantics*. Proc. of the 40th Annual Meeting of the Association for Computational Linguistics (ACL). Pp. 247-254. Philadelphia.
- Sekine S. (2004). *Definition, dictionaries and tagger for Extended Named Entity Hierarchy*. Proc. of the Language Resources and Evaluation Conference (LREC). Lisbon.
- Tanabe L. et al. (2005). GENETAG: a Tagged Corpus for Gene/Protein Named Entity Recognition. BMC Bioinformatics, 6 (Suppl 1).
- Teufel S. and Elhadad N. (2002). *Collection and linguistic processing of a large-scale corpus of medical articles*. Proc. of the Language Resources and Evaluation Conference (LREC). Las Palmas.
- Wermter J. and Hahn U. (2004). An Annotated German-Language Medical Text Corpus as Language Resource. Proc. Of the 4th Conference on Language Resources and Evaluation (LREC). Lisbon.
- 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
- Zweigenbaum P. et al. (2001). *Building a text corpus for representing the variety of medical language*. Proc. of the 10 World Congress on Medical Informatics, London.