MedLex+: An Integrated Corpus-Lexicon Medical Workbench for Swedish

Dimitrios Kokkinakis
Maria Toporowska Gronostaj
Göteborg University

This paper reports on the work carried out developing MedLex+, a medical corpus-lexicon workbench for Swedish. This project, which is still under active development, has been going on for some years now within the Department of Swedish language at Göteborg University. At the moment, the workbench incorporates:

- *an annotated collection of medical texts-including 20 million tokens and 45,000 documents,*
- *a number of language processing software programs, including tools for collocation extraction, compound segmentation and thesaurus-based semantic annotation, and*
- *a lexical database of medical terms-containing 5,000 medical entries. MedLex+ is a multifunctional lexical resource due to a structural design and content which can be easily queried. The medical workbench is intended to support lexicographers compiling lexicons and also lexicon users more or less initiated in the medical domain. MedLex+ can also assist researchers working on either lexical semantics or natural language processing (NLP) applications with focus on medical language. The linguistically and semantically annotated medical texts in combination with a set of smart queries turn the corpora into a rich repository of semasiological and onomasiological knowledge about medical terms and their linguistic, lexical and pragmatic properties. These properties are recorded in the lexical database with a cognitive profile. The MedLex+ workbench seems to offer a constructive help in many different lexical tasks.*

1. Introduction

The medical corpus-lexicon workbench for Swedish (henceforth MedLex+) has evolved from work on the project *Active Lexicon* conducted at the University of Gothenburg, Department of Swedish Language and at the *Semantic Mining in Biomedicine*[1] a research collaboration project in the European Network of Excellence programme. While the first project dealt with compiling an electronic lexicon, aimed to support non-native speakers of Swedish expected to work in the Swedish health care sector, the second one was mainly language technology oriented. Medical texts collected in the course of these two projects and the lexical resources developed therein, in conjunction with the gained insights on the state of the art concerning Swedish medical lexicons, have been the main incitements for building MedLex+. The work on MedLex+ is still progressing.

MedLex+ is a multifunctional lexical resource due to its structural design and content which can be easily queried. It is intended to support lexicographers in their work on compiling lexicons as well as lexicon users in a familiar look-up of words and even more advanced information searches. Furthermore, it can serve a wide group of researchers working in the fields of lexical semantics and natural language processing (NLP) applications with focus on medical language. To be able to fulfil these tasks fully MedLex+ needs to be further expanded and

[1] The long-term goal of *Semantic Mining in Biomedicine* (EU Network of Excellence, 507505) was the development of generic methods and tools supporting the critical tasks of the field; data mining, knowledge discovery, knowledge representation, abstraction and indexing of information, semantic-based information retrieval, and knowledge-based adaptive systems for provision of decision support for dissemination of evidence based medicine. More information can be found at *http://semanticmining.coling-uni-jena.de/*.

this is our main objective in our current work and future plans. There is no doubt that an effective expansion of the workbench depends to a large extent on mutual feedback between, on one hand, domain-dependent text mining research activities and applications that require such type of input and, on the other hand, the growing content of corpus-based medical lexicons.

At the moment MedLex+ incorporates: (1) an annotated collection of medical texts, (2) a number of language processing modules including tools for collocation extraction, compound segmentation and thesaurus-based semantic annotation and also (3) a constantly growing lexical database of medical term, henceforth referred to as MedLex lexicon. The MedLex+ environment supports querying the domain corpus by a combination of multiple search criteria. For that purpose we use as a backend the IMS Corpus Workbench (Christ 1994), while results are displayed in a concordance format, familiar for lexicographers. It is assumed here that an extensive domain corpus and intelligent corpus tools are a prerequisite for further elaboration of multifunctional lexicons and their morphological, syntactic, semantic and thematic modules, hence bridging the gap between general language and medical language resources.

The remaining of this paper is structured in the following way. Section 2 reports on corpus related work. We provide a number of quantitative and qualitative characteristics of the corpus, its structural and linguistic annotation and also show how the linguistic features enable non-trivial access to the data in the form of *semantic concordances*. Section 3 is devoted to the description of the content of the medical lexicon and its database structure. We conclude with some comments on the advantages of the workbench and its further development.

2. The MedLex corpus

With the information overload in the life sciences, there is an increasing need for corpora, preferably annotated (e.g. with biological entities), which is the driving force for data-driven language processing applications and the empirical approach to language study. The GENIA corpus (Kim et al. 2003) is one such corpus, extensively used in biomedical research. GENIA is a manually curated corpus of 2000 MEDLINE/PubMED abstracts with over 100,000 annotations for biological terms.

In order to meet the needs of our research within the language technology field and also support activities in lexicography and terminography,[2] we have collected a Swedish medical corpus, MedLex, the first large, structurally and linguistically annotated, Swedish medical corpus (cf. Kokkinakis, 2006). The MedLex corpus consists of a variety of text-documents related to various medical subfields. It does not focus on a particular medical genre, primarily, due to the lack of very large Swedish resources within a particular specialized subdomain. All text samples, currently 25 million tokens (50,000 documents) are fetched from web pages during the past couple of years. The MedLex corpus includes 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 and blogs etc. All texts have been converted to text files and have been both structurally and linguistically annotated. The structural characteristics of each document are preserved using a simple eXtensible Markup Language (XML) scheme. The scheme covers the source of origin (web address) of each document, its main heading and the date of its publication (where possible). We are currently investigating means to adapt this simple structure as well the linguistic annotation (next section) to a standardized format such as the TEI-lite encoding scheme (Burnard, Sperberg-McQueen 2006), the XCES (Corpus Encoding Standard) or the DocBook specifications.

[2] Here, *lexicography* is understood as the process of making dictionaries of general-language words while *terminography* is concerned exclusively with compiling collections of the vocabulary of special languages.

2.1. *Linguistic annotation*

The linguistic annotation of the MedLex corpus consists of a series of lexical and shallow[3] semantic pieces of meta-information automatically added to each document in the collection. Most of these are provided by *domain-independent* tools which have, at least in some cases, undergone several adaptations to the language of the domain. The tools can be roughly divided into three main groups: (1) in the first annotation layer, there are tools that deal with the processing and transformation of the physical structure of a document, (2) in the second annotation layer, there are tools performing shallow semantic analysis, while (3) in the third annotation layer, there are tools that work on the morphosyntactic characteristics of the texts.

Thus text documents pass a basic and modular processing pipeline which starts by converting each document, which can be of any format (e.g. .pdf, .doc, .html, .xsl), to text format. These files are then tokenized, undergo zoning (sentence and paragraph splitting), part-of-speech annotation (including the recognition of multi-word expressions), lemmatization and shallow parsing.

For the part-of-speech annotation, we use the TnT tagger (Brants 2000) and the Swedish MULTEXT tagset (*http://spraakbanken.gu.se/parole/tags.phtml*). Medical language, as any other specialized language, presents some characteristics that differentiate it from general language on the sentence and/or lexical level, namely use of idiosyncratic expressions, medical jargon, terminology etc. Since the tagger is not trained on texts from the medical domain, its lexicon has been enhanced with medical terminology, roughly 20,000 new entries, in particular nouns such as names of diseases, common drug names and chemical substances. For instance, special attention has been paid to nouns that end in "-it" *lymfadenit* (lymphadenitis) and "-as" *aminoxidas* (amine oxidase), because the general language part-of-speech tagger erroneously annotated such forms as verbs. The partial syntactic analysis is based on the Cass-parser, *Cascaded analysis of syntactic structure* and originates from the work by Abney (1997). 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. The Swedish grammar used by the parser has been developed by Kokkinakis and Johansson Kokkinakis (1999). It has been adapted and modified to make the parser aware not only of the features of the morphosyntactic description but also of the ones provided by the shallow semantic analysis.

The shallow semantic annotation is based on a set of thematic tags taken from the Medical Subject Headings thesaurus (MESH® *http://mesh.kib.ki.se/*). MeSH is the controlled vocabulary thesaurus of the NLM (U.S. National Library of Medicine) and has been used for different purposes in a variety of applications and settings, primarily for indexing and organization of bibliographic citations (medical literature). The semantic tags referring to subdomains dealing with *anatomy*, *organisms*, *diseases*, *psychiatry and psychology*, *chemicals and drugs* as well as to *diagnostic and therapeutic techniques and equipment*, being some of the top nodes in the MeSH thesaurus, expose the knowledge-based medical facts which are in several respects crucial for the lexicographer's work. This purely medical information is further extended with more general semantic information, as generic named entity recognition is performed on the texts providing tags referring to eight main types of entities: *person*, *location*, *organization*, *object/artifact*, *event*, *work*, *time* and *measure expressions* (cf. Kokkinakis 2004). 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. This combination of semantic information makes it easier to access the higher level construction and to view semantic patterns in the corpus (see section 2.2 and Figure 2). It also supports: (1) disambiguation and defining of polysemous medical terms, (2) detection of meaning-sensitive collocational patterns, and (3) recognition and retrieval of relevant medical multi-word terms.

The IMS Corpus Workbench is a set of tools for the manipulation of large, linguistically annotated text corpora and it is used as a backend for the corpus processing and the querying of the MedLex

[3] We call this layer *shallow* because only a subset of the vocabulary in the text is subjected to semantic annotation.

corpus. All linguistic annotations previously outlined are integrated into a uniform representation which facilitates the querying of the metadata in a flexible manner. The results are displayed in a traditional KWIC-concordance format. A characteristic of IMS-CW is its ability to efficiently handle an unrestricted number of attributes per corpus position as well as the use of regular expressions over attribute values.

2.2. *Case example: Semantic concordances*

There have been a number of approaches trying to organize corpora in meaningful and sophisticated manners with focus on easy access to linguistic evidence and to patterns of interest. Miller et al. (1993) discuss the notion of "semantic concordance", that is, a textual corpus and a lexicon combined in such a way that every substantive word in the text is linked to its appropriate sense in the lexicon. Another approach, but without any use of semantic information, is the work by Tapanainen and Järvinen (1998), in which they describe a tool producing and utilizing functional syntactic information which is presented as mutual dependencies between lemmatized words.

Even though Miller's semantic concordance is not yet quite feasible at the current stage of the development of MedLex+, we have put a great effort into semantically annotating as many words as possible that (might) indicate medical notions in the MedLex corpus and also into providing basic morphosyntactic annotations to all tokens in the corpus as previously described. The semantic annotation with the tagset corresponding to the top nodes in the MESH thesaurus has been refined in the course of the work, as the tagset has been enhanced with tags from the lower nodes in the MESH hierarchies as well as from a number of specialised medical lexical resources obtained from the Internet, including medical eponym lists and sub-domain specific word lists. Figure 1 illustrates the results of the query "[lem='.*barn'][]{1,5}[pos='V.*'][word='i|av'][]{1,2}[sem='MESH-C*']" which combines lexical, morphosyntactic and semantic criteria. The query can be paraphrased as: "get KWIC contexts that start with the lemma *barn* (child) possibly as the head of a compound, followed by 1 to 5 tokens (roughly words) and then by any verb form which is followed by the word *av* (of/by) or *i* (in), one or two words and the semantic annotation *MESH-C* (disease or symptom)".

LEFT Context	KWIC	RIGHT Context
tal absolut sett - omkring 80	**barn och 1 200 vuxna drabbas av dessa tumörer**	varje år . Anette Bromert , b
de mot 21 förväntade . Antalet	**barn som drabbats av den svåra RSV-smittan**	har sjunkit dramatiskt jämför
viteter som underlättar för de	**barn som redan lider av astma , allergier**	eller annan överkänslighet .
räldrarna lever med risken att	**barnet kan dö av sin missbildning**	. Barnets nära relationer med
t gäller ju i princip bara små	**barn som kan smittas av någon med bältros**	. - De flesta är positiva til
liga förbättringen ser man hos	**barn som drabbats av lymfatisk leukemi**	. Den mjuka madrassen ger en
vårt att på förhand säga vilka	**barn som riskerar att drabbas av allvarlig sjukdom**	och i värsta fall dö på grund
nte var rätt väg eftersom även	**barn utanför idrottsrörelsen drabbats av plötslig hjärtdöd**	. Då sattes hon upp på listan
ör tror och hoppas vi att inga	**barn ska behöva födas i Sverige med hiv-infektion**	i framtiden . Därför var det
är det ungefär hälften av alla	**eksembarn som blivit av med sina eksem**	i 13-årsåldern . Enligt Mona
e även andra defekter . Ett av	**barnen som behandlats i Frankrike fick leukemi**	, som ansågs ha samband med g
id Göteborgs universitet . Ett	**barn hade drabbats av cerebral pares**	följt av spastisk hemiplegi .
flora i tarmen hos nyfödda och	**barnens risk att senare i livet drabbas av atopiska eksem**	. Forskare vid Stanford Unive
LL . Fyra personer , varav tre	**dagisbarn , har insjuknat i smittsam hjärnhinneinflammation**	i Göteborg under den senaste
die som visat goda resultat på	**barn som drabbats av återfall i sjukdomen**	eller inte svarat på annan be
. Granskningen visade att tre	**barn hade dött och fem troligen drabbats av allvarliga hjärnskador**	på grund av missbedömningar m
biofilm finns i mellanörat hos	**barn som lider av kronisk öroninflammation**	. I en större epidemiologisk
ändandet av kortisonsalvor som	**barn leda till att man aldrig blir av med sina eksem**	eftersom huden blir känsligar
va dö pga sjukdomen . Många av	**barnen är undernärda de lider av parasiter/malaria och diarréer**	. Många av de barnen utveckla
i för behandling av fullgångna	**barn som har drabbats av syrebrist , asfyxi**	, i samband med förlossning .
rekommenderad att vaccinera de	**barn som föds i Sverige mot tbc**	med s.k. BCG-vaccin (Bacillu
starkare kortisonkräm) än att	**barnet ska behöva lida av sin kläda**	och sitt eksem . Själva studi
dubbelt så stor risk än andra	**barn att drabbas av stressbesvär som ont**	i magen och problem med sömne
lvarlig för äldre personer och	**spädbarn , som riskerar att drabbas av svår lunginflammation**	med hög feber och andnöd . Sm
globalt projekt för att stödja	**barn som drabbats av effekterna av AIDS**	. Sommarledigheten är lika vi
ngelmans syndrom . Ungefär ett	**barn på mellan fyrahundra och femhundra drabbas av typ 1-diabetes**	, enligt Novo Nordisks egna s

Figure 1. Towards a semantic concordance

The results of the query cumulate in a significant way grammatical, lexical and encyclopaedic knowledge and pave the way for more advanced text exploration techniques. Thus the above concordance lines with explicitly marked semantic tags for terms referring to diseases help to detect typical medical collocational patterns, like *drabbas av* SJUKDOM (be stricken with ILLNESS) or *lida av* SJUKDOM (suffer from DISEASE) (cf. Sköldberg and Toporowska Gronostaj 2008). Furthermore, the enhanced tagset explicitly displayed, helps to automatically retrieve medical semantic patterns in line with Fillmore's frames (Fillmore et al. 2003). A sample of a semantic concordance with the enhanced tagset is presented in Figure 2 for the verb *operera* (operate, perform surgery). For more details concerning the extended annotation tagset and the advantages of semantic pre-processing of the medical corpora see Borin et al. (2007).

To conclude this section the following advantages of lexical annotation needs mentioning: (1) relevant semantic frames and semantic roles can be acquired from medical corpora with minimal human supervision, which supports detection of semantic and syntactic valency patterns for lexicographic purposes as well as different NLP tasks, including both text understanding and text generation, (2) semantically annotated nouns promote disambiguation of predicates, and (3) corpus based extraction of lexical units carrying related meaning (e.g. *operera bort, avlägsna, ta bort*—to remove—).

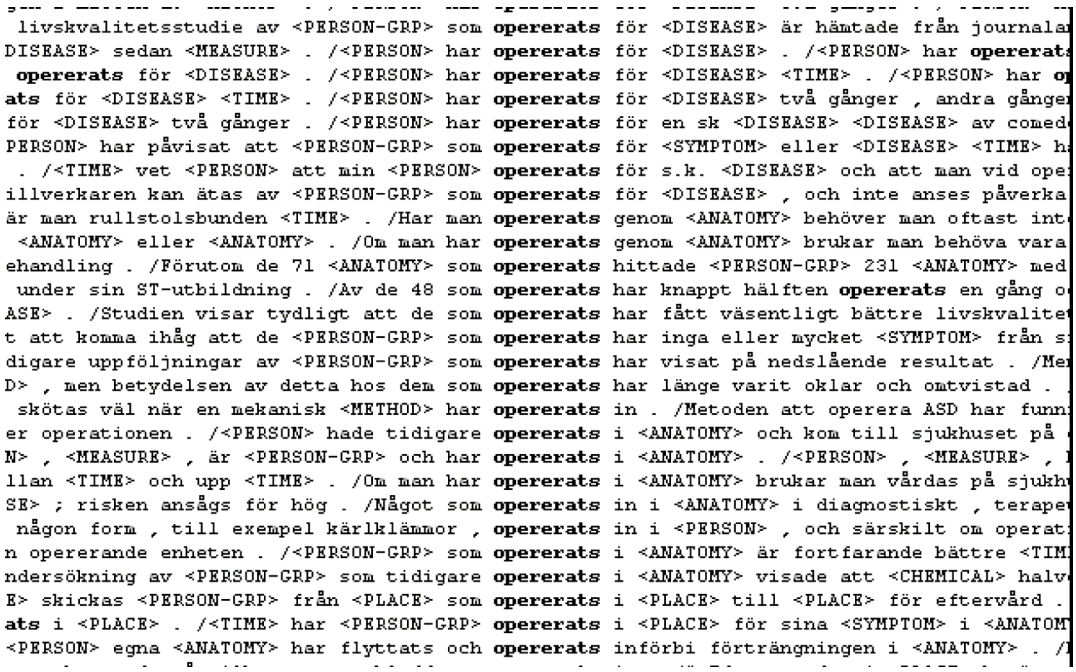
livskvalitetsstudie av <PERSON-GRP> som **opererats** för <DISEASE> är hämtade från journala
DISEASE> sedan <MEASURE> . /<PERSON> har **opererats** för <DISEASE> . /<PERSON> har **opererats**
opererats för <DISEASE> . /<PERSON> har **opererats** för <DISEASE> <TIME> . /<PERSON> har **op**
ats för <DISEASE> <TIME> . /<PERSON> har **opererats** för <DISEASE> två gånger , andra gånger
för <DISEASE> två gånger . /<PERSON> har **opererats** för en sk <DISEASE> <DISEASE> av comed
PERSON> har påvisat att <PERSON-GRP> som **opererats** för <SYMPTOM> eller <DISEASE> <TIME> h
. /<TIME> vet <PERSON> att min <PERSON> **opererats** för s.k. <DISEASE> och att man vid ope
illverkaren kan ätas av <PERSON-GRP> som **opererats** för <DISEASE> , och inte anses påverka
är man rullstolsbunden <TIME> . /Har man **opererats** genom <ANATOMY> behöver man oftast int
<ANATOMY> eller <ANATOMY> . /Om man har **opererats** genom <ANATOMY> brukar man behöva vara
ehandling . /Förutom de 71 <ANATOMY> som **opererats** hittade <PERSON-GRP> 231 <ANATOMY> med
under sin ST-utbildning . /Av de 48 som **opererats** har knappt hälften **opererats** en gång o
ASE> . /Studien visar tydligt att de som **opererats** har fått väsentligt bättre livskvalite
t att komma ihåg att de <PERSON-GRP> som **opererats** har inga eller mycket <SYMPTOM> från s
digare uppföljningar av <PERSON-GRP> som **opererats** har visat på nedslående resultat . /Me
D> , men betydelsen av detta hos dem som **opererats** har länge varit oklar och omtvistad .
skötas väl när en mekanisk <METHOD> har **opererats** in . /Metoden att operera ASD har funn
er operationen . /<PERSON> hade tidigare **opererats** i <ANATOMY> och kom till sjukhuset på
N> , <MEASURE> , är <PERSON-GRP> och har **opererats** i <ANATOMY> . /<PERSON> , <MEASURE> ,
llan <TIME> och upp <TIME> . /Om man har **opererats** i <ANATOMY> brukar man vårdas på sjukh
SE> ; risken ansågs för hög . /Något som **opererats** in i <ANATOMY> i diagnostiskt , terape
någon form , till exempel kärlklämmor , **opererats** in i <PERSON> , och särskilt om operat
n opererande enheten . /<PERSON-GRP> som **opererats** i <ANATOMY> är fortfarande bättre <TIM
ndersökning av <PERSON-GRP> som tidigare **opererats** i <ANATOMY> visade att <CHEMICAL> halv
E> skickas <PERSON-GRP> från <PLACE> som **opererats** i <PLACE> till <PLACE> för eftervård .
ats i <PLACE> . /<TIME> har <PERSON-GRP> **opererats** i <PLACE> för sina <SYMPTOM> i <ANATOM
<PERSON> egna <ANATOMY> har flyttats och **opererats** införbi förträngningen i <ANATOMY> . /

Figure 2. Semantic concordance

3. The MedLex lexicon

The MedLex lexicon is a short name for a multifunctional lexical database with focus on medical language, being one of the modules in the MedLex+ workbench. The content of the lexical database has not been created from scratch, as some lexical data have been retrieved from existing lexical resources, such as a Swedish learners' dictionary, *LEXIN's Svenska ord,* and a general Swedish dictionary, *Nationalencyklopedins ordbok*. The repository of relatively common medical entries from these dictionaries has been enlarged with new ones, retrieved from the MedLex corpus. At the moment, there are about 5,000 medical entries in the MedLex lexicon, capturing a wide spectrum of semasiological and onomasiological properties of the entries. This integrative and more cognitive approach makes the content of the MedLex lexicon unique, as compared to the types of content provided in existing medical databases and dictionaries for Swedish. The cognitive approach enhances the multi-functionality of the database, as it is assumed here that its multi-functionality is to a considerable extent determined

by the scope of its lexical content and the set of available selective query functions. In what follows, we provide a short overview of the content of the MedLex lexicon.

3.1. *Lexicon content*

The MedLex lexicon is basically a monolingual lexical resource for Swedish with some minor bilingual information, in the form of translations of head words to English, and whenever relevant to Latin. The inclusion of English and Latin equivalents reinforces the understanding of the Swedish terms for speakers being in command of English or Latin medical terminology. Access to English equivalents can also support non-native speakers of English in some simple production or reception tasks and it can also be used for cross language information retrieval. The medical entries in the MedLex lexicon are described in terms of a wide range of information with focus on semasiological and onomasiological characteristics of head words. The screen dump for the medical reading of the head word *depression*, available in the Appendix, gives an insight into the types of lexical records included in the MedLex database. These records are either a head word or sense oriented. The following records are linked to head words: *lemma*, *pronunciation, part of speech, inflection* and *frequency of the lemma*. While the lexical information provided by these records is self-explanatory and requires no comments, the numerical specification being part of the frequency information deserves some explanation. The word's frequency, if indicated, approximates the lemma's position within the 10,000 most frequent words extracted from a general language corpus including Swedish newspaper articles. As word frequency is represented in terms of thousand words intervals (starting with 0), the number shows the interval for the lemma in question. The frequency calculations are based on the principles discussed by Allén (1972).

The remaining records in the MedLex database form a section which is meaning/sense oriented. This section is further subdivided into four subsections with focus on (1) sense-related information, (2) linguistic samples, (3) other semantic information and (4) translations.

In the sense-related section, the following records are listed: *definition, definition comment*, *source, valency, style*, *grammatical comment*, *phrases, guide word and thematic path*. The types of information specified by particular records are implied by the labels, so just a few of them will be commented in what follows. Let's begin with the source record. As the database is to a certain extent a compilation of available lexical resources, the name of the lexical resource providing lexical background data, e.g. lemma form, inflection, definition etc. is put in the source record. As far as valency is concerned, syntactic patterns for verb arguments, taken from a learners' dictionary (Lexin), are provided. A list of recurrent collocations including a head word, sorted according to syntactic structures, is made available in the field labelled Phrase. Inclusion of information on a guide word and thematic path are recent extensions to the MedLex lexicon. The main purpose of guide words is to highlight a word's meaning(s) in a rough manner. For instance, for the polysemous Swedish noun *depression* its medical sense is explicated with ILLNESS as guide word. A thematic path for the mentioned noun includes reference to the following medical sub-domains imported from the MESH thesaurus: *Psychiatry and the Psychology Behaviour => Behavioural symptoms => Depression*. These nodes capture the onomasiological spectrum with the top node *Psychiatry and the Psychology* and *Depression* in the lowest node. The choice of MESH categories has been motivated by the fact that they are also used as semantic tags for annotating the MedLex corpus. Whenever a semantic tag in the corpus is identical to a corresponding item in a thematic path or guide word, a basic prerequisite for linking lexical data in the two resources is met and in consequence, a very first step towards the integrated semantic concordance has been undertaken.

The section devoted to samples contains mainly authentic sentences manually extracted from the MedLex corpus or the Internet, exemplifying the described sense of the head word and its syntactic context. Since Swedish is a compounding language, a list of recurrent compounds is provided, as these capture relevant semantic relations. Access to an explicit list, automatically generated, of compounds and their frequency in the MedLex workbench offers valuable guidance for the lexicographer when selecting relevant data.

In the section marked semantic information, there are records indicating *synonyms, antonyms*, *related concepts*, *domain level* and a *link to MESH*. Records listed as related concepts refer to categories such as hyperonyms, hyponyms and cohyponyms as well as to semantically related derivations of the head word. The latter are preceded by information on part of speech.

Information on the domain level is an approximation of whether a word belongs to the basic vocabulary, more advanced lay vocabulary or specialized language vocabulary. As far as determination of the domain level is concerned, the differentiation of the levels has been based on following criteria. The medical vocabulary listed in the learners' dictionary Lexin's *Svenska ord* gets the rank of basic vocabulary and is encoded as MDCN 0 in the MedLex lexicon. The supplementary medical vocabulary retrieved from other general dictionaries aimed at lay native speakers of Swedish, is marked as advanced level (MDCN 10), while vocabulary extracted from scientific papers aimed at medical experts is regarded as specialized language vocabulary (MDCN 20).

The overlap between the lexical and encyclopaedic knowledge becomes particularly evident in the work on medical terms. To facilitate the transition between the two, the semasiological description in the MedLex lexicon is linked to a corresponding onomasiological description provided by the MeSH thesaurus. The link points to a relevant medical sub-domain and paves the way to advanced medical information presented both in English and Swedish. Thus the bilingual information available in MESH can be seen as a complement to translation equivalents provided in the final section of the MedLex lexicon.

The above-mentioned diversity of information types in the MedLex lexicon lays the foundation for a number of NLP applications that require access to semantic and onomasiological content. It is also supportive of automatic compiling of medical lexicons which can be precisely tailored to the level of the user's medical knowledge and lexical needs. Some reflections on how the MedLex+ workbench can support compiling a medical collocational lexicon for Swedish is presented in the papers by Sköldberg and Toporowska Gronostaj (forthcoming 2008). As already mentioned, it is our conviction that the enriched content of the lexicon together with the enrichment of corpora can improve the content of lexicons for general and specialized language for human use and different NLP tasks.

4. Summary and conclusions

In this paper we have provided a brief description of MedLex+, an integrated corpus-lexicon workbench for the Swedish medical language. MedLex+ is primarily intended as a multifunctional lexical resource capable of supporting both human users and NLP applications in the sense of a *foreground lexicon* (Kilgarriff 1997). From what has been reported in the paper, it follows that the MedLex+ workbench has an immense potential for further effective enhancing the medical lexical database. Both lexicographers and lexicon users are endowed with tools and search queries which promote semantic mining within the medical domain. A conclusion that can be drawn is that the model elaborated for lexical infrastructure, MedLex+, seems to be promising. Although still in progress, MedLex+ already incorporates a number of analysis and acquisition tools and relatively large domain-specific corpus. Semantic concordances have a great potential for organizing the data into meaningful ways, facilitating easier access to argument structure, providing the means for enhancing the frame-based description of language elements, for example with regard to verbs significant for the medical domain. The future work will be focused on improving the stringency within the semantic records in the lexical database in accordance to the findings in the medical corpus. Particular attention will be drawn to further elaboration of the collocational module.

References

Abney, S. (1997). "Part-of-Speech Tagging and Partial Parsing". In Young, S.; Bloothooft, G. (eds.) *Corpus-Based Methods in Language and Speech Processing* 4. Dordrecht: Kluwer. 118-136.

Allén, S. (1972). *Tiotusen i topp*. Stockholm: Almqvist & Wiksell.

Borin, L.; Toporowska Gronostaj, M.; Kokkinakis, D. (2007). "Medical Frames as Target and Tool". In *Proceedings of the 16th Nordic Conference of Computational Linguistics*. Tartu, Estonia.

Brants, T. (2000). "TnT—A Statistical Part-of-Speech Tagger". In *Proceedings of the Sixth Applied Natural Language Processing Conference (ANLP)*. Seattle.

Burnard, L; Sperberg-McQueen, C. M. (2006). *TEI Lite: Encoding for Interchange: an introduction to the TEI Revised for TEI P5 release* [on-line]. URL: *http://www.tei-c.org/release/doc/tei-p5-exemplars/html/teilite.doc.html* [Access date: 25 March, 2008].

Christ, O. (1994). "A modular and flexible architecture for an integrated corpus query system". In *Proceedings of the Computational Lexicography* (*COMPLEX '94)*. Budapest, Hungary.

DocBook v5 [on-line]. URL: *http://www.docbook.org/xml/5.0/*. [Access date: 25 March, 2008]

Fillmore, C. J.; Johnson, C. S.; Petruck, M. R. L. (2003). Background to FrameNet. In *International Journal of Lexicography*. Oxford: Oxford University Press. Vol. 16. Nº 3.

Kilgarrif, A. (1997). *Foreground and Background Lexicons and Word Sense Disambiguation for Information Extraction. ITRI-97-04*. Brighton: University of Brighton.

Kim, J. D. et al. (2003). GENIA corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics* 19 (1). Oxford: Oxford University Press. i180-i182.

Kokkinakis, D.; Johansson Kokkinakis, S. (1999). A Cascaded Finite-State Parser for Syntactic Analysis of Swedish. In *Proceedings of the 9th European Chapter of the Association of Computational Linguistics (EACL)*. Bergen, Norway.

Kokkinakis, D. (2004). Reducing the Effect of Name Explosion. Proceedings of the 4th Language Resources and Evaluation (LREC) Workshop. In *Beyond Named Entity Recognition Semantic Labelling for NLP tasks*. Lisbon, Portugal.

Kokkinakis, D. (2006). Collection, Encoding and Linguistic Processing of a Swedish Medical Corpus—The MedLex Experience. In *Proceedings of the 5th Languages Resources and Evalutaion (LREC)*. Genoa, Italy.

Lexikon för invandrare, Svenska ord. Stockholm: Skolverket. 1992.

MEDLINE/PubMed database [on-line]. URL: *http://www.nlm.nih.gov/pubs/factsheets/medline.html*. [Access date: 25 March, 2008].

MESH [on-line]. URL: *http://www.nlm.nih.gov/mesh/*. [Access date: 25 March, 2008].

Miller, G. A. et al. (1993). A Semantic Concordance. In *Proceedings of the Workshop on Human Language Technology*. Princeton. 303-308.

Nationalencyklopedins ordbok. (1995-1996). Språkdata, University of Gothenburg. Höganäs: Bra Böcker.

Sköldberg, E.; Toporowska Gronostaj, M. (forthcoming). Modell för beskrivning av kollokationer i ett medicinskt lexicon (MedLex). In *Proceedings of the Nordiska föreningen för lexikografi (NFL, 2007)*. Akureyri, Iceland.

Sköldberg, E.; Toporowska Gronostaj, M. (2008) From subdomains and parameters to collocational patterns—on the analysis of Swedish medical collocations. In *Proceedings of the 13th EURALEX International Congress*. Barcelona, Spain.

Tapanainen, P.; Järvinen, T. (1998). Dependency Concordances. *International Journal of Lexicography* 11. 187-203.

XCES [on-line]. URL: *http://www.xml-ces.org/* [Access date: 25 March, 2008].

Appendix. The medical sense of the noun *depression* in the MedLex database

Lemma Information – Found (2)					
Lemma	depression			Updated	2008-03-27
Inflection	depression –en –er (depression depressionen depressioner)				
Frequency	gnral–8	Part-of-speech	subst.		
Sense Information (sense id–nr. 3066)					
Definition	sjukligt tillstånd med svår (långvarig) nedstämdhet				
Definition comment	ofta omfattande ångest, handlingsförlamning och livströtthet				
Source	LEXIN				
Valency					
Style		Grammatical comment			
Phrase	V+N: få en depression; drabbas av depression; gå in i depression; ha en depression; lida av depression; komma ur en depression; ta sig ur en depression; ADJ+N: djup depression; svår depression; lätt depression; mild depression; kronisk depression;		Comment		
Thematic Path	Psychiatry and Psychology, Behaviour, Behavioural symptoms, Depression		Guide Words	Illness	
Linguistic Samples					
Examples	Depression har blivit en av våra stora folksjukdomar och chansen att du känner någon som är drabbad eller kommer drabbas är väldigt stor. ; Många människor blir nedstämda någon gång i livet. Om nedstämdheten djupnar och går över i en depression behöver du hjälp från sjukvården.;				
Idioms					
Compounds	utmattnings~depression; vinter~depression; förlossnings~depression; ålders~depression; depressions~medel; depressions~symtom				
Semantic Information					
Synonyms	nedstämdhet; nedtryckhet				
Antonyms					
Related Concepts	hyponymer: utmattningsdepression, vinterdepression, förlossningsdepression, åldersdepression; verb: deppa; adj.: deprimerad, deprimerande, deppig, depressiv;				
Domain Level	MDCN–0				
MeSH Link	F01.145.126.350				
Translation					
English	depression				
Latin	depressio				

Textual Occurrences (Frequencies – fr. the MEDLEX Corpus) Based on 25 million tokens	
Frequency	*Found 37:* depression (2767) utmattningsdepression (68) andningsdepression (54) vinterdepression (28) förlossningsdepression (27) benmärgsdepression (10) postpartumdepression (10) vårdepression (9) ångest/depression (5) manodepression (5) nedstämdhet/depression (3) cns-depression (2) mani/depression (2) säsongsdepression (2) utmattningssyndrom/utmattningsdepression (2) nsatt-behandling-mot-depression (1) mano-depression (1) läkemedeldepression (1) kt/utmattningsdepression (1) humörsvikt/depression (1) förstagångsdepression (1) dysfori/depression (1) desorientering/depression (1) demens/depression (1) oro/ångest/depression (1) potentiation/depression (1) psykiatricentrum/depression (1) ångest/spänning/depression (1) Ångest/depression (1) ångest-depression (1) vasodepression (1) utmatningsdepression (1) ungdoms-depression (1) tomhetsdepression (1) s-patienter-med-mild-depression (1) sjukdom/depression (1) binjuredepression (1)
Statistical Collocations (fr. the MEDLEX Corpus) Based on 25 million tokens and pointwise MI	
Collocations	*Found 84; [showing the first 50]:* unipolära depressionerna (17.7) förebåda utmattningsdepression (16.4) uppmärksammar depressionen (12.9) unipolära depressioner (12.7) montgomery-Åsberg depression (12.4) medelsvåra depressioner (12.1) maskerad depression (11.4) uttalad andningsdepression (11.1) egentliga depressioner (10.9) egentlig depression (10.9) djupa depressioner (9.6) tablettbehandlad depression (9.3) hamilton depression (9.3) atypiska depressioner (8.8) fobier depression (8.7) måttliga depressioner (8.6) obehandlade depressioner (8.5) svårare depressioner (8.4) långvariga depressioner (8.3) svårbehandlad depression (8.3) endogen depression (8.3) utlösa depressioner (8.2) psykos depression (8.2) djupare depressioner (7.9) avsnitten depression (7.9) svåra depressioner (7.8) diagnosen utmattningsdepression (7.7) diagnostiserad depression (7.7) sömnsvårigheter depression (7.6) nedstämdhet depression (7.6) lindra depressioner (7.5) sömnstörningar depression (7.4) upprepade depressioner (7.3) själva depressionen (7.1) reaktiv depression (7.0) obehandlad depression (6.9) differentialdiagnoser depression (6.9) recidiverande depression (6.8) demens depression (6.8) major depression (6.4) pectoris depression (6.4) förvirring depression (6.3) allvarliga depressioner (6.2) hjärtsvikt depressioner (6.2) minne depression (5.9) netdoktor depression (5.8) trötthet depression (5.5) behandla depressioner (5.5) avsnittet depression (5.4) samtidig depression (5.4)