A Semantically Annotated Swedish Medical Corpus

Dimitrios Kokkinakis
Department of Swedish Language, Språkdata
University of Gothenburg
Box 200, SE-405 30 Göteborg, Sweden
E-mail: dimitrios.kokkinakis@svenska.gu.se

Abstract

With the information overload in the life sciences there is an increasing need for annotated corpora, particularly with biological and biomedical entities, which is the driving force for data-driven language processing applications and the empirical approach to language study. Inspired by the work in the *GENIA Corpus*, which is one of the very few of such corpora, extensively used in the biomedical field, and in order to fulfil the needs of our research, we have collected a Swedish medical corpus, the *MEDLEX Corpus*. MEDLEX is a large structurally and linguistically annotated document collection, consisting of a variety of text documents related to various medical text subfields, and does not focus at a particular medical genre, due to the lack of large Swedish resources within a particular medical subdomain. Out of this collection we selected 300 documents which were manually examined by two human experts who inspected, corrected and/or accordingly modified the automatically provided annotations according to a set of provided labelling guidelines. The annotations consist of medical terminology provided by the Swedish and English MeSH® (Medical Subject Headings) thesauri as well as named entity labels provided by an enhanced named entity recognition software.

1. Introduction

Provision of application and domain-dependent labelled language resources, such as annotated corpora, is a crucial key for progressing R&D in the human language technology (HLT) field. Such resources constitute an indispensable part for evaluation, software prototyping and design validation. The manually verified linguistic annotation of electronic text material (corpora) is a prerequisite for the development and evaluation of standard language technology tools, such as taggers, and the process is highly relevant for a number of applications including information extraction, text mining and information retrieval. The issue that our work combines the collection and annotation of corpora in a specialized field, such as medicine, and for a lesser-spoken languages, Swedish, provides the means for which the promotion of the continuous development and growth of language technology research, for resource development and for the implementation of practical applications is maintained. While at the same time we enhance the diversity of the resource flora by creating a unique annotated material that fills the gap that currently exists in the Swedish language resource terrain.

In this paper, we start by providing a brief state of affairs description of a large Swedish medical corpus, the *MEDLEX Corpus*. Out of this material, 300 documents/articles have been selected and passed a thorough inspection process by two human curators (one linguistic and one domain expert) who inspected, corrected and/or accordingly modified the automatically annotated sample according to a set of annotation guidelines. The methodology adapted for the semantic annotation of the corpus is exemplified. The annotations consists of a mixture of medical terminology provided by the Swedish and English MeSH® (Medical Subject Headings) thesauri as well as named entity labels provided by enhanced named entity recognition software. The latter applies fine-grained entity labels implemented in a hybrid system combining a rule-based system, multiword name lexica and document statistics. A discussion on how the results of the annotation and inspection process can be further used for evaluation purposes in terminology and named entity recognition, as well as a valuable teaching and working material in advanced courses in language technology is exemplified. The first version of the material is planned to be released in the summer of 2008.

2. The MEDLEX Corpus

With the information overload in the life sciences there is an increasing need for annotated corpora (particularly 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. Cohen *et al.* (2005) present a survey of biomedical corpora and the authors discuss the importance of format and annotation standards for usability and usefulness in research activities. Striving towards this direction, 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 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. All text samples (20 million tokens, 50,000 documents) are fetched from heterogeneous web pages during the past couple of years, 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 and blogs etc. All texts have been converted to text files, and have been both structurally and linguistically annotated. The 300 documents selected for manual inspection are part from this collection.

2.1 A 300-document Annotated Sample

Inspired by the work in the GENIA-corpus which has been extensively used in many bio-NLP[1] related activities during the last years and also by the fact that there is no such similar resource for Swedish we decided to initiate such a project, by extracting a sample from the MEDLEX Corpus which has been undergone a thorough manual inspection of its annotated content. The textual sample consists of 300 articles from the official magazine of the *Swedish Medical Association*, *Läkartidningen*, <www.lakartidningen.se>, particularly articles from the subsections *New Findings* (Nya Rön), and *Clinical research and Science* (Klinik och Vetenskap). Articles from the first section are usually short 1/2-2 pages long, translated from English leading (bio)medical journals. Articles from the second section are much longer, usually 3-8 pages and written by professional medical doctors and scientists, native speakers of Swedish.

For the annotation we used MeSH, Medical Subject Headings (edition 2006), as it is a free resource, which makes it potentially attractive as a component to build on and explore. We apply both the Swedish and, for reasons explained in the next section, its corresponding English original source. Since we are interested in exploring the thesaurus at its fullest, we have applied various techniques for increasing the coverage of the resources (Section 3). Moreover, we have annotated the sample with a set of extended named entities (Section 4). The total size of the sample is 342,128 tokens.

3. MeSH®

MeSH is the controlled vocabulary thesaurus of the U.S. National Library of Medicine (NLM). The annotation we have applied are based on both the English and the Swedish translation of the year 2006 MeSH. The motivation for integrating the English hierarchy in our work has been the fact that it is fairly common that Swedish texts, intended both for professional and lay audience, contain portions of short or longer English segments. Moreover, the use of English simplex or compound terms in Swedish texts is also very common. This is probably due to the authors' unfamiliarity with the appropriate Swedish translation; by the influence or "contamination" from the English language, particularly in orthographic variation, e.g. use of *ph* instead of *f* (e.g. *lym**f**o-lym**ph**o*); use of *th* instead of *t* (e.g. *hypo**th**yreos-hypo**t**yreos*; ***th**orax-**t**orax*) and use of *c* instead of *k* (e.g. *brady**c**ardi-brady**k**ardi*); by spelling errors that might have a direct correspondence to English terms, the overuse of hyphen (e.g. *tetra-cyklin* instead of *tetracyklin*), or possibly because of an author finding the English spelling more appropriate or "correct". MeSH is a subset of the Unified Medical Language System Metathesaurus (UMLS), the world's largest domain-specific thesaurus and it is used for subject analysis of biomedical literature, particularly for indexing the MEDLINE/PubMed, the premier bibliography of NLM, a large repository of research papers from the medical domain. MEDLINE/PubMed contains bibliographic citations and abstracts from over 4,000 journals.

The MeSH hierarchy consists of 16 different semantic groups, or sub-hierarchies but not all corresponds to branches medical terminology, since some of the hierarchies are more homogeneous and strongly associated with medicine than others, e.g. *Anatomy* compared to *Information Science*. Therefore for the annotation of the sample we use the first five groups A-F, Anatomy [A], Organisms [B], Diseases [C], Chemicals and Drugs [D], Analytical, Diagnostic and Therap. Techniques/Equipment [E] and Psychiatry and Psychology [F].

3.1 Normalization of MeSH

The original MeSH database has undergone a number of normalization steps in order to be able to apply it for automatic annotation. The main steps included:

- Changing the order of the head and modifier complements as well as term variants with commas, in the original material to the word order one would expect in free text
- All inflected/non-inflected entries were coded into a neutral non-inflected variant
- Addition of optional inflectional morphological features and variants using regular expressions patterns to all entries; e.g. *vaccin**(en|et|er|erna)?*** (the vaccine; the vaccines)
- Addition of variant numeric forms; variants to the Roman numbers e.g. for the use of "III" the addition of "3" (e.g. for kollagen typ III we added kollagen typ 3), and also to the Arabic numbers e.g. for the use of "2" the addition of "II" (e.g. for Typ 2-diabetes we added Typ II-diabetes) were added
- Addition of derivational variants using a number of derivational patterns, particularly the frequent in Swedish forms of making adjectives from nouns, such as *farmakologi* and its derivational *farmakologisk* and nouns from nouns, such as *diagnos* and its derivational variant *diagnosis*
- Addition of variant forms based on empirical observations, particularly multiword terms, such as common text patterns for organisms as well as anatomical terms of Latin origin for which the text realization is usually found in a shortened form were added, particularly terms describing muscles, nerves, arteries and bacteria, e.g. for the MeSH

[1] *Bio-NLP* is the field of research that seeks to create tools and methodologies for sequence and textual analysis that combine bioinformatics and NLP technologies in a synergistic fashion (*cf.* Yandell & Majoros, 2002).

term *arteria carotis* we added *art carotis* and *a. carotis* and for *Staphylococcus aureus* we added *staph aureus* and *s. aureus*

- Case folding was applied to all terms, except the acronyms. This was necessary in order not to introduce new forms of ambiguity, since the 100% elimination of case information could introduce new ambiguities between homographs uppercase/low case words. For instance, *kol* [D01.268.150] (carbon) and *KOL* [C08.381.495.389] (Chronic Obstructive Pulmonary Disease).

4. Generic Named Entity Recognition

Following the paradigm proposed by Sekine (2004), we apply a fine-grained NER system for Swedish capable of recognizing eight main categories: person, including personhood (male/female), location, organisation, event, object, work & art, time and measure, and over sixty subtype named entities, including a large set of different types of measure subgroups relevant to the domain, such as: pressure, frequency, weight, dosage, volume and temperature, *cf.* Kokkinakis (2005).

5. Document Pre and Post-Processing

It is a well-known fact that even within the same text, a term can take many different forms. Tsujii & Ananiadou (2005) discuss that "a term may be expressed via various mechanisms including orthographic variation, usage of hyphens and slashes […], lower and upper cases […], spelling variations […], various Latin/Greek transcriptions […] and abbreviations […]." This rich variety for a large number of term-forms is a stumbling block especially for text mining, as these forms have to be recognised, linked and mapped to terminological and ontological resources; for a review on normalization strategies see Krauthammer & Nenadic (2004). Consider the following two examples for terms extracted from the MEDLEX Corpus which clearly illustrates the term variability in authentic corpora, the original MeSH annotation in the first case is *Typ 2-diabetes* (Diabetes Mellitus, Type 2) and in the second *COX-2-hämmare* (Cyclooxygenase 2 Inhibitor)

(a) diabetes typ 2, diabetes typ II, typ 2-diabetes, typ II diabetes, typ II-diabetes, typ2 diabetes, typ-2 diabetes, typ2-diabetes, typ-2-diabetes, 'diabetes mellitus, rimligen typ 2'…

(b) Cox 2, cox-II hämmare, Cox II, COX-2 hämmare, COX-2-hämmare, cox 2-hämmare, COX2-hämmare, cox 2-hämmare, Cox-2 hämmarna …

In order to capture cases as the previous one, we have generated permutation of the multiword terms in MeSH and added the new forms in the database. However, even greater problem and challenge is posed by solid compound terms *not* in MeSH and for which compound analysis as it is described in the following section, is necessary.

5.1 Compound Analysis

Compounds pose a serious problem for many tasks when processing Swedish with the computer, particularly in applications that require morphological segmentation, such as Information Retrieval. In Swedish, compounds are written almost exclusively as one orthographic word (solid compounds) and are very productive. Therefore, for potential compound terms where there are no entries in MeSH covering these forms, heuristic compound segmentation is necessary. Inspired by the work of Brodda (1979) we have implemented a domain-independent, finite-state based segmenter that builds on the idea of identifying "unusual" grapheme clusters (usually consonants) as means of denoting potential compound limits. The segmentation algorithm we have developed is a non-lexical, quantitative one and it is based on the distributional properties of graphemes, trying to recognize grapheme combinations, indicating possible boundaries. It proceeds by scanning word forms from left to right, trying to identify clusters of character combinations (n-grams) that are non-allowable when considering non-compound forms, and which carry information on potential token boundaries. The grapheme combinations have been arranged into groups of 2 to 8 characters. For instance, an example of a two-character cluster is the combination *sg* which segments compounds such as *virus||genom* (virus genome) and *fibrinolys||grupp* (fibrinolysis group); a three-character cluster is the combination *psd* which segments compounds such as *lewykropps||demens* (Lewy Body Dementia); a four-character cluster is *ngss* which segment compounds such as *sväljnings||svårighet* (swallowing difficulty) and so forth. Special attention has been given to compounds where the head or modifier is a very short word (2-3 characters long), such as *lår* (thigh), *sår* (wound), *hår* (hair), *tå* (toe), *yt* (surface), *syn* (sight), *hud* (skin) and *gen* (gene). For such cases we have manually added clusters of short characteristic contexts taken from the MEDLEX Corpus, usually 4-6 characters, before or after the short words. Compound splitting into its parts enables partial or whole annotation with MeSH codes and substantial improvement of indexing.

5.2 Elliptic Coordinations - Gapping

For maximum performance, the input texts can be optionally pre-processed in various ways (see the previous discussion) in order to resolve certain frequent types of coordinated constructions with ellipsis (also known as *gapping*). These can mainly be of three types:

- *solidCompound binder –partialCompound* (e.g. *binjurebarken och –märgen* i.e. adrenal cortex and adrenal medulla)
- *partialCompound– binder solidCompound* (e.g. (*rygg– och nackvärk* i.e. back pain and neck pain)
- *multiW1 multiW2– binder multiW1 multiW3–Term*

(*typ 1– och typ 2–diabetes* i.e. type 1 diabetes and type 2 diabetes)

Here, *binder* refers to a conjunction such as *och*/and or *eller*/or. When such patterns are identified, the solid compound is automatically segmented and the elliptic, partial compound gets the head of the complete compound. This means that in the example *rygg- och nackvärk*, the compound *nackvärk* is segmented as *nack||värk* and *värk*, the head of the compound, is added as the head for *rygg*, and thus the whole phrase becomes *ryggvärk och nackvärk*. Here '||' denotes the border between the head and the modifier of the compound. In order to achieve this type of labelling, compound segmentation, as described previously, is applied and then the text is processed with a module that recognizes and restores candidate discontinuous structures. As soon as the segmentation is performed, the restoration of such structures becomes a trivial task using simple pattern matching. Note, that in case of more than one segmentation points, the rightmost segmentation is considered for the restoration. For instance, *stroke- och hjärtinfarktregister* (stroke registry and infarction registry) becomes after compound segmentation *stroke- och hjärt||infarkt||register*, with two segmentation points. But since the rightmost segmentation point is considered, the coordination will take the form *stroke||register och hjärt||infarkt||register*. Moreover this resolution approach is not limited to binary coordinations but to *n-ary*. For instance *alfa-, beta- och gammaglobulin* (alpha, beta and gamma globulin) becomes after compound segmentation *alfa-, beta- och gamma||globulin* and finally *alfa||globulin, beta||globulin och gamma||globulin*. 368 of such cases could be found in the 300 document sample.

5.3 Approximate String Matching

We can safely assume that official, edited vocabularies will not be able to identify all possible terms in a text. There are a lot of cases that could be considered as MeSH-term candidates but are left unmarked, particularly in the case of misspellings. Approximate string matching is fundamental to text processing for identifying the closest match for any text string not found in the thesaurus. Since we are interested to identify as many terms as possible and with high accuracy, such technique seems very practical for achieving this goal. String matching is an important operation in information systems because misspelling is common in texts found in various web pages, particularly blogs. Therefore, we also calculate the orthographic similarity between potential candidates (≥ 7 characters long) and the MeSH content, We have empirically observed that the length of 7 characters is a reliable threshold, unlikely to exclude many misspellings. As measure of orthographic similarity (or rather, difference) we used the Levenshtein distance (LD; also known as edit distance) between two strings. The LD is the number of deletions, insertions or substitutions required to transform a string into another string. The greater the distance, the more different the strings are. We chose to regard 1 as a trustworthy value and disregarded the rest (misspelled terms and MeSH terms usually differ in one character) although there were a few cases for which the value of 2 could provide compatible results. For instance, the misspelled *accneärr* (Acne Keloid) which could be matched to *akneärr* with LD=2. Hence by this approach, and after manual inspection, we actually chose to add the very frequent spelling errors in the thesaurus itself. The method is also applied *on the fly* while indexing arbitrary texts.

5.4 Integration of Acronyms

Long full names in (bio-) medical literature are almost always abbreviated, most frequently by the use of acronyms, which implies the creation of new sets of synonyms. Such abbreviations can introduce ambiguity since they might overlap with other abbreviations, acronyms or general Swedish or English vocabulary, as in *hemolytiskt uremiskt syndrome (HUS)* (Hemolytic-Uremic Syndrome), where *HUS* also stands for the Swedish common noun house (i.e. house). Therefore, discovering acronyms and relating them to their expanded forms is an essential aspect of text mining and terminology management. Shultz (2006) claims that online interfaces do not always map medical acronyms and initialisms to their corresponding MeSH phrases. This may lead to inaccurate results and missed information if acronyms and initialisms are not used in search strategies. Acronyms are rather rare in MeSH and freely available acronym dictionaries in Swedish are currently non existent, while they are rather frequent in biomedical texts. Therefore, we applied a simple, yet effective, pattern matching approach to acronym identification, using a set of hand-coded patterns. The pattern matching approach is applied after the annotation of a text with MeSH labels. Appropriate annotations in conjunction with orthographic markers in the near vicinity of an MeSH-annotation drive the recognition of acronyms, throughout a document. Note that it is generally perceived that acronyms are usually introduced once in a text and then frequently used in the same document instead of the expanded form; this means that it is not safe to simply use an identified acronym in one document for the annotation of a seemingly similar acronym in another document. However, it is rather safe to consistently use the same *meaning* of an acronym throughout a single document. The applied approach has certain similarities with work by Pustejovsky *et al.* (2001) and Schwartz & Hearst (2003), but here we apply more patterns with more variation and not merely the *Aaa Bbb Ccc (ABC)* where *Aaa*, *Bbb* and *Ccc* are words in a multiword term. A handful of simple heuristic pattern matching rules can capture a large number of unknown to the resource acronyms and thus assign appropriate MeSH labels. In previous studies based on Swedish data the most frequent acronym patterns where of the form: *D (A)* 66,2%, *D, A,* 14,2% and *A (D)* 5,7%, here *D* stands for the expanded form of an acronym *A*; *cf.* Kokkinakis & Dannélls, 2006.

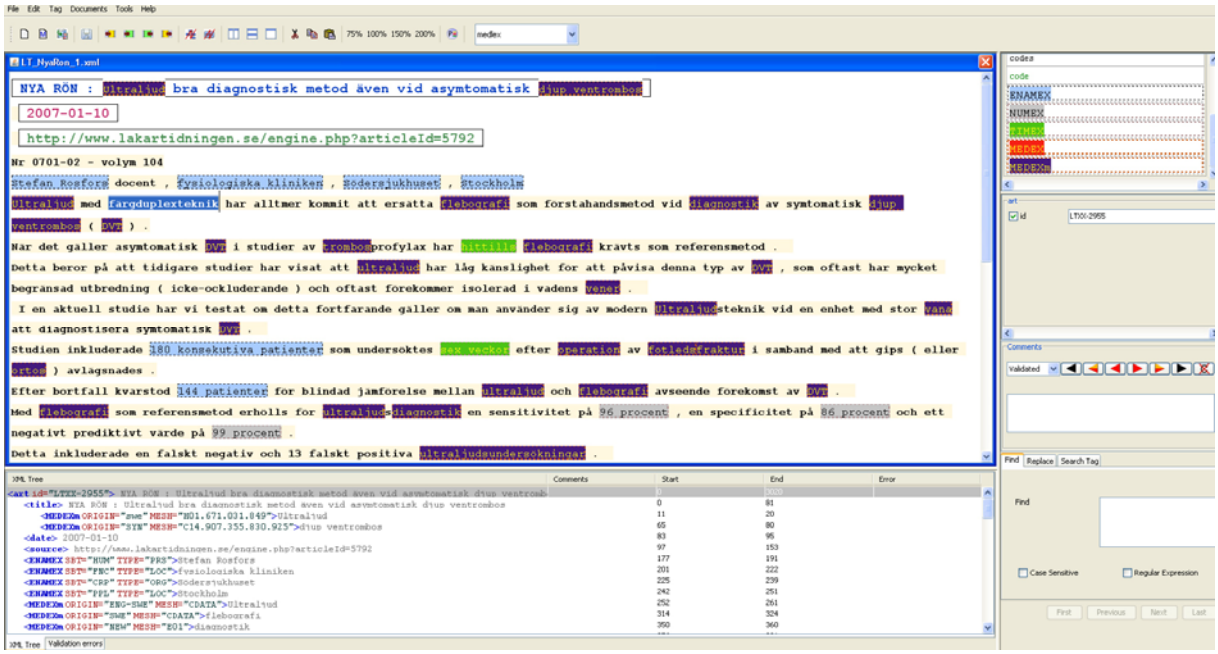

Figure 1: Interactive inspection using the CADIXE editor; the top left view is the annotated text (each XML-tag is associated with a *style sheet*), the attribute zone is to the right and the XML-code generated is shown at the bottom.

6. Annotation and Manual Inspection

The only restriction that is posed on the documents to be annotated is that the texts are tokenized (some basic form of separation between graphic words and punctuation). Moreover, and for maximum performance, the input has been processed in accordance to the previous discussion (e.g. elliptic coordinations). The annotation process uses XML-elements each of which contains suitable attributes specifying the nature of an annotation. Each MeSH term is annotated using a simple metadata scheme with three attributes. The first attribute designates the alphanumeric MeSH code (id), the second the origin of the tag (src) and the third whether the term occurrence is negated or not. The origin's attribute of a MeSH-tag can take one of the following values:

swe for a term originating from the Swedish MeSH
e.g. *<mesh id="C08…" src="swe">astma</mesh>*
eng for a term originating from the English MeSH
e.g. *<mesh id="D11…" src="eng">ephrins</mesh>*
syn for a synonym
e.g. *<mesh id="C20…" src="syn">*allergier*</mesh>*
acr for a newly identified acronym; e.g. *<mesh id="C10…" src="acr">GBS</mesh>*, for *Guillain-Barres syndrome*
mdf a modified MeSH term, such as derivations and "empty" suffixes; e.g. *<mesh id="C23…" src="mdf">syndrom**typ**</mesh>*
new which stands for terms added to MeSH, e.g. brand names of medicines and misspelled terms; e.g. *<mesh id="C14…" src="new">ischmi</mesh>*

The automatically annotated data have been manually inspected by the two human experts using the CADIXE XML Annotation Editor (Bisson, 2005), a user friendly interface which nicely integrates style sheets and XML DTDs for the convenient inspection and modification of the annotated documents (Figure 1). Some of the problematic cases we could identify in the annotated sample and which we have not dealt with during the various pre-processing steps had to do with

- the use of multiword compounds instead of solid compounds, e.g. in MeSH there is the solid compound *socialfobi* (*Phobic disorder*) but in the texts we could find a multiword variant *social fobi*
- a number of elisions were also observed, e.g. in MeSH there is the term *chikungunya virus* but in the texts we could only find *chikungunya*.
- a number of terms are only partially covered by MeSH *kronisk <trötthet>* (chronic fatigue); while for some multiword terms as well as solid compounds, only a fraction was covered, e.g. *lindrig <tyreoidea>rubbning;* e.g. *<color> duplex <sonography>*.
- a number of potential terms could not be marked, simply since MeSH lacks a description e.g. *kalcipotriol*, *rimonabant*, which hints on the limitations of MeSH in terms of its coverage.

There was a number of spuriously identified concepts, due to homography with non-medical words, such as *huvuddelen* (part of the head), which more frequently used in an adverbial position, i.e. 'mainly'; *leder* (joints), which was used as the homograph verb 'to lead' and *tunga* (tongue), which was used as the homograph adjective 'heavy'. Note, that for homography between verbs and nouns or adjectives and nouns, part-of-speech tagging can be of great help for distinguishing the two forms from each other but distinguishing different senses of noun required more advanced processing in the form of word

sense disambiguation.

6.1 Sample Statistics

Table 1 shows some statistics taken from the 300-document sample. There were 26,412 MeSH-annotations for categories A-F, out of which 19,757 were full annotations, that is the whole string is completely matched by a one or more MeSH labels. Here "full" incorporates not only simplex terms, but also compound terms not in MeSH (non-lexicalised compounds) that were analysed during compound analysis, and for which "complete" annotation could be still obtained. For instance, both *fetma* (obesity) and *patient* (patient) are "lexicalised" in MeSH, however the compound *fetmapatienter* (obesity patients) is not, therefore after compound analysis (Section 5.1) we obtain the two strings which are annotated accordingly (and thus considered here as full annotation), e.g. *<mesh id="C18...">fetma </mesh><mesh id="M01…"> patienter</mesh>*.

1,435 of the 26,412 annotations had only the end-part labelled, usually the head of a compound e.g. *vuxen<mesh id="F04...">psykiatri</mesh>* (adult psychiatry). 4,710 of the 26,412 had only the start-part labelled, usually the modifier of a compound, e.g. *<mesh id="D06...">insulin</mesh>infusion* (insulin infusion). Most of the rest of the 510 annotation were of the form that only a part of the string was matched, such as *icke-<mesh id="F01...">språk</mesh>lig* (non-verbal).

Source	#	Example
Swedish MeSH	11,370	<astma>
Synonyms	5,046	<akne>
English MeSH	1,828	<collagen>
New Entries	993	<diarré>
Acronyms	315	<DVT>
Modified Entries	195	<stroke>*liknande*
Both Eng./Swe.	10	<hand>
Elliptic Coordinations	386	HDL- och LDL-‖kolesterol

Table 1. Distribution of MeSH tags according to *source of origin*

The most frequent labels in the sample (at level 0) were *E05-Investigative Techniques* (2088), *E01-Diagnosis* (1354), *C23-Pathological Conditions, Signs and Symptoms* (1338), *F01-Behavior and Behavior Mechanisms* (1179) and *C10-Nervous System Diseases* (984).

7. Conclusions

We have outlined our work on developing a Swedish medical corpus sample of 300 scientific documents. The annotated document sample is manually inspected and is freely available for research purposes. The texts can be used as a *gold-standard* sample for actvities in the Swedish language technology field. The material can be used by students and researchers in experimenting with computational models, methods and tools for evaluation purposes in terminology recognition and named entity recognition for Swedish, which is an important subtask in information extraction and question-answering applications. With a compilation of such resource, it is possible to say, in an objective way, which annotation programs and methods perform better, which worse, which parts present particular problems to which programs etc. Therefore the creation of such a sample, that can be shared for evaluation exercises for Swedish language technology tools, is a valuable resource, of great importance which does not exist elsewhere in Sweden (to the best of the author's knowledge). In the near future we intend to successively continue with the enrichment of the sample with other types of linguistic features, including part-of-speech and coreference chains and put efforts on extending the sample with more documents.

Acknowledgements

This work has been supported by the Centre of Language Technology (CLT) small project grant, University of Gothenburg (<http://www.clt.gu.se>).

References

Bisson G. (2005). CADIXE: a XML Annotation Editor. The CADERIGE Project. <http://caderige.imag.fr>

Brodda B. (1979). Något om de svenska ordens fonotax och morfotax: Iakttagelse med utgångspunkt från experiment med automatisk morfologisk analys. *PILUS nr 38*. Department of Swedish, Stockholm University. (In Swedish).

Cohen K.B., Ogren P.V., Fox L. and Hunter L. (2005). Empirical Data on Corpus Design and Usage in Biomedical Natural Language Processing. *AMIA Annual Symp Proc.* (pp. 156–160). Washington, USA.

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. OUP.

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. (2006). *Collection, Encoding and Linguistic Processing of a Swedish Medical Corpus - The MEDLEX Experience*. Proceedings of the 5th Conference on Language Resources and Evalutaion (LREC). Genoa, Italy.

Kokkinakis D. and Dannélls D. (2006). *Recognizing Acronyms and their Definitions in Swedish Medical Texts*. Proceedings of the 5th Conference on Language Resources and Evalutaion (LREC). Genoa, Italy.

Krauthammer M. and Nenadic G. (2004). Term identification in the biomedical literature. *Journal of Biomedical Informatics. Special issue: Named entity recognition in biomedicine.* 37(6), 512 – 526.

Pustejovsky J. *et al.* (2001). Automation Extraction of Acronym-Meaning Pairs from MEDLINE Databases. *Medinfo 2001*. 10 (Pt 1), 371-375.

Schwartz A. and Hearst M. (2003). A simple algorithm for identifying abbreviation definitions in biomedical texts. *Proceedings of the Pacific Symposium on Biocomputing (PSB)*. Hawaii, USA.

Sekine S. (2004). *Definition, dictionaries and tagger for Extended Named Entity Hierarchy*. Proceedings of the 4th Conference on Language Resources and Evaluation (LREC). Portugal.

Shultz M. (2006). Mapping of medical acronyms and initialisms to Medical Subject Headings (MeSH) across selected systems. *Journal Med Libr Assoc.* 94(4): 410–414.

Tsujii J. and Ananiadou S. (2005). Thesaurus or logical ontology, which one do we need for text mining? *Language Resources and Evaluation, Springer Science and Business Media B.V*. 39:1, 77-90. Springer Netherlands.

Yandell, M. D., & Majoros, W.H. (2002). Genomics and natural language processing. *Nature Reviews Genetics*. 3, 601-610, doi:10.1038/nrg861.