MeSH® - From a Controlled Vocabulary to a Processable Resource

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

Abstract

Large repositories of life science data in the form of domain-specific literature, textual databases and other large specialised textual collections (corpora) in electronic form increase on a daily basis to a level beyond the human mind can grasp and interpret. As the volume of data continues to increase, substantial support from new information technologies and computational techniques grounded in the form of the ever increasing applications of the *mining paradigm* is becoming apparent. These emerging technologies play an increasingly critical role in aiding research productivity, and they provide the means for reducing the workload for information access and decision support and for speeding up and enhancing the knowledge discovery process. In order to accomplish these higher level goals and support the mining approach however, a fundamental and unavoidable starting point is the identification and mapping of terminology from the textual, unstructured data onto biomedical knowledge sources and concept hierarchies. In this paper, we provide a description of the work regarding terminology recognition using the Swedish MeSH® thesaurus and its corresponding English original source. We explain the various transformation and refinement steps applied to the original database tables into a fully-fledged processing oriented annotating resource. Particular attention has been given to a number of these steps in order to automatically map the extensive variability of lexical terms to structured MeSH® nodes. Issues on annotation and coverage are also discussed.

1. Introduction

Identification, classification and mapping of terminology from the textual, unstructured data onto biomedical knowledge sources and concept hierarchies, such as domain-dependent thesauri, nomenclatures and ontologies is the first, but crucial step, for a deeper semantic analysis and exploration of the unstructured textual content (Ananiadou & McNaught, 2006; Crammer *et al.*, 2007; Krauthammer & Nenadic, 2004; Névéol *et al.*, 2007; Vintar *et al.*, 2003). The task is considered as one of the most challenging research topics within the *biomedical natural language processing* community (bio-NLP), 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 (Yandell & Majoros, 2002). Ananiadou & Nenadic (2006) point out that processing and management of terminology is one of the key factors for accessing the information stored in literature, since information across scientific articles is conveyed through terms and their relationships. Indexing, which is one of the main target activities of this mapping, is an indispensible step for efficient information retrieval engines and applications. A step that is realized as *the* most time consuming activity for librarians, *cf.* Névéol *et al.* (2005). Moreover, thesauri and ontologies are considered the backbone for various data and knowledge management systems. In our work, we take the point that such resources *do* exist in a digital form. We will use 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 and therefore there is no need to create a thesaurus from scratch. Thesaurus learning and fully automatic, corpus-based thesaurus construction, as alternative methodologies for knowledge management application design, are out of the scope of this paper. However, a number of techniques for *thesaurus enrichment*, and how newly acquired terminology can be related to the original MeSH hierarchy are envisaged and will be described.

2. The MeSH Thesaurus

MeSH® is the controlled vocabulary thesaurus of the NLM, from the U.S. National Library of Medicine. The implementation of the terminology annotator we use is based on both the English and the Swedish translation of the year 2006 MeSH® thesauri. 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. MeSH® is a subset of the UMLS (Unified Medical Language System) Metathesaurus, 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 database of research papers from the medical domain (bibliographic citations and abstracts from over 4,600 journals). MeSH® has been used for different purposes in a variety of applications and settings. Cooper & Miller (1998) developed and evaluated three different methods for mapping terms in clinical text to MeSH® terms. Their results varied between 17-44% for precision and 40-66% for recall. Rosario *et al.* (2002) used MeSH® for the analysis of compound nouns, namely placing words from a noun compound into categories, and then using the category membership to determine the relation that holds between the nouns. Rechtsteiner & Rocha (2004) report on validation of gene expression clusters;

while Djebbari *et al.*, 2005 apply MeSH® to the identification of biological concepts. Moreover, Douyere *et al.*, (2004) enhance MeSH® in order to adapt the terminology to a wider field of health internet resources instead of scientific articles, by introducing two new concepts, namely 'resource types' and 'metaterms'. Finally, Struble & Dharmanolla (2004) investigated the use of MeSH® for discovering groupings within a collection of medical literature in PubMED.

3. MeSH Transformation/Normalization

A tool designed for *automatic indexing* is faced with at least two interrelated issues that need to be taken under serious consideration, in order to successfully map free text to control vocabularies and hierarchies such as MeSH. The first issue (Sections 3.1-3.8) has to do with the necessary adaptations of the resource content to a format suitable for text processing. Necessary, since it has been claimed by a number of researchers that many term occurrences cannot be identified in text if straightforward dictionary/database lookup is applied (*cf.* Hirschman *et al.*, 2003). Therefore a number of conversion and normalization steps have to be applied to the original material. Secondly, is the fact that we have to efficiently deal with language variability by using, or even better, combining various techniques such as stemming (Jacquemin & Tzoukermann, 1999) and approximate string matching (Tsuruoka & Tsujii, 2003), see Section (4). Normalization is thus necessary before the actual application of the MeSH content due to the nature of the original data, in which the head word usually precedes its modifiers, e.g. *Lung Diseases, Obstructive* with the Swedish translation *Lungsjukdomar, obstruktiva*. Thus, a great effort has been put into the normalization of MeSH, since, compared to English and UMLS, for which a number of supporting software tools are available as part of the SPECIALIST[1] lexicon, the situation for Swedish is diametrically different and there are no similar tools available.

3.1 Head and Modifiers

The first step applied to the MeSH database was to change the order of the head and modifier complements, usually variants with commas, in the original material to the word order one would expect in free text. There are several hundreds of such cases in the database due to obvious terminological and lexicographic purposes, e.g. easier means of sorting based on head words, which had to be changed in order to be able to apply the lexical content to text data. For instance, *Vacciner, orala* (Vaccines, edible) has been changed to *orala Vacciner*.

3.2 Inflected Variants

The second step was to normalize all inflected entries into a neutral *non-inflected* variant by applying a number of morphology stripping patterns. This was necessary since there is a combination of both inflected and uninflected terms in the database. Although Swedish is morphologically more complex than English there is a small number of inflectional patterns that cover the vast majority of inflected variants in the MeSH database, e.g. adjectives are usually inflected with the addition of *–t* (depending on the gender) and *–a* (depending on the number). Special patterns were constructed for a small number of noun terms with irregular inflection patterns, e.g. *öra* (ear) and *öron* (ears). Thus taking the example in Section 3.1, *orala Vacciner* was normalised to *oral vaccin*.

Case folding was applied to all terms at this stage, except those consisting of uppercase letters, 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 uppercase/low case words, for instance, *kol* [D01.268.150] (carbon) and *KOL* [C08.381.495.389] (Chronic Obstructive Pulmonary Disease).

3.3 Inflectional Morphology and Variant Numeric Forms

At this stage, each entry is either of the form: $term_x$ => *mesh.tag(s)* or $term_x$' '$term_z$ => *mesh.tag(s)* in case of multiword terms. Depending on the suffixes of the adjectives and/or nouns in MeSH, we heuristically added inflectional morphological features and variants using regular expressions patterns to all entries, and in the case of multiword terms to both head and modifier(s). Thus, each term has been actually encoded as: $term_x(m_1|\ldots|m_n)?$ => *mesh.tag(s)* or $term_x(m_1|\ldots|m_n)?$' ' $term_z(m_1|\ldots|m_n)?$ => *mesh.tag(s)*, where m_1 to m_n are different optional inflectional suffixes. Here optionality is denoted by '?' and disjunction by '|'. This is a necessary step since we want to apply our system on raw, unstemmed, texts and therefore we did not want to pose any particular restrictions to the nature of the input, which can be both ungrammatical and contain spelling errors, particularly at the declination level. Therefore, apart from the *grammatically correct* inflectional patterns we also added wrong gender inflection patterns since we have noticed that particularly for the category [C] terms, grammatical gender (*–et* or *–en*) usually occurs in either forms, e.g. *adrenalin**et*** and *adrenalin**en***. Thus taking the example in Section 3.1 once again, *oral vaccin* is transformed to *oral(t|a)? vaccin(en|et|er|erna)?.*

We also added 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*).

3.4 Derivational Variants and Empty Suffixes

A small number of derivational patterns are also considered in order to add new entries through this affixation process. Particularly emphasis was put on

[1] The SPECIALIST lexicon contains for each term, variants such as acronyms, abbreviations, synonyms, derivational variants, spelling errors, inflectional variants and meaningful combinations of all these.

productive forms of making noun-to-adjective or noun-to-participle derivations with the suffixes *–sk* and *–nde*, as well as forms of making noun-to-noun derivations with the suffixes *–ik*, *–ing* and *–ion*.

In parallel, we developed a component utilizing a set of "empty" suffixation patterns to various MeSH groups. During the development cycle of our system we identified a number of such group-dependent markers that do not substantially change the meaning of a term, but simply act either as placeholders, or sometime adding a slightly more detailed nuance to the term in question. For instance, for the main heading *[B] Organisms* a common suffix is 'type', *bakterie**typ*** (bacteria type), for *[E] Techniques and Equipment* a common suffix is 'method', *gastrostomi**metod*** (gastrostomy method) while for the main heading *[C] Chemicals and Drugs* a frequent suffix is 'tablet', *tetracyklintablett* (tetracycline tablet).

3.5 Variant Forms Based on Empirical Observations

Although a number of vocabularies, taxonomies and thesauri have been developed for the healthcare and biomedical domain, no single vocabulary can ever provide complete coverage of the information that can be found in free text. Therefore support from empirical studies provides an indispensable way to enhance existing taxonomies Almeida *et al.* (2003) argue that existing vocabularies may serve as content starting points to avoid duplication of work. During the development cycle of the MeSH-annotator we observed a number of typographical variant cases of existing terms that we successively added to the database in an automatic fashion. Thus, apart from the cases discussed in the previous sections, we also added a large number of variants of multiword terms using pattern matching. Some of the most common text patterns had to do with various organisms as well as anatomical terms of Latin origin for which the text realization is usually found in an abbreviated form, particularly terms describing *muscles*, *nerves* and *arteries*. For instance, for *vena cava inferior* it is more probable to find *v. cava inferior* or *v. cava inf.* while for *helicobacter pylori* it is more likely to find *h pylori* or *h. pylori*.

Automatically extracted paraphrase variant forms of solid MeSH compounds were also added after manual inspection, e.g. for the compound *njurcancer*, added the *cancer i njuren* (Kidney Neoplasm). Compounds are usually corresponding to multiword phrases in other languages, such as English. A large number of such forms (roughly 500) were extracted from the MEDLEX corpus, sometimes referred to as permutations (Jacquemin, 2001) and added to MeSH. Moreover, orthographic variant forms of compounds, compounds with acronyms and elisions were also added, e.g. *restless legs syndrome* added *restless legs*.

3.6 Errors in the Swedish MeSH

Some errors in the original material were also identified and added to the database, while other discrepancies were also minimized and normalized. For instance, there are both singular and plural forms for some head terms, such as *aortapulmonal **septumdefekt*** (aortopulmonary septal defect) in singular and *ventrikel**septumdefekter*** (heart septal defects, ventricular) in plural. Similarly, there were cases of both definite and indefinite forms, such as *bakterier i **urinen*** (bacteriuria) and *blod i **urin*** (hematuria), which were also normalised according the morphological normalization process described earlier.

3.7 Medical Brand Names

Apart from the work described in the previous section, particularly Section 3.5, we have also explored other means for enhancing the content of MeSH. One such technique is to look-up names of medicines, i.e. the trade name given by the manufacturers, and automatically map these names to their generic ones, if such a name is present in MeSH. This was accomplished using a handful handcrafted pattern matching rules. For instance, *Cipramil*®, a medicine used for treatment of depression and anxiety disorders, contains the active ingredient *citalopram*. *Citalopram* can be found in MeSH under the nodes *D02.092.831.170, D02.626.320* and *D03.438.127.187*, but not *Cipramil*® which has been added by this matching process with the same alphanumeric coding as its generic name. In the MEDLEX Corpus there were 940 unique brand name contexts for which we could obtain MeSH annotations for the generic name of a brand names. For a small number of these candidates (9%) only a part of the generic compound substance name was annotated by MeSH (e.g. *zoledron<mesh id="D01…">syra</mesh> (Zometa)*) and we chose *not* to add such name in the database. There were also cases in which neither the generic name or the trade name could be associated with a MeSH term (e.g. *tadalafil (Cialis)*.

3.8 External Lexical Resources

Although MeSH and its enhancements, as described previously, are a good starting point for mapping free text to structured codes, texts contain a lot of other types of medical terminology that needs to be considered. MesH lacks for instance information on (at least) two types of important and frequent occurring terminology: names of pharmaceutical products (drugs) and anatomical Greek and Latin terms. For that purpose, we have added several thousand names of pharmaceutical products, particularly names of drugs, applying a generic annotation with simply the code [*D*] which in MeSH stands for *Chemicals and Drugs*, unless the generic name of the drug could be mapped to a structured MeSH code, according to the previous discussion. The pharmaceutical names have been obtained from a reference book of all medicines that are approved and used in Sweden (<http://www.fass.se>), while terminology of Greek/Latin origin, particularly anatomical terms, have been obtained from the Karolinska institutet (<http://www.karolinska.se>) in this case the generic annotation with code [*A*], which in MeSH stands for *Anatomy*, has been used.

4. Text Processing

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; *cf.* Krauthammer & Nenadic (2004). For instance, the official Swedish MeSH *Typ 2-diabetes* is found in the MEDLEX corpus with the following variants: *diabetes typ 2*, *diabetes typ II*, *typ 2-diabetes*, *Typ 2-diabetes*, *typ II diabetes*, *typ II-diabetes*, *typ2 diabetes*, *typ-2 diabetes*, *typ2-diabetes*, *typ-2-diabetes*, '*diabetes mellitus, rimligen typ 2*',...

In order to capture cases as the above, we have generated permutations of the multiword terms in MeSH, supported by corpus evidence, and added the new forms in the database. However, even greater problem and challenge is posed by solid compound terms not in MeSH, and the next two sections discuss how these can be effectively covered.

4.1 Compound Segmentation

Compounds pose a serious problem for many tasks when processing Swedish with the computer, particularly in applications that require morphological segmentation. 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, compound analysis is necessary. Inspired by the work of Brodda (1979) we have implemented a domain-independent, finite-state based analyser 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 *viru**s**||**g**enom* (virus genome) and *fibrinoly**s**||**g**rupp* (fibrinolysis group); a three-character cluster is the combination *psd* which segments compounds such as *lewykrop**ps**||**d**emens* (Lewy Body Dementia); a four-character cluster is *ngss* and *gssp* which segment compounds such as *sväljni**ngs**||**s**vårighet* (swallowing difficulty) and *mässlin**gs**||**sp**ecifik* (measles specific), 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), *tum* (thumb), *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, enhancing technologies such as semantic relation mining.

4.2 Elliptic Coordinations - Gapping

The only requirement we pose prior to annotation is that the texts are tokenized (some basic form of separation between graphic words and punctuation). However, for maximum performance, the input texts can be optionally pre-processed in various ways (see previous section) in order to resolve certain frequent types of coordinated constructions with ellipsis, also called gapping. These can be of three types:.

- *solidCompound binder –partialCompound*
  e.g. binjurebarken och –märgen (*adrenal cortex and adrenal medulla*)
- *partialCompound– binder solidCompound*
  e.g. rygg– och nackvärk (*back pain and neck pain*)
- *multiW1 multiW2– binder multiW1 multiW3–Term*
  e.g. typ 1– och typ 2–diabetes (*type 1 diabetes and type 2 diabetes*)

Here, *binder* refers almost exclusively to a conjunction such as *och/and* or *eller/or*, while a few case with the adverb *som* (as/like) as a binder were also found in the corpus. 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 *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*. By applying the process to the MEDLEX Corpus, 25,000 coordinations could be detected, 6,000 of those didn't receive a MeSH label and approx. 2,000 consisted of either simplex words or the compounds were not segmented by our segmenter. A random sample of 300 of those showed that 12 (4%) were restored erroneously due to complex non-elliptic compounds with multiple segmentation points, for which our method chose the rightmost one which appeared to be (partially) wrong, e.g. *fyra||stadiet eller åtta||cells||stadiet* (four-cell stage or eight-cell stage).

4.3 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 or 3 could provide compatible results. For instance, the misspelled *accneärr* (*Acne Keloid*) which could be matched to *akneärr* with LD=2. By this approach and after manual inspection we actually chose to add the very frequent spelling errors in the thesaurus itself (e.g. *aerophagi* with LD=1 to the term *aerophagy*). The method is also applied *on the fly* while indexing arbitrary texts.

4.4 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)*, where *HUS* also stands for the Swedish common noun *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 MeSH-annotations 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 the 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. For instance, the pattern *<MeSH-term> (UUUl)*[2] which can match *RNA-interferens (RNAi)* (RNA Interference). In previous studies based on Swedish data the most frequent acronym patterns found were 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).

5. Annotation and Coverage

Each identified MeSH term is annotated using a simple metadata scheme, based on XML technology 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 (neg, with values *yes* or *no*), this attribute is currently not used but is planned for use in the near future. 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>*

[2] U as *UPPER*-case word and l as *low*-case word.

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>*

In order to empirically investigate the coverage of the resources, by applying the previously discussed transformations and text processing, 50 random articles published by *Läkartidningen*, the Swedish Medical Association's official magazine (<http://www.lakartidningen.se>) during 2006-07 under section "Nya Rön" (New Findings), which usually contain a large portion of terminology, were automatically extracted and annotated. These documents are part of a manually inspected annotated corpus we are currently building (*cf.* Kokkinakis, 2008). In this sample we discovered some cases that had a negative effect on the evaluation results and for which we have not dealt with during the previous steps. For instance, 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*. Also other forms of elisions were observed, e.g. in MeSH there is the term *chikungunya virus* but in the texts we could only find *chikungunya*.

Manual inspection of the obtained results was performed, which showed a coverage of 74,7% considering a possible total of 2,516 terms, out of which 1,688 were completely covered, true positives (including new acronyms and terms as described earlier), and 391 were partially covered by MeSH. These were scored as half correct (0.5) if half of the term was covered by MeSH, e.g. *kronisk <trötthet>* (chronic fatigue); 0.3 if one third or less was covered, e.g. *lindrig <tyreoidea>rubbning* and 0.7 if more than two thirds were covered, e.g. *<color> duplex <sonography>*. A total of 437 terms (17.3%) were false negatives, left unannotated, e.g. *kalcipotriol*, *rimonabant*, which hints on the limitations of MeSH in terms of its coverage. The number of false positives, spuriously identified concepts, was low, 46. The majority of these cases are due to homography with non-medical words and are highly context dependent, 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. The number of acronyms (36), the new terms (67) and the large number of terms originating from the English MeSH (265) in the sample designates that the effort put into the pre-processing stages pays off both quantitatively and qualitatively. Some other, but less frequent cases had to do with the problem of multiple occurrences of acronyms with the same surface form within the same document but with different semantics. For instance the case of *BNP* in the fragment shown below:

> *…beslutsgränsvärdet för BNP (»brain natriuretic peptide« eller »B-type natriuretic peptide«) torde härröra från Maiselgruppens undersökning av patienter som söker akut för dyspné (breathing not properly , »BNP«) ."* (Läkartidningen, vol. 103:19, (2006).

Another frequent problem has to do with compound segmentation. Specifically for cases for which the modifier ends in double consonant and the head starts with the same consonant as well, for instance, *galläckage* (bile leaking) and *skabbehandling* (treatment for scabies). Here, the first example actually stands for *gall+läckage* and the second for *skabb+behandling*. Our compound analyser did not segmented the first compound while the second was segmented as *skab||behandling* and *skab* could not be matched to the actual MeSH entry *skabb* since the approximate string matching approach ignores string less than 7 characters long. Another difficulty arises when near synonyms of MeSH terms are used in text. For instance, *viktökning* (Weight Gain) is a MeSH term but the near synonym *viktuppgång* is not. In these cases a general thesaurus such as WordNet (Fellbaum, 1998) can be helpful, and once again compound analysis can play an important role for aiding the matching process between *ökning* and *uppgång*. Another important issue has to do with simplex head words that exist in MeSH only as heads of solid compounds. For instance, *retinopati* (retinopathy) exists in MeSH only in compound forms such as *prematuritetsretinopati* (retinopathy of prematurity) but not as a simplex term that can be found in corpora. Finally, it is noticeable that there are a number of rather frequent (lay) words that for reasons we are not aware of, are not covered by MeSH, such as *mage* (stomach) and *kropp* (body). However, such words do not contribute to the evaluation previously presented.

6. Conclusions

We have discussed the transformation of a medical controlled thesaurus for Swedish and English to a resource that can be applied to raw data producing high quality results. Extensive effort has been put on various aspects of the normalization process in order to cover for a large range of phenomena that have implications on coverage. Our experiments revealed some incompleteness of the Swedish MeSH® w.r.t. applying it to real data, since a number of *obvious* medical terms were left unrecognized. At the same time, simple steps

(normalization) have the ability to considerably increase coverage and thus aid the enhancement of the current gaps. Swedish is a compound language and compound analysis is a crucial step for fast accessing the partially annotated segments which can aid the enhancement and thus quality of the results. In the future we intend to utilize new means of enhancing MeSH® and look deeper into the evaluation's false negatives which might be a bit higher depending on who and how the judgment is performed since not all MeSH® categories are homogeneous.

7. References

Almeida F., Bell M. and Chavez E. (2003). Controlled health thesaurus for the CDC web redesign project. *AMIA Annu Symp Proc.* (p. 777).

Ananiadou S. and McNaught J. (2006). Text mining for biology and biomedicine. Artech House Books.

Ananiadou S. and Nenadic G. (2006). Automatic terminology management in biomedicine. In S. Ananiadou & G. McNaught (Eds.), *Text mining for biology and biomedicine*. Artech House Books.

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).

Cooper G.F. and Miller R.A. (1998). An experiment comparing lexical and statistical method for extracting MeSH terms from clinical free text. *Journal of Am Med Inform Assoc.* 5, 62–75.

Crammer K., Dredze M., Ganchev K., Talukdar P.P. and Caroll S. (2007). Automatic code assignment to medical text. Biological, Translational and Clinical Language processing (BioNLP 2007) (pp. 129-136). Prague.

Djebbari A., Karamycheva S., Howe E. and Quackenbush, J. (2005). MeSHer: identifying biological concepts in microarray assays based on PubMed references and MeSH terms. *Bioinformatics.* 21(15), 3324-3326; doi:10.1093/bioinformatics/bti503.

Douyere M., Soualmia LF, Neveol A, Rogozan A, Dahamna B., Leroy JP., Thirion B. and Darmoni SJ. (2004). Enhancing the MeSH thesaurus to retrieve French online health resources in a quality-controlled gateway. *Health Info Libr J.* 21(4), 253-61.

Fellbaum C. (1998). *WordNet: an Electronic Lexical Database*. Cambridge, Mass. MIT Press.

Hirschman L., Morgan A.A. and Yeh A.S. (2003). Rutabaga by any other name: extracting biological names. *Journal of Biomedical Informatics*. 35: 247-259. Elsevier.

Jacquemin C. and Tzoukermann E. (1999). NLP for term variant extraction: A synergy of morphology, lexicon and syntax. In T. Strzalkowski (Ed.), *Natural Language Information Retrieval* (pp. 25-74). Kluwer: Boston.

Jacquemin, C. (2001). *Spotting and discovering terms through natural language processing.* Cambridge, MA, USA: MIT Press,

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.

Kokkinakis D. (2008). A Semantically Annotated Swedish Medical Corpus. *Proceedings of the 6th Conference on Language Resources and Evalutaion (LREC)*. Marrakech, Marocko.

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

Névéol A. V. Mary, A. Gaudinat, C. Boyer, A. Rogozan and SJ. Darmoni. (2005). A Benchmark Evaluation of the French MeSH® Indexing Systems. *Springer's Lecture Notes in Computer Science*. S. Miksh, J. Hunter, E. Keravnou, (Eds.): Artificial Intelligence in Medicine. Proc. AIME. Pp. 251-255.

Névéol A., Mork J.G. and Aronson A.R. (2007). Automatic Indexing of Specialized Documents: Using Generic vs. Domain-Specific Document Representations. *Biological, Translational and Clinical Language processing (BioNLP 2007)*. Pp. 183-190. Prague.

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

Rechtsteiner A. and Rocha L.M. (2004). MeSH key terms for validation and annotation of gene expression clusters. In A. Gramada & E. Bourne (Eds), *Currents in Computational Molecular Biology. Proceedings of the Eight Annual International Conference on Research in Computational Molecular Biology (RECOMB 2004)*. (pp 212-213).

Rosario B. Hearst M.A. and Fillmore C. (2002). The Descent of Hierarchy, and Selection in Relational Semantics. *Proceedings of the ACL-02*. Pennsylvania US.

Schwartz A. and Hearst M. (2003). A Simple Algorithm for Identifying Abbreviation Definitions in Biomedical Texts. *Proceedings of the Pacific Symposium on Biocomputing (PSB)*.

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

Struble C.A. and Dharmanolla C. (2004). Clustering MeSH® Representations of Biomedical Literature. *ACL Workshop: Linking Biological Literature, Ontologies, and Databases, BioLINK 2004*. Pp. 41-48. Boston US.

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.* Vol. 39, no 1, 77-90.

Tsuruoka Y. and Tsujii J. (2003). Probabilistic term variant generator for biomedical terms, *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*. Toronto, Canada

Vintar S., Buitelaar P. and Volk, M. (2003). Semantic relations in concept-based cross-language medical information retrieval. Proceedings of the Workshop on Adaptive Text Extraction and Mining. Cavtat-Dubrovnik, Croatia.

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