A Swedish Scientific Medical Corpus for Terminology Management and Linguistic Exploration

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

Ulla Gerdin
Centre for Epidemiology
The National Board of Health and Welfare
SE-106 30, Stockholm, Sweden
E-mail: ulla.gerdin@socialstyrelsen.se

Abstract

This paper describes the development of a new Swedish scientific medical corpus. We provide a detailed description of the characteristics of this new collection as well results of an application of the corpus on term management tasks, including terminology validation and terminology extraction. Although the corpus is representative for the *scientific medical domain* it still covers in detail a lot of specialised sub-disciplines such as *diabetes* and *osteoporosis* which makes it suitable for facilitating the production of smaller but more *focused* sub-corpora. We address this issue by making explicit some features of the corpus in order to demonstrate the usability of the corpus particularly for the quality assessment of subsets of official terminologies such as the *Systematized NOmenclature of MEDicine - Clinical Terms* (SNOMED CT). Domain-dependent language resources, labelled or not, are a crucial key components for progressing R&D in the human language technology field since such resources are an indispensable, integrated part for terminology management, evaluation, software prototyping and design validation and a prerequisite for the development and evaluation of a number of sublanguage dependent applications including information extraction, text mining and information retrieval.

1. Introduction

Large domain-specific literature databases, textual collections and repositories of scientific data are necessary for empirically based linguistic investigations and for progressing R&D in the human language technology field. Several such data collections have been described in the literature in the past, mainly for French, German and English particularly in the biomedical domain, MEDLINE/PubMed being the largest and best known of these text databases, (e.g. Zweigenbaum *et al.*, 2001; Kim *et al.*, 2003; Cohen *et al.*, 2005; Kim & Tsujii, 2006). However, nothing similar is currently available for Swedish apart from a listing of indexed references to Scandinavian articles in the medical field in the database SveMed+ <http://micr.kib.ki.se/>. The biomedical literature is growing at a double exponential pace and electronic open access to the full texts, including graphics and figures has just started to rise. Sophisticated linkages between publications and data repositories or other supplementary materials increase the amount of information available still further (Hunter & Cohen, 2006). This fact has a direct implication to the need of biomedical, medical and clinical processing and filtering systems where terminology management is a key component for knowledge management, competitive intelligence, hypothesis generation and decision support (*cf.* Bodenreider, 2006; Chen *et al.*, 2005).

We start by providing a general description of the corpus (Section 2). The extraction process from various formats to a standardised, text-based utf8 format, including the linguistic annotation is given in Section 3, while linguistic exploration can be currently carried out using a semantically-oriented concordance facility (Section 5). The corpus has been also used for various terminology management activities, namely terminology validation and terminology extraction and Section 4 provides some notes on these two activities.

2. Corpus Description

The corpus we have developed comprises the electronic archives of "Läkartidningen" the *Journal of the Swedish Medical Association* (<http://www.lakartidningen.se/>) one of the most reliable sources for accurate scientific medical findings and medical research in Sweden. The Journal has a long history and tradition in publishing high quality scientific medical articles in Swedish (including health economic analysis reports, historical medical outlooks, medical (bio) medical, clinical and pharmaceutical news, etc.) and was launched at the end of 1903 with the name "Allmänna svenska läkartidningen", *General Swedish medical journal*.

Since 1996, volume 93, the archive's content is available in electronic form (*pdf-files*), while since 2006, volume 103 and beyond, the electronic editions are also produced using other formats (*.xml* and *.html)* which are easier to manage from a natural language processing perspective. All electronic issues have been manually indexed with the Swedish MeSH thesaurus. Open electronic access to the full texts was granted a couple of years ago and searchable editions of all the material since 1996 are available from: <http://ltarkiv.lakartidningen.se/> while the processed archive (corpus) is searchable through a suitable interface that can be found here:: <http://www.medicinskkorpus.se/login.phtml> (free access). The archive's content is not only an important source of knowledge for health care professionals (specialists, clinicians), but also a source of information for general interest readers who wish to acquaint themselves with new findings and developments in different parts of the

	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Articles	2345	2116	2089	1779	1908	1940	2159	2150	2201	1802	1909	2001	1815	1676
Only web-articles	---	---	---	---	---	---	---	---	---	---	32	2	93	93
Commented arts	---	---	---	---	---	---	---	---	---	---	11	89	187	292
Tokens	2,027	1,988	2,207	2,085	2,011	2,108	2,023	1,763	1,845	1,517	1,594	1,654	1,761	1,716
Tokens without punct./numerical	1,754	1,723	1,914	1,807	1,745	1,822	1,756	1,528	1,589	1,308	1,374	1,418	1,509	1,468
Hapax legomena	70802	69029	73679	70414	66679	67939	63227	57036	61453	53620	54192	56522	60202	57589
Type/token ratio	0,73	0,72	0,69	0,71	0,70	0,68	0,66	0,69	0,72	0,76	0,73	0,74	0,74	0,73
Avg word length	5,95	5,90	5,82	5,86	5,91	5,88	5,92	5,92	5,98	5,97	5,97	6,01	6,00	5,97
Total MeSH tags	239kb	230kb	248kb	233kb	231kb	241kb	230kb	198kb	220kb	180kb	192kb	202kb	214kb	205kb

Table 1. Some characteristics of the corpus; (average word length in a Swedish balanced corpus is 5,3).

medical knowledge domain. The corpus we have currently assembled and processed comprises at the time of writing this paper to 28,110 different documents/ articles, approximately 28,2 million tokens. Table 1 shows some descriptive characteristics of the corpus.

3. Corpus Processing

The archive comes in many format flavors. Earlier issues in *pdf*-format while the most recent ones in *.html* and *.xml*. Although the non-pdf editions of the archive were rather unproblematic for the subsequent automatic language processing, the pdf-files posed certain difficulties due to the complexity of the layout of the journal's pages and the different pdf-versions that the material was encoded in. However, all material has been transformed to a unified utf8 text-format. The extraction was made in a semi-automatic fashion with manual verification, since our aim was to preserve as much as possible of the logical text flow and eliminate the risk for losing valuable information such as each article's title, publication details for each issue etc. An example of a typical page under the heading "Korrespondens" *Correspondence*, is given in (Figure 1).

Figure 1: A snapshot of the metadata extraction process.

By identifying and annotating the title of each article we can also benefit from the already MEDLINE-like MeSH-indexed versions of the electronic material. This way we can take advantage of the manually assigned indexes and ease the creation of various specialized sub-corpora. It is also possible to compare how well automatic indexing performs compared to the manual, an exercise we have left for the future.

3.1 Metadata Extraction and Corpus Genres

From the electronic editions we have extracted a number of metadata that in the future we intend to enhance and encode using the Text Encoding Initiative (TEI5) Lite standard <http://www.tei-c.org/index.xml>. or other XML-based formats such as SciXML, a representation for the logical structure and essential formatting of research papers (Rupp *et al.*, 2006). For this reason we have currently chosen to only use a rather flat and simplistic metadata structure using generic identifiers, such as "figDesc", "title", "issue" and "date" (publication date) which have a direct resemblance to TEI5 labels (Figure 2).

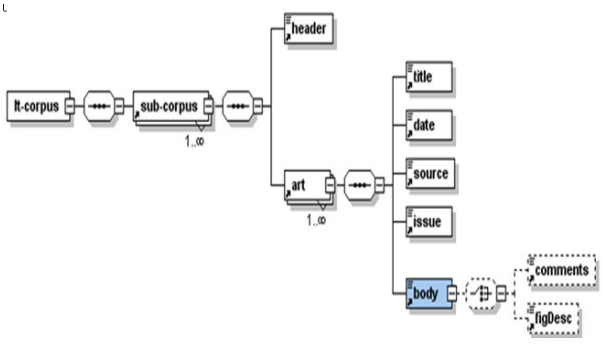

Figure 2. Flat metadata scheme

Explicit identification of genre is important for large digital collections since it provides useful information as to the document's purpose and may enhance searching, retrieval and better matching of users' needs.

Since we have also extracted metadata regarding genre (Figure 1) for the majority (85%) of the documents it is now easier to make genre selection upon searching for a term or a particular phenomenon. Genres include "New Findings", "Clinic and Science", "Patient safety", "Medicine and society", "Debate" etc. The extraction of all metadata was performed using the ABBYY PDF Transformer <http://pdftransformer.abbyy.com/> which could

extract all text zones with good performance.

3.2 Linguistic Processing

The entire corpus has been tokenized and sentences have been identified using generic software enhanced with domain specific patterns; e.g. a typical pattern in the corpus is references to bibliographic citations in the text such as "*J Biological Psychiatry doi: 10.1016/j. biopsych. 2007.03.005 (2007)*". A generic statistical part-of-speech tagger, TnT, (Brants; 2000) trained on general Swedish, was also applied after enhancement using a domain specific vocabulary of over 4000 new word forms. These forms were carefully chosen after manual review of randomly annotated domain text samples. They were word forms that seem problematic for the part-of-speech tagger's general language model and for which the tagger constantly annotated erroneously. For instance, various sublanguage words ending in '-ens', which is a typical genitive suffix in general Swedish, for which a large number of exceptions (forms that are not genitives) are predominant in the medical language; such as "*inkontinens*" (incontinence), "*prevalens*" (prevalence) and "*insufficiens*" (insufficiency); and various word forms ending in '-it', which is a typical verb suffix in general Swedish, for which a large number of exceptions (forms that are not verbs) are predominant in the medical language; such as "*perikardit*" (pericarditis), "*myokardit*" (myocarditis) and "*uveit*" (uveitis).

The processing included also lemmatization as well as named entity recognition using a detailed entity hierarchy (Kokkinakis, 2004). Named entities are particularly helpful for reducing the quantity of generated n-grams from the statistical analysis of the corpus (e.g. term extraction) for which named entities, numerical and time expressions can be filtered out. Named entities can be also used as features when searching in the available concordance interface (Figure 4). Finally the corpus has been automatically annotated with the Swedish and English MeSH thesauri as well as part of the Swedish SNOMED CT. Section (4) gives a description on how the terminology annotation using MeSH and SNOMED CT is accomplished, and which mechanisms are applied for capturing the term variation; for a detailed description see Kokkinakis (2009) and Kokkinakis & Gerdin (2009).

3.3 Subcorpora

Since all articles are indexed by MeSH it is rather straightforward to apply various techniques that enable the creation of more homogeneous subsets. For instance we have used the vector space model (Salton *et al.*, 1975) and the MeSH labels as features for document representation and off-the-shelf tools for clustering experiments (HCE; Seo & Shneiderman, 2002).

4. Terminology Management

We have used the corpus for term validation of *subsets*[1] of the Swedish SNOMED CT translation in the areas of *diabetes* and *heart problems*. It is well known that even within the same text, a term can take many different forms, and we have developed methods to deal with the variation. This rich variety for a large number of term-forms is a stumbling block especially for natural language processing, as these forms have to be recognized, linked and mapped to terminological and ontological resources (Krauthammer & Nenadic, 2004; Tsujii & Ananiadou, 2005; Tsuruoka *et al.*, 2008).

We have both used subsets of documents as outlined in section 3.3 to test the coverage of the SNOMED CT subsets as well as the whole document corpus since the results we obtained using subcorpora with respect to coverage were rather poor. Still, however, using the whole corpus and despite our efforts to apply various mechanisms for term variation the results we obtained showed rather low coverage figures. Based on 2,841 terms in the two subsets only 373 or 13,12% of these terms could be identified in the whole corpus.

4.1 Term Variation

A large number of variation patterns have been developed and extensively tested on the corpus. The most important variations we have dealt with, and which in all cases is meaning preserving, included: *morphological* variation (e.g. inflection); *permutations* of various types (e.g. structural which captures the link between a compound noun/term and a noun phrase containing a right-hand prepositional phrase); *compounding* (the inverse of the above, in which a noun phrase containing a right-hand prepositional phrase is re-written to a single-word compound); *modifications* and *substitutions* of various types (i.e. transformations that associate a term with a variant in which the head word or one of its argument has an additional modifier such a hyphenation, substitution of Arabic to Roman numbers, deletion of punctuation from length multiword terms etc); *partial matching* of a term (by applying automatic compound segmentation and try to match part(s) of the compound segmented). Figure 3 shows an example of term variation for the term *stress* based on the corpus content. In the term exploration window you can see the observed frequency distribution of the term occurrences from 1996-2009. The term could be found in 385 variant forms, including a large number of compounds in which *stress* is either a head or modifier.

4.2 Term Extraction

We have also used the corpus for automatic term recognition and we have tested a number of term recognition methods that have been suggested in the literature for unigrams, bigrams, trigrams and n-grams. The methods we tested for unigrams was the *weirdness measure*

[1] Because SNOMED CT is a large terminology it is sometimes necessary to define subsets for various use cases and specific audiences; *cf.* Patrick *et al.*, 2008. Subsets are sets of concepts, descriptions and/or relations that share a specified common characteristic or common type of characteristic and are thus appropriate to a particular user group, specialty, organization, dialect and context (for constraining choices, e.g. diabetes or osteoporosis datasets).

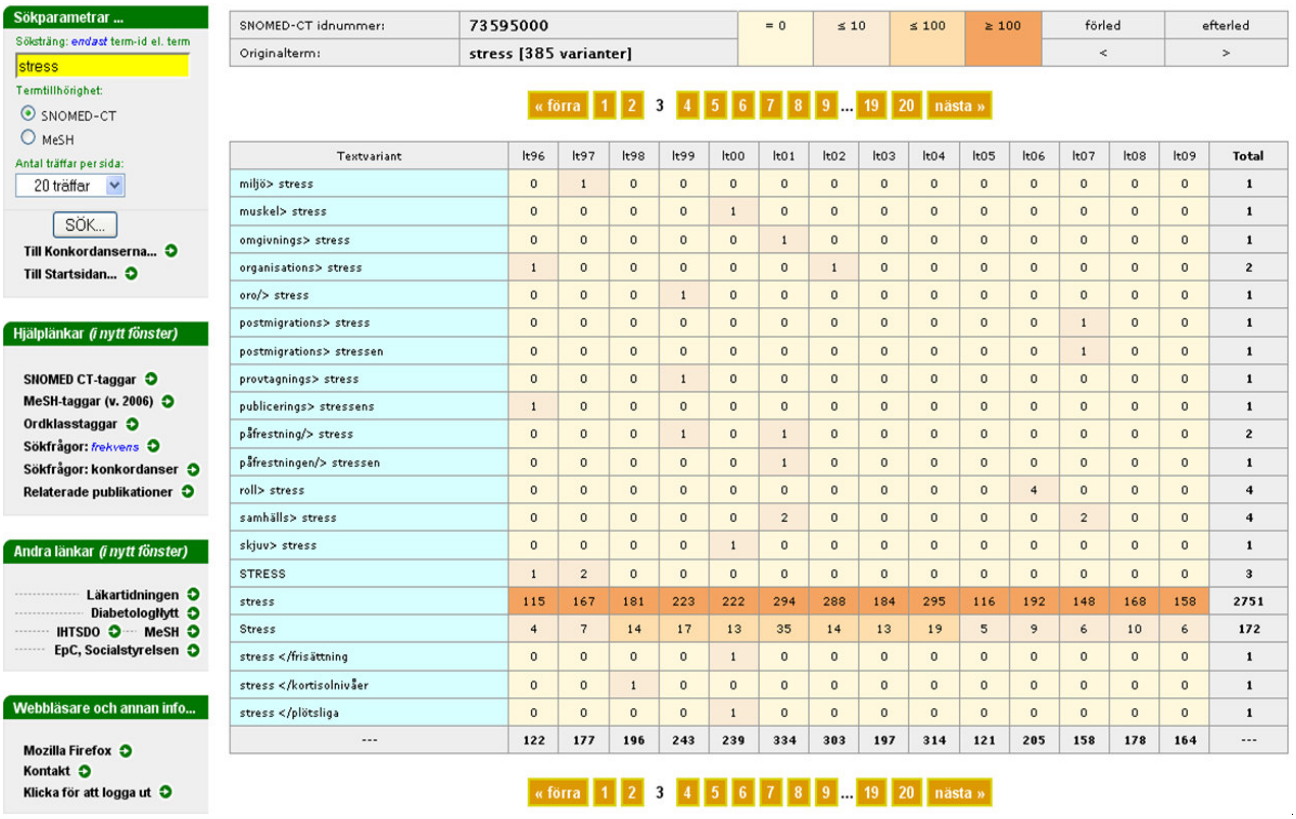

Figure 3. The term stress [73595000] (here shown 20 out of 385 variants).

by Gillam *et al.* (2005); various methods for significant bigram and trigram terms by Banerjee & Pedersen (2003) and a method for multiword terms, the *C-value* described in Frantzi *et al.*, (2000).

The results from the unigram processing (top-100 candidates) were not satisfactory and revealed a couple of major drawbacks. A large number of acronyms, and adjectival modifiers (e.g. "diabetisk" *diabetical*) were suggested as candidates, while a large number of the proposed nouns were part of multiword terms, (e.g. mellitus). The majority of the top ranked bigram and trigram candidates were rather reliable terms and Pointwise Mutual Information was the measure that returned most reliable results (looking at the top-100). Finally, for multiword terms the C-value method was applied which utilizes a linguistic filter to extract word sequences likely to be terms, particularly simple and complex noun phrases based on part of-speech tags sequences. The majority of the proposed candidates longer than 4 tokens were actually English terms such as *intrinsic cardiac nervous system* and *latent autoimmune diabetes in adults*. Although a detailed evaluation of each term extraction algorithm has not been performed for all candidate terms extracted, it is noteworthy that the results obtained by the C-value were rather poor with respect to ≥ 4 tokens long candidates. Furthermore, we didn't proceed to apply the NC-value, an extension to C-value, which incorporates information of context words into term extraction. We believe that the syntactic patterns used by the C-value method are insufficient to carry out term recognition in Swedish basically because the *noun noun* pattern is not common in a compounding languages as Swedish, compared to English, in which single word compound is the norm. Perhaps other methods are more suitable and may be explored in the future (*cf.* Sclano & Velardi, 2007).

5. Linguistic Exploration through Semantically Enabled Concordances

Currently the corpus can be exploited using standard concordance facilities enhanced by the possibility to combine various features such as part-of-speech labels, lemmatised word forms and labels from an extended named entity hierarchy. Moreover, the corpus has been indexed with the Swedish and English Medical Subject Headings (MeSH) thesaurus and for the sake of term validation with subsets of the ongoing translation of SNOMED CT. As a backend for the concordance facilities we are using the powerful indexing mechanism provided by the IMS Open Corpus Workbench (CWB; Christ, 1994). Figure 4 shows an example in which the search pattern has been: *[pos='A.*'][snm='73595000']*, which can be paraphrased as *return all contexts from the 1997 issue of the corpus where the first word's part-of-speech is an adjective or participle followed by a word (or part of a compound word) with the SNOMED CT-label 73595000 i.e. the code for the term 'stress'.*

6. Conclusions

In this paper, we have provided a description of a Swedish corpus in the domain of scientific medicine. The corpus has been already applied for *term validation* and *term recognition*, which was the primary reason for its compilation. For term validation we developed different

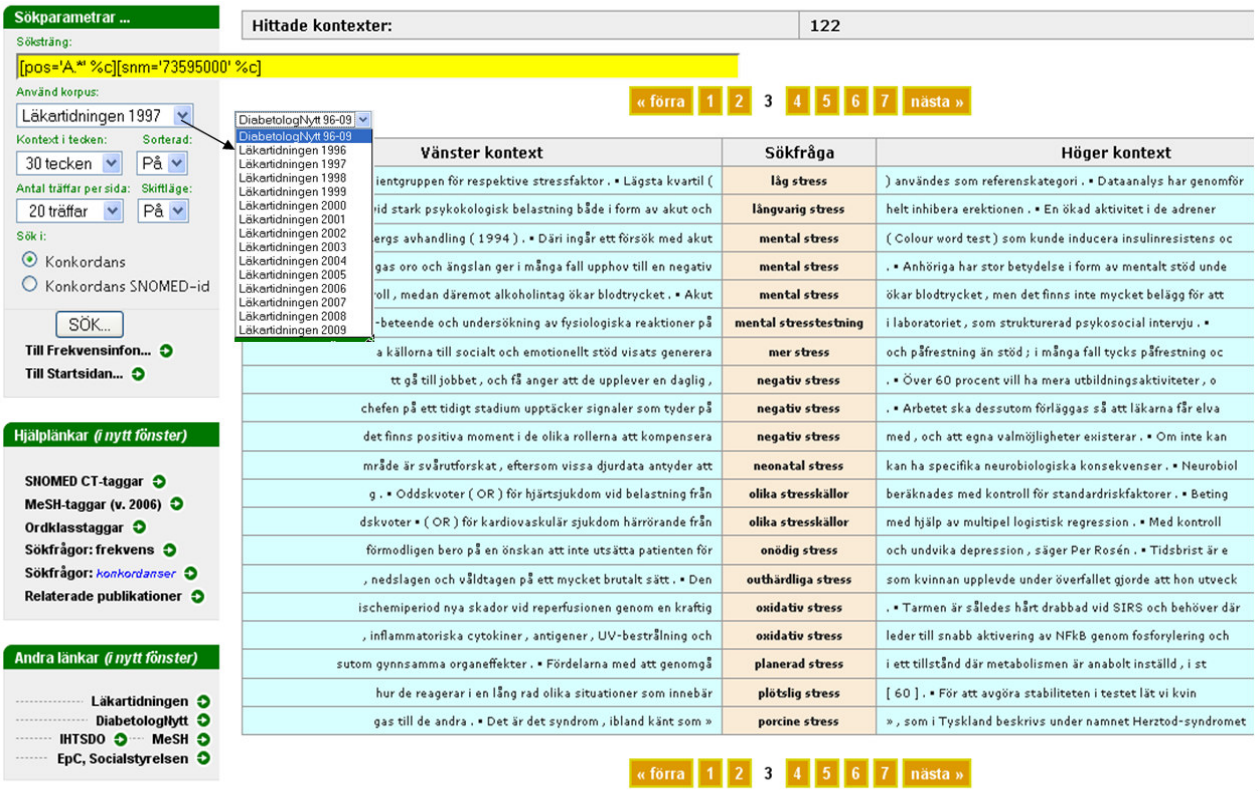

Figure 4. Linguistic exploration through concordances.

methods to cope with term variation. Still, however, only a very small fraction of the official terms could be found in the corpus and we can only speculate for an answer. Are the recommended terms so unusual? Are they paraphrased in other more complex ways? Are parts of standard terminologies based on introspection? Or is the corpus too limited or unsuitable for obtaining a better coverage? At the moment we do not have answers to such questions, presumably other means of qualitative types of analyses are necessary in order to obtain a clear picture of this performance, which are out of the scope for this paper.

Nevertheless, we believe that the quality of the corpus content can be used for other types of empirical research and linguistic exploration such definition extraction (*cf.* Westerhout, 2009) and fact extraction (*cf.* Agichtein & Gravano, 2000), issues we would like to explore in the future.

7. Acknowledgements

We would like to thank the editors of the Journal of the Swedish Medical Association for making the electronic versions available for the creation of the corpus.

8. References

Agichtein E. and Gravano L. (2000). *Snowball: Extracting Relations from Large Plain-Text Collections*. The 5th ACM Conference on Digital Libraries. Texas, US. Pp. 85-94.

Banerjee S. and Pedersen T. (2003). *The Design, Implementation and Use of the Ngram Statistics Package*. The 4th Conference on Intelligent Text Processing and Computational Linguistics. Mexico.

Bodenreider O. (2006). Lexical, terminological and ontological resources for biological text mining. In *Text Mining for Biology and Biomedicine*. Ananiadou S. and McNaught J., eds. (Norwood, MA: Artech), pp. 43–66.

Brants T. (2000). *TnT - A Statistical Part-of-Speech Tagger*. 6th Conference on Applied Natural Language Processing. Seattle, Washington, USA. 224–231.

Chen H., Fuller S.S., Friedman C. and Hersh W. (2005). Knowledge Management, Data Mining, and Text Mining in Medical Informatics. *Medical Informatics*. Vol 8. Springer.

Christ O. (1994). *A modular and flexible architecture for an integrated corpus query system*. COMPLEX'94, Budapest.

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.

Frantzi K., Ananiadou S. and Mima H. (2000). Automatic recognition of multi-word terms: the C-value/NC-value method. *J. on Digital Libraries*, Vol. 3:2. Pp. 115-130.

Gillam L., Tariq M. and Ahmad K. (2005). Terminology and the construction of ontology. *Terminology*, 11:55–81.

Hunter L. and Cohen K.B. (2006) Biomedical language processing: what's beyond PubMed? *Molecular Cell* 21:589-594.

Kim J.-D., Ohta T., Tateisi Y. and Tsujii J. (2003). GENIA Corpus - a Semantically Annotated Corpus for Bio-textmining. *Bioinformatics* Vol. 19:1. Pp 1180–1182.

Kim J.-D. and Tsujii J. (2006). Corpora and their Annotation. In *Text Mining for Biology and Biomedicine*. Ananiadou S. and McNaught J., eds. (Norwood, MA: Artech), pp. 179-211

Kokkinakis D. (2004). *Reducing the Effect of Name Explosion.* LREC Workshop: Beyond Named Entity Recognition, Semantic labelling for NLP tasks. 4th Language Resources and Evaluation Conference (LREC). Lissabon, Portugal.

Kokkinakis D. (2009). Lexical granularity for automatic indexing and means to achieve it – the case of Swedish MeSH®. *Information Retrieval in Biomedicine: Nat. Language Processing for Knowledge Integration*. Prince V. and Roche M. (eds). IGI Global. pp. 11-37.

Kokkinakis D.and Gerdin U. (2009). *Issues on Quality Assessment of SNOMED CT® Subsets - Term Validation and Term Extraction*. Proceedinsg of RANLP-2009 Worksop: Biomedical Information Extraction. Borovets, Bulgaria.

Krauthammer M. and Nenadic G. (2004). Term identification in the biomedical literature. *J Biomed Inform.* 37(6):512-26.

Nenadié G., Ananiadou S. and McNaught J. (2004). *Enhancing automatic term recognition through recognition of variation.* 20th Conf. on Computational Linguistics. Switzerland.

Patrick J. et al. (2008). Developing SNOMED CT Subsets from Clinical Notes for Intensive Care Service. *Health Care & Informatics Review Online*. Open Access.

Rupp CJ., Copestake A., Teufel S. and Waldron B. (2006). *Flexible Interfaces in the Application of Language Technology to an eScience Corpus.* UK e-Science Programme All Hands Meeting. Nottingham, UK.

Salton G., Wong A. and Yang C.S. (1975). A Vector Space Model for Automatic Indexing. *Com of the ACM*, vol. 18:11, pp 613–620. <http://www.cs.uiuc.edu/class/fa05/cs511/Spring05/other_papers/p613-salton.pdf>

Sclano F. and Velardi P. (2007). *TermExtractor: a Web Application to Learn the Common Terminology of Interest Groups and Research Communities*. 9th Conf. on Terminology & AI. Sophia Antinopolis.

Seo J. and Shneiderman B. (2002). Interactively Exploring Hierarchical Clustering Results. *IEEE Computer*, Volume 35:7, pp. 80-86.

Tsujii J. and Ananiadou S. (2005). Thesaurus or Logical Ontology, Which One Do We Need for Text Mining? *J. of Language Resources and Evaluation.* Pp 77-90. Vol. 39:1.

Tsuruoka Y., McNaught J. and Ananiadou S. (2008). Normalizing biomedical terms by minimizing ambiguity and variability. *BMC Bioinformatics* 2008, 9(Suppl 3):S2.

Westerhout E. (2009). *Definition extraction using linguistic and structural features*. Proceedings of the RANLP Workshop on Definition Extraction. Bulgaria.

Zweigenbaum1 P., Jacquemart P., Grabar N. and Habert B. (2001). *Building a Text Corpus for Representing the Variety of Medical Language.* MEDINFO 2001, Pp. 290-294. IOS.