Issues on Quality Assessment of SNOMED CT® Subsets – Term Validation and Term Extraction

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

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

Abstract

The aim of this paper is to apply and develop methods based on Natural Language Processing for automatically testing the validity, reliability and coverage of various Swedish SNOMED-CT subsets, the *Systematized NOmenclature of MEDicine - Clinical Terms* a multiaxial, hierarchical classification system which is currently being translated from English to Swedish. Our work has been developed across two dimensions. Initially a Swedish electronic text collection of scientific medical documents has been collected and processed to a uniform format. Secondly, a term processing activity has been taken place. In the first phase of this activity, various SNOMED CT subsets have been mapped to the text collection for evaluating the validity and reliability of the translated terms. In parallel, a large number of term candidates have been extracted from the corpus in order to examine the coverage of SNOMED CT. Term candidates that are currently not included in the Swedish SNOMED CT can be either parts of compounds, parts of potential multiword terms, terms that are not yet been translated or potentially new candidates. In order to achieve these goals a number of automatic term recognition algorithms have been applied to the corpus. The results of the later process is to be reviewed by domain experts (relevant to the subsets extracted) through a relevant interface who can decide whether a new set of terms can be incorporated in the Swedish translation of SNOMED CT or not.

Keywords

Quality Assessment; Term Validation; Automatic Term Recognition; SNOMED CT; Scientific Medical Corpora.

1. Introduction

The purpose of the current paper is to provide an introduction and description of the methodology for the validation and quality assessment of the ongoing Swedish translation of the *Systematized NOmenclature of MEDicine – Clinical Terms* (SNOMED CT). The translation of SNOMED CT is part of the Swedish strategy for e-health and is expected to facilitate both interoperability between health- and social care systems and communication between health- and social care professionals in clear and unambiguous concepts and terms. SNOMED CT is a very large and systematically organized computer processable collection of health and social care terminology. The main aim of our work is to develop and apply Natural Language Processing (NLP) techniques for automatically mapping structured SNOMED CT concepts to unrestricted texts in order to evaluate the validity and reliability of the translated terms. Also, algorithms for suggesting new candidate terms are currently being explored and may benefit the translation work.

The material used in this work is based on large samples of scientific medical data that cover a broad spectrum of medical subfields. Currently, the medical corpus consists of two main parts. The first part consists of the electronic editions from the latest 14 year publications of the Journal of the Swedish Medical Association, *Läkartidningen*, (<http://www.lakartidningen.se/>). The second part of the corpus consists of electronic editions of a Swedish diabetes journal, *DiabetologNytt*, (<http://diabetologNytt.se/>). The corpus is used as a testbed for exploring and measuring the coverage and quality related to the translated concept instances as well as for applying various term extraction techniques, before new services based on SNOMED CT are launched.

By applying an empirical approach to the validation of the Swedish translation of various SNOMED CT subsets, we aim to explore issues related to the:

- provision of concrete actions for evaluation of the quality of the translated recommended terms;
- identification of potential problems or shortcomings related to the choice of recommended terms;
- design of activities to overcome such potential deficiencies by e.g. suggesting sets of potential term candidates for future inclusion in the resource;
- follow-up monitoring to ensure effectiveness of corrective steps;
- (possibility to) measuring the quality of translations and comparing over time; as SNOMED CT and the corpus evolve.

Workshop Biomedical Information Extraction 2009 - Borovets, Bulgaria, pages 27–34

This paper will put emphasis on the first three of these issues. The rest of this document provides a short, general overview of SNOMED CT (Section 2) and its characteristics, the textual resources developed for the task (Section 3), as well as methodological issues related to the validation process of subsets of the Swedish translation of SNOMED CT. Moreover, a number of automatic term recognition (or term extraction/mining) techniques have been tested for the purpose of suggesting candidate terms to be included in SNOMED CT after inspection by domain experts.

2. SNOMED CT®

SNOMED CT, the *Systematized Nomenclature of Medicine Clinical Terms*, is a common computerized language, a so called "compositional concept system" which means that concepts can be specialized by combinations with other concepts, e.g. by post-coordination which describes the representation of a clinical meaning using a combination of two or more concept identifiers; [1]. This way a single expression consisting of several concepts related by attributes, such as *finding site* and *severity* can be created; e.g. [patient] [currently] has [severe] [fracture] of [left] [shaft of femur]; [2]

During the coming years SNOMED CT will be used by all clinical and information systems in the Swedish healthcare sector in order to facilitate both interoperability between healthcare systems and communications between healthcare professionals in clear and unambiguous terms. Its primary purpose is to be used as the standard reference terminology with Electronic Health Record systems (EHR). According to AHIMA [3], SNOMED CT provides a common language that enables consistency in capturing, storing, retrieving, sharing and aggregating health data across specialties and sites of care.

Table 1. The 19 top-level SNOMED CT-hierarchies.

Body structure	Physical object
Clinical finding	Procedure
Environments geo locations	Qualifier value
Event	Record artifact
Linkage concept	Situation with explicit content
Observable entity	Social context
Organism	Special concept
Pharmaceutical biologic product	Specimen
Physical force	Staging and scales
	Substance

SNOMED CT is a clinically derived terminology, the content of which has been developed by clinical groups, mainly by the College of American Pathologists (CAP, <http://www.cap.org/>). SNOMED CT combines the content and structure of the SNOMED Reference Terminology/RT with the United Kingdom's National Health Service – NHS – Clinical Terms version 3. SNOMED CT covers most areas of clinical information and according to the international release of July 2008, it includes more than 315,000 active concepts, where each concept is claimed to have a semantic, logic-based definition stated in description logic[1]. SNOMED CT concepts are organized into 19 top-level hierarchies (Table 1), each subdivided into several sub-hierarchies. Moreover, SNOMED CT contains over 806,000 English language descriptions (human-readable phrases or names associated with concepts) and more than 945,000 logically-defining relationships. Each concept may have more than one descriptor, and may appear in more than one hierarchy e.g. *pneumonia* is an *infectious disease* and a *lung disease*. SNOMED CT provides a rich set of inter-relationships between concepts. Hierarchical relationships define specific concepts as children of more general concepts. For instance, *kidney disease* is defined *as-a-kind-of disorder of the urinary system*. In this way, hierarchical relationships provide links to related information about the concept. This last example shows that *kidney disease* has a relationship to the concept that represents the part of the body affected by the disorder (i.e., *the urinary system*).

2.1 IHTSDO and SNOMED CT

In April 2007 the International Health Terminology Standards Development Organization (IHTSDO, <http://www.ihtsdo.org>) acquired the intellectual property rights of SNOMED CT and its antecedents from the College of American Pathologists. IHTSDO is a non-profit association under Danish Law and it is established by a group of nine founding nations (Australia, Canada, Denmark, Lithuania, The Netherlands, New Zealand, Sweden, the United States and the United Kingdom). By acquiring the SNOMED CT, the IHTSDO and its member countries, will help to ensure the continued maintenance and evolution of SNOMED CT as well as its availability on an international scale. The IHTSDO assumed responsibility for the ongoing maintenance, development, quality assessment, and distribution of SNOMED CT. In Sweden the Swedish National Board of Health and Welfare (*Socialstyrelsen*, <http://www.socialstyrelsen.se>) runs the projects that in a few years time will provide a Swedish translation and a release centre with methods, routines, support and organization for national maintenance of SNOMED CT.

3. Materials and Methods

3.1 Corpus

This section provides a description of the material developed and used for this work which comprises two

[1] See [4] for a critical review of the SNOMED CT's logic based definitions of concepts.

major components: a new, large, Swedish electronic scientific medical textual corpus and two subsets of SNOMED CT, namely one related to *diabetes* and one to *heart problems*.

For the first phase of the validation process of the Swedish translations it was a prerequisite to have the appropriate textual collection to use as a testbed for indexing with SNOMED CT and then make it available to domain experts for further analysis. The archives of the *Journal of the Swedish Medical Association* are one of the most reliable sources for such exploration. Since 1996, volume 93, the archive's content exists in the form of pdf-files, while the last four years, volumes 103-106, electronic editions are also produced using other, easier to process formats such as *.xml* and *.html*. Table 2 shows some characteristics of this corpus which currently comprises 26 945 different articles and 25.5 million tokens (roughly 21.8 million words, tokenised excluding punctuation).

Table 2. Characteristics of the Swedish Medical Association Journal corpus.

YEAR	ARTICLES		
1996	2342	TOKENS	2 058 797
		WORDS	1 759 496
1997	2122	TOKENS	2 015 640
		WORDS	1 727 694
1998	2090	TOKENS	2 234 777
		WORDS	1 918 119
1999	1779	TOKENS	2 108 235
		WORDS	1 810 314
2000	1909	TOKENS	2 036 670
		WORDS	1 747 848
2001	1940	TOKENS	2 132 462
		WORDS	1 825 819
2002	2159	TOKENS	2 051 456
		WORDS	1 759 481
2003	2150	TOKENS	1 791 249
		WORDS	1 531 787
2004	2201	TOKENS	1 873 986
		WORDS	1 592 125
2005	1802	TOKENS	1 540 840
		WORDS	1 310 321
2006	1984	TOKENS	1 656 268
		WORDS	1 408 627
2007	2042	TOKENS	1 712 602
		WORDS	1 448 409
2008	1915	TOKENS	1 793 178
		WORDS	1 518 303
2009	510	TOKENS	532 776
		WORDS	451 832

0K 500K 1000K 1500K 2000K 2500K

As a complement to this material we have also integrated yet another subdomain specific corpus from a Swedish Diabetes Journal, *DiabetologNytt*. This corpus, which is much smaller than the previous one, also covers published issues from 1996 up to the beginning of 2009 and consists of 861 different articles and 950,000 tokens (820,000 words).

3.2 Corpus Processing

Although the non-pdf editions of the Swedish Medical Association's Journal are rather unproblematic for the subsequent NLP processing, the pdf-files pose certain difficulties due to the complexity of the layout of the journal's pages and the different pdf-versions that the material is encoded in. However, all material has been transformed to a unified UTF-8 text-format. The extraction was made in an 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 and publication details of each issue. By identifying and annotating the title of each article we can also benefit from the already MEDLINE-like MeSH-indexed version of the material which can be found at: <http://ltarkiv.lakartidningen.se/>. This way we can take advantage of the manually assigned indexes and ease the creation of various specialized subcorpora, e.g. *diabetes*. Sentence identification, tokenization and lemmatization were also part of this step. In order to reduce the quantity of generated n-grams from the statistical analysis of the corpus (section 5) we have also applied named entity recognition on the corpus in order to filter out named entities as well as numerical and time expressions.

Figure 1: Snapshot of the Swedish Medical Association Journal's layout.

3.3 SNOMED CT Subsets

Because SNOMED CT is a large terminology it is sometimes necessary to define *subsets* for various use cases and specific audiences; *cf.* [5]. 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 (UK *vs.* American English) and context (for constraining choices, e.g. *diabetes* or *osteoporosis* datasets). SNOMED CT provides such a

mechanism that is of particular interest at the translation stage, its implementation and actual use; *cf.* the SNOMED CT - User Guide, page 6-4. Thus, the creation and maintenance of appropriate subsets, navigational hierarchies, and application filtering techniques reduce the problem of "noise" results and eliminates inconsistencies, making the data easier to analyse; [6]

Previous evaluation of the terminology to various subsets has resulted into high figures in terms of coverage. Elkin *et al.* [7] found that 92.3% of terms used in medical problem lists could be exactly represented by SNOMED CT. Ruch *et al.* [8] reports a precision of over 80% on assigning SNOMED concepts to MEDLINE abstracts, while comparable results are also reported by [9] and [10].

4. Term Validation

Even within the same text, a term can take many different forms. Tsujii & Ananiadou [11] 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 natural language processing, as these forms have to be recognized, linked and mapped to terminological and ontological resources; for a review on normalization strategies see [12].

Another related issue is the fact that a number of necessary adaptations of the resource content itself have to take place in order to produce a format suitable for text processing, for instance indexing. 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.* [13]). Therefore a number of conversion and normalization steps have to be applied to the original data. These steps are necessary before the actual implementation of a SNOMED CT-validator due to the nature of the original data. Therefore, a great effort has been put into defining ways to deal with the variety of term realization in the data, both in the textual and lexical (taxonomic) one. Some of the many possible variation types are further described in [14: 161-219]. In short this variation, which should be in all cases *meaning preserving*, includes:

1. *morphological* variation, such as the use of inflection and derivational patterns, e.g. plural forms.
2. *permutations* of various types, such as certain forms of syntactic (structural) variations which capture the link between a term, e.g. a compound noun, such as *skin neoplasm*, and a noun phrase containing a right-hand prepositional phrase, such as *neoplasm of/in/on the skin*. Naturally, both the compound and the noun phrase should then convey the same meaning, unless these variants are lexicalized. Note that compounds in Swedish are written as a single orthographic unit, i.e. *hudtumör* ('skin neoplasm').
3. *compounding*, which is the inverse of the above, in which a noun phrase containing a right-hand prepositional phrase is re-written to a single-word compound or in the case of a two word term written as a single compound, e.g. *glomerulär filtration* ('glomerular filtration') and *glomerulusfiltration.*
4. *modifications and substitutions* of various types, that is transformations that associate a term with a variant in which the head word or one of its argument has an additional modifier, hyphenation, e.g. *b cell* vs. *b-cell*; the substitution of Arabic to Roman numbers, e.g. *NYHA type 2* vs. *NYHA type II* or the deletion of a part of a lengthy multiword term (usually function words, punctuation or other modifiers), e.g. *diabetes mellitus type 1* vs. *diabetes type 1*.
5. *coordination*, an unambiguous transformation that associates two or more terms with a composite variant. Sometimes two or more entities are coordinated by their heads, e.g. *interleukin-1 och -6* actually *interleukin-1 och* ***interleukin****-6* ('interleukin-1 and 6') and sometimes by their arguments, e.g. *hjärt- och njursvikt* actually *hjärt****svikt*** *och njursvikt* ('heart and kidney failure'). Note that in Swedish such coordinations contain an obligatory hyphen at the end of each shortened form.
6. *partial matching* of a term, by applying automatic compound segmentation, e.g. *insulinnivå* ('insulin level'); here the compound *insulinnivå* has been segmented as *insulin+nivå*.
7. *acronyms*; e.g. *ventricular tachycardia (VT).*
8. *ellipsis* and *coreference* of various types, e.g. "*…chromosome 17. This chromosome is…*".
9. *lexico-semantic patterns*, e.g. *oftalmologisk undersökning* vs. *ögonundersökning* ('ophthalmologic examination').

4.1 Term Validation Results in Subcorpora

We have developed methods to test the first six of the previously discussed variation types using two SNOMED

CT subsets. The first belonging to the area of *diabetes* (92 terms) and the second to the area of *heart problems* (2756 terms). Thus for instance, according to the previous discussion on term variation, SNOMED CT *single word compound terms* have been automatically segmented and a new set of *noun plus prepositional phrase* alternatives have been created and tested against the corpus. In the case of *two word terms* we both changed the order of the constituents and also created a compound of the two constituents. In the case of *three word terms* with a preposition between we automatically created *single word compounds*. In the case of *permutations* we tested lengthy terms by re-ordering or even deleting individual "content empty" items, usually punctuation, conjunctions, function words and a few cases adjectival modifiers.

Table 3. Term variation in the *diabetes* subcorpus.

term variations	occurrences	example
original form	4352	insulin, stress, kolesterol
morphological variation	619	insuliner, kolesterolet
permutation	3337	typ2 diabetes *[diabetes mellitus typ 2]*
compounding	91	njurtransplantation *[transplantation av njure]*
modification – addition	13	koronar bypassoperationen *[koronar bypass]*
modification – deletion	11	fotpulsar saknades *[pulsar saknas i fot]*
modification – other	72	bt-diast *[diastoliskt blodtryck]*
partial matching	4555	*[insulin]*behandling
not in the subcorpus	32 (34.8%)	normal vibrationskänsla

Table 4. Term variation in the *heart problems* subcorpus.

term variations	occurrences	example
original form	8142	ventrikeltakykardi, angina
morphological variation	653	hjärtinfarkter, st-höjningen
permutation	67	arytmogen högerkammardysplasi *[arytmogen dysplasi i höger kammare]*
compounding	60	hjärttumör *[tumör i hjärta]*
modification – addition	35	ekg visade sinusrytm *[ekg : sinusrytm]*
modification – deletion	129	akut koronart syndrom *[akut koronart syndrom, aks]*
modification – other	17	av-block iii *[av-block 3]*
partial matching	763	*[hjärtinfarkt]*patienter
not in the subcorpus	2523(91.5%)	reumatisk pulmonalklaffstenos

Tables 3 and 4 provide information on the distribution of the variation for the two subsets in two different subcorpora. The one consisting of articles on *diabetes*, including the Swedish Diabetes Journal's texts, while the other consists of articles in the domain of *heart problems*. Acronym matching has been also been performed but due to frequent ambiguities between acronyms we decided not to suggest acronyms as variant forms. For instance, 'VT' in the corpus can stand for: *ventricular tachycardia*; *tidal volume* or *official in charge 'vakthavande tjänsteman'*. Of course, a possible solution could be to only suggest unambiguous candidates occurring over a certain threshold.

5. Automatic Term Recognition

Automatic term recognition (or term extraction/mining) techniques can be divided into two broad categories, the *unithood-based* and the *termhood-based* ones; [15]. Unithood refers to the attachment strength or stability of syntagmatic combinations or collocations. Some well studied, common measures of this approach are Pointwise Mutual Information (the co-occurrence frequencies of the constituents of complex terms are utilised to measure their dependency), the Log-Likelihood (which attempts to quantify how much more likely one pair of words is to occur compared to others) and the chi-square ($\chi 2$) test. Termhood refers to the degree that a linguistic unit actually represent (or is related to) a domain-specific concept. A common measure for termhood is the C-value/NC-value ([15]). For instance, in the eye-pathology domain "soft contact lens" is a valid term which has both high termhood and unithood, while its substring "soft contact" has high unithood and low termhood; example from [16].

Thus, the application of term extraction consists of two fundamental steps in which unithood as an important pre-requisite for termhood [17]; to identify term candidates from text (unithood), and to filter through the candidates, to separate terms from non-terms (termhood).

5.1 Term Recognition Methods

We have tested a number of methods that have been suggested in the literature. The methods we tested included a method for unigrams (the weirdness measure [18]), various methods for bigrams and trigrams [19]) and one method for multiword terms (C-value [15]).

5.1.1 Unigram Term Recognition

Gillam *et al.* [18] describe a method called *weirdness*, which compares the relative frequency of a term candidate in a domain specific corpus against its relative frequency in a general corpus, a reference corpus. A candidate that is significantly more frequent in the domain specific corpus becomes a potential term candidate. In order to cope with words that do not occur in the general language corpus the description of weirdness incorporate a simple smoothing technique, *add-one*, that adjust frequencies according to a renormalization factor.

$$\tau(w) = \frac{N_{GL} f_{SL}}{(1 + f_{GL}) N_{SL}}$$

In the above formula *w* stands for a word type, *fSL* for word the frequency in a domain corpus, *fGL* for the word frequency in a general corpus, *NSL* is the total number of words in the domain corpus and *NGL* is the total number of words in the general corpus, which in our case was a 45 million token newspaper corpus. Table 5 shows the top-10 results for unigram candidates for the *diabetes* and *heart* subdomain.

Table 5. Top-ranked unigram candidates.

candidate (w) *diabetes*	τ(w)	candidate (w) *heart*	τ(w)
metformin	2866.6	troponin	3272.2
diabetesteam	3034.6	koronar	3737.4
UKPDS	3112.9	kardiell	4047.6
SFD	3146.5	mortalitet	4776.4
hyperglykemi	3404.1	kardiomyopati	4885.0
neuropati	3650.4	randomiserad	4916.0
diabetisk	4266.3	pectoris	5800.0
mellitus	5083.7	ekokardiografi	6001.6
hypoglykemi	6584.3	systolisk	6187.7
NDR	8028.8	ischemisk	6637.4

A manual review of the top-100 extracted candidates revealed a couple of major drawbacks with this approach. For the first the number of acronyms (e.g. UKPDS) proposed was high while the percentage of adjectival modifiers (e.g. *diabetisk* 'diabetical') suggested as candidates was also very frequent. Naturally, also a long number of the proposed nouns were part of multiword terms, (e.g. *mellitus*) particularly English.

5.1.2 Bigram and Trigram Term Recognition

Table 6. Top-ranked bigram and trigram candidates.

PMI(x,y) *diab.*	score	PMI(x,y,z) *diab.*	score
24-h ambp	16.8	ligamentum carpi transversum	32.0
stenoserande tendovaginit	16.8	perkutan transluminal angioplastik	31.3
tendovaginitis stenosa	16.5	limited joint mobility	31.2
dorsalis pedis	16.5	hyperinsulinemisk euglykemisk klamp	29.2
tibialis posterior	16.5	insulin-like growth factor	27.6
PMI(x,y) *heart*	**score**	**PMI(x,y,z) *heart***	**score**
cord stimulation	15.6	hypokalemisk periodisk paralys	29.3
external counterpulsation	15.6	torakal epidural anestesi	27.6
sarkoplasmatisk retikel	15.4	främre nedåtstigande gren	27.3
spinal cord	15.4	international normalized ratio	27.2
sexminut gångtest	15.4	vena cava inferior	27.1

We have tested a number of bigram and trigram recognition measures implemented in the Text-NSP package [19].

The method that seems to achieve the most reliable results compared to other measures was Pointwise Mutual Information (PMI), which measures the strength of association between two or three words. Intuitively, PMI tells us how informative the occurrence of one word is about the occurrence of another word and co-occurrence frequencies of the constituents of complex terms are utilised to measure their dependency.

$$I_a(x,y,z) = \log_2 \frac{P(xyz)}{P(x)P(y)P(z)}$$

Table 6 shows the top-5 results for bigram and trigram candidates for the *diabetes* and *heart* subdomain.

5.1.3 Multiword Terms

The majority of the studies examined in the literature concerns two-word terms since they are considered the most important and typical in a core terminology [20]. However hybrid approaches such as the C-value/NC-value try to combine linguistics (term candidates and term formation patterns), statistics (ranking based on term length, frequency of occurrence and frequency of nested terms) and contextual information (re-ranking term candidates based on co-occurrence with significant context words) in order to suggest *multiword terms*.

We have applied the C-value method [15] on our corpus to extract multi-word terms. For the linguistic analysis we used the TnT part of speech tagger [21] trained on general Swedish corpora and enhanced with a few hundred new words which were problematic for the tagger. For instance, new words ending in *-ns* were annotated by default as genitives but in the corpus such words are rather nominatives, *insufficiens* (insufficiency) and *prevalens* (prevalence). Other words were exclusively found as adjectival modifiers in general corpora rather than nouns in the medical corpora. Alternative morphosyntactic descriptions were added for these forms in the lexicon, e.g. the homograph *terminal* (as noun – predominant in general corpora or as adjective – predominant in the medical corpus). The linguistic filter was used to extract word sequences likely to be terms, particularly simple and complex noun phrases based on part of-speech tags sequences. Our filter included common nouns, adjectives, and participles as well as 'foreign words', i.e. English or Latin words that the TnT-tagger annotates as such.

$$C\text{-}value(a) = \begin{cases} log_2|a| \cdot f(a) & a \text{ is not nested,} \\ log_2|a|(f(a) - \frac{1}{P(T_a)} \sum_{b \epsilon T_a} f(b)) & otherwise \end{cases}$$

In the above, *a* is the candidate string, *f(a)* is its frequency of occurrence in the corpus, *Ta* is the set of extracted candidate terms that contain *a* and *P(Ta)* is the number of these candidate terms. The C-value is a domain-independent method used to automatically extract multi-word terms from a corpus. It aims to get more accurate terms than those obtained by the pure frequency of occurrence methods.

Table 7 shows the results of this method for which the majority of the *Swedish* candidates extracted were 2-3 tokens long with a very few exceptions for candidates with 4 tokens. The majority of candidates longer than 4 tokens were *English* terms, e.g. "intrinsic cardiac nervous system" and "latent autoimmune diabetes in adults".

Table 7. Top 5-ranked multi-word term candidates (*diabetes* on the top, *heart problems* on the bottom of the table).

C-value	f(a)	f(b)	P(Ta)	Candidate
86.7903708047807	80	5	5	god metabol kontroll
42.8458792580563	40	67	67	diabetes mellitus typ
39.5500423920519	37	4	4	förbättrad metabol kontroll
25.2680826393665	24	2	2	ny diagnostisk kriterium
17.5777966186898	17	2	2	förbättrad glykemisk kontroll
54.7736698207386	51	8	7	refraktär angina pectoris
51.4150551096675	48	6	5	akut koronar syndrom
34.6062870930455	33	3	2	instabil angina pectoris
33.324572756266	32	5	3	stabil angina pectoris
16.4791843300216	16	1	1	tidig invasiv behandling

6. Discussion and Conclusions

In this paper, we investigated two major issues related to the quality assessment of a large terminological medical resource that is currently translated to Swedish. The two issues were *term validation* and *term recognition*. We started by developing a large scientific medical corpora to be able to apply various methods and algorithms for both purposes. Priority was given to the term validation purpose and thus it was important to develop different methods to cope with term variation. The results showed that simple means can enhance the recognition of term variants that otherwise would have been neglected during the automatic processing.

Particularly helpful have been the partial matching and various forms of structural variation. Table 8 illustrates an example for which the recommended SNOMED-CT *diabetes mellitus typ 2* has 59 (40+19) occurrences while the dominated variant *typ 2-diabetes* has 1004 (966+38) occurrences. Still, subdomain specific corpora showed that only a fraction of the recommended terms in the two subsets actually appear in the subcorpora For the diabetes subcorpora we could only find 65,2% of the terms, while for the heart problems subcorpora the corresponding figure was much lower, namely 8,5%.

Table 8. Variation for the term *diabetes mellitus typ 2*; frequencies in parenthesis are based on the entire corpus.

typ 2-diabetes (966)	typ II-diabetes (32)	TYP2-DIABETES (2)
typ 2 diabetes (838)	***diabetes mellitus typ 2 (19)***	Typ2-diabetes (2)
diabetes typ 2 (250)	Diabetes typ 2 (12)	Typ-2 diabetes (2)
Typ 2 diabetes (79)	diabetes typ II (6)	Typ II Diabetes (1)
typ-2 diabetes (60)	diabetes typ-2 (6)	Typ II-diabetes (1)
typ2-diabetes (48)	typ II diabetes (5)	TYP 2 DIABETES (1)
Diabetes mellitus typ 2 (40)	Typ II diabetes (3)	typ2 diabetes (1)
Typ 2-diabetes (38)	diabetes av typ 2 (3)	diabetes typ2 (1)

In order to assess the validity of this finding it is imperative to continue testing in much larger scale, starting by using the *whole* collected corpus we have at our disposal so far. The ability to tackle different term variation phenomena is a crucial step for enhancing the performance of automatic term recognition and term management systems [22].

With respect to the second leg of our study, that is the evaluation of term recognition approaches and determining the relevance of extracted terms is an issue we let domain experts to decide how valid the candidate terms are and we intend to engage such experts for the task. Gold standards do not exist although term recognition has been a research enterprise with a long tradition. Moreover evaluation of term recognition is a highly subjective problem domain. However, suitable inspection interfaces (Figure 2) can enhance the work of the experts and relevant feedback can provide us with enough data in order to assess the validity and correctness of the various term extraction algorithms.

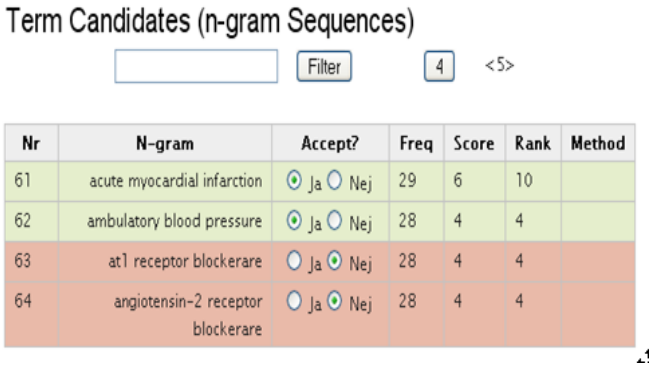

Figure 2: Term inspection interface.

Although a thorough evaluation of each term extraction algorithm has not been performed yet in a large scale, it is noteworthy that the results obtained by the C-value were rather poor with respect to ≥4 tokens long extracted 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 will be explored in the future, both with respect to multiword terms [23] and further term variation features [24].

7. Acknowledgements

We would like to thank the editors of the Journal of the Swedish Medical Association and DiabetologNytt for making the electronic versions available to this study.

8. References

[1] K. Spackman and J. Gutai. Compositional Grammar for SNOMED CT Expressions in HL7 Version 3. 2008.

[2] B. Luff and M. Bainbridge. Common User Interface Clinical Applications and Patient Safety. SNOMED CT and Interoperable Healthcare. Birmingham, UK. 2008.

[3] AHIMA. American Health Information Management Association). Statement on Implementation of SNOMED-CT. 2007. Accessed 2009-01-18, from: <http://www.ahima.org/dc/positions/documents/MicrosoftWord-StatementonImplementationofSNOMEDRevandAppro12-1-2007.pdf>

[4] S. Schulz, B. Suntisrivaraporn and F. Baader. SNOMED CT's Problem List: Ontologists' and Logicians' Therapy Suggestions. Medinfo 2007, Studies in Health Technology & Informatics. IOS Press. 2007.

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

[6] D. Walker. GP Vocabulary Project – stage-2; SNOMED CT®; Report-2. 2004. Accessed 2009-01-20, from <http://www.adelaide.edu.au/health/gp/rsearch/ current/vocab/2_02_2.pdf>

[7] P.L. Elkin *et al.* Evaluation of the Content Coverage of SNOMED CT: Ability of SNOMED Clinical Terms to Represent Clinical Problem Lists. Mayo Clin Proc. 81(6):741-748. 2006.

[8] P. Ruch, J. Gobeill, C. Lovis and A. Geissbühler. Automatic medical encoding with SNOMED categories. BMC Medical Informatics and Decision Making 2008, 8 (Suppl 1). 2008.

[9] Y.A. Lussier, L. Shagina and C. Friedman C. Automating SNOMED coding using medical language understanding: a feasibility study. Proc AMIA Symp. 2001; 418–422. 2001.

[10] C. Friedman, L. Shagina, Y. Lussier and G. Hripcsak. Automated encoding of clinical documents based on natural language processing. J Am Med Inform Assoc 2004, 11(5):392-402. 2004.

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

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

[13] L. Hirschman, A.A. Morgan and A.S. Yeh. Rutabaga By Any Other Name: Extracting Biological Names. Journal of Biomed. Informatics. Vol. 35. Pp. 247-259. Elsevier. 2003.

[14] C. Jacquemin. Spotting and Discovering Terms through Natural Language Processing. MIT Press. 2001.

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

[16] I. Korkontzelos, Klapaftis I. and S. Manandhar. Reviewing & Evaluating Automatic Term Recognition Techniques. 6th International Conference on Natural Language Processing, GoTAL 2008, Gothenburg, Sweden. 248-259. 2008

[17] W. Wong, and M. Bennamoun. Determining the Unithood of Word Sequences using MI and Independence Measure. 10th PACLING. Australia. 2007.

[18] L. Gillam, M. Tariq and K. Ahmad. Terminology and the construction of ontology. Terminology, 11:55–81. 2005.

[19] S. Banerjee and T. Pedersen. The Design, Implementation and Use of the Ngram Statistics Package. Fourth International Conference on Intelligent Text Processing and Computational Linguistics. Mexico, 2003.

[20] M.T. Pazienza, M. Pennacchiotti and F.M. Zanzotto. Terminology extraction: an analysis of linguistic and statistical approaches. In Knowledge Mining, Studies in Fuzziness & Soft Computing, Vol.185, Springer. 2005.

[21] T. Brants. TnT - A Statistical Part-of-Speech Tagger. Sixth Applied Natural Language Processing Conference – ANLP. Seattle, WA. 2000.

[22] G. Nenadié, S. Ananiadou and J. Mcnaught. Enhancing automatic term recognition through recognition of variation. 20th Conf. on Computational Linguistics. Switzerland. 2004.

[23] F. Sclano and P. Velardi. TermExtractor: a Web Application to Learn the Common Terminology of Interest Groups and Research Communities. Ninth Conf. on Terminology & AI. Sophia Antinopolis. 2007.

[24] K. Verspoor, D. Dvorkin, K. Bretonnel Cohen and L. Hunter. Ontology quality assurance through analysis of term transformations. Bioinformatics 25(12):i77-i84; doi:10.1093/bioinformatics/btp195. 2009.