ELSEVIER

International Journal of Medical Informatics 67 (2002) 85–95

International Journal of Medical Informatics

www.elsevier.com/locate/ijmedinf

Assessing the consistency of a biomedical terminology through lexical knowledge

Olivier Bodenreider [a,*], Anita Burgun [b], Thomas C. Rindflesch [c]

[a] US National Library of Medicine, 8600 Rockville Pike, MS 43, Bethesda, MD 20894, USA
[b] Laboratoire d'Informatique Médicale, University of Rennes, Rennes, France
[c] US National Library of Medicine, Bethesda, MD, USA

Abstract

Objective: We investigate the use of adjectival modification as a way of assessing the systematic use of linguistic phenomena to represent similar lexical or semantic features in the constituent terms of a vocabulary. *Methods:* Terms consisting of one or more adjectival modifiers followed by a head noun are selected from disease and procedure terms in SNOMED. Frequently co-occurring adjectival modifiers are systematically combined with the contexts (i.e., terms minus modifier) of each modifier. The existence of these combinations is checked in both SNOMED and the entire UMLS Metathesaurus; the term corresponding to the context alone is similarly checked. Relationships among terms sharing a context and between each of these terms and their context are studied. *Results:* Four pairs of modifiers were studied: (*acute*, *chronic*), (*unilateral*, *bilateral*), (*primary*, *secondary*), and (*acquired*, *congenital*). The numbers of contexts studied for each pair ranged from 73 to 974. The percentage of contexts associated with both modifiers ranged from 5 to 50% in SNOMED and from 10 to 60% in UMLS. The presence of the context term varied from 31 to 64% in SNOMED and from 43 to 79% in UMLS. Finally, 172 occurrences (9%) of synonymy between a modified term and the context term were found in SNOMED. One hundred and forty-five such occurrences (8%) were found in the entire Metathesaurus.
Published by Elsevier Science Ireland Ltd.

Keywords: Unified Medical Language System; SNOMED; Lexical knowledge; Terminology

1. Introduction

Several dozen biomedical terminologies, generally consisting of pre-coordinated terms, are available and contribute to capturing not only the vocabulary of biomedicine, but also, in part, biomedical knowledge. Large biomedical terminologies are usually the result of a team effort sustained over a long period of time. Therefore, it is not surprising that, besides limited coverage [1], one issue often identified in biomedical terminological resources is their lack of consistency [2]. However, the study of inconsistency is often limited to structural aspects such as the

* Corresponding author
E-mail addresses: olivier@nlm.nih.gov (O. Bodenreider), anita.burgun@univ-rennes1.fr (A. Burgun), tcr@nlm.nih.gov (T.C. Rindflesch).

1386-5056/02/$ - see front matter. Published by Elsevier Science Ireland Ltd.
PII: S1386-5056(02)00051-5

presence of circular hierarchical relationships [3,4], or to the features of existing terms such as their semantic categorization [5,6]. In this study, we define consistency as the consistent use of linguistic phenomena to represent similar lexical or semantic features in the constituent terms of a vocabulary.

Terms for the disease adrenocortical insufficiency can be used as an illustration. The origin of adrenocortical insufficiency can be either primary (when a lesion of the adrenal cortex is responsible for the reduction of the secretion of adrenal hormones) or secondary (when the regulation of the secretion, not the adrenal cortex, is deficient). Therefore, to reflect the possible causes of the disease, the two terms *primary adrenocortical insufficiency* and *secondary adrenocortical insufficiency* are expected to be present in a clinical terminology.

From a terminological perspective, *primary* and *secondary* are modifiers of the broader term *adrenocortical insufficiency*. The absence of one of the two narrower terms would thus represent an inconsistency in the terminology. From a linguistic perspective, *primary* and *secondary* are two adjectives modifying the noun phrase *adrenocortical insufficiency*, and the terms *primary adrenocortical insufficiency* and *secondary adrenocortical insufficiency* are two hyponyms of *adrenocortical insufficiency*. More generally, we hypothesize that two terms of the form $modifier_1$-*context* and $modifier_2$-*context* are co-hyponyms of the term *context*. Therefore, in a consistent terminology, the terms $modifier_1$-*context* and $modifier_2$-*context* should be (1) both present and (2) in hierarchical relation with the term *context*.

In a previous study, we applied lexical knowledge to suggest hyponymic relationships among medical terms [7]. More precisely, we used the property that adjectival modifiers usually introduce a hyponymic relationship to suggest possible hyponymic relationships between modified and unmodified terms (e.g., *secondary cardiomyopathy* and *cardiomyopathy*). We found that less than half of the hyponymic relationships suggested by this method were actually recorded as hierarchical relationships in the Unified Medical Language System® (UMLS®). This method was used to suggest some 20,000 possibly missing relationships to be reviewed by UMLS editors. We also argued that patterns based in particular on additional knowledge about the modifiers might help assess certain hyponymic relationships automatically. For example, if *chronic ischemic enteritis* is a hyponym of *ischemic enteritis*, knowing that *acute* is an antonym of *chronic* allows the inference that *acute ischemic enteritis* is also a hyponym of *ischemic enteritis*.

Following-up on this study, we decided to apply lexical knowledge to the analysis of biomedical terminologies, with the aim of assessing the consistency of a terminology. In other words, our hypothesis is that lexical knowledge may help discover inconsistencies in a vocabulary, either lexical (inconsistent use of linguistic phenomena in terms) or structural (inconsistent organization of the terms). The goal of this study is not to automatically assess consistency. Rather, we propose an unsupervised method to detect potential inconsistencies, which can support and focus the effort of human editors of a medical vocabulary.

2. Background

The main contribution of this paper is not to propose a novel technique, but rather to apply existing techniques to a novel objective, namely assessing the consistency of a terminology. The techniques used are based on the

study of term variation and have previously been used, for example, for creating semantic classes. The general framework of this study is that of word affinities derived from a corpus described by Grefenstette [8]. We use first-order techniques (“what other words are likely to be found in the immediate vicinity of a given word”) and second-order techniques (“which words share the same environments”). Because our objective is to assess consistency rather than create semantic classes, we do not, however, use third-order techniques, in which semantic clusters are derived from lists of similar words produced by second-order techniques.

The study of term variation is central to that of terminology, and methods have been proposed for identifying and representing term variants [9]. Daille et al. [10] report that, on a medical corpus, insertion and juxtaposition (including the juxtaposition of an adjectival modifier to the left of a noun phrase) account for roughly half of the variation. Terminologies consisting of pre-coordinated terms are somewhat similar to sense enumerative lexicons, and the two share similar limitations. Studying term variation helps to reveal the compositional nature of the terms and can therefore be understood as a generative approach to terminology [11].

The grouping of semantic variants is called semantic normalization and serves as the basis for creating semantic classes. Habert and Fabre [12] use dependency trees to analyze term variants from a corpus and acquire semantic classes. Nazarenko et al. [13] successfully applied this technique to SNOMED. Although used in a different objective and limited in its scope, our method shares many of the techniques and corpora used by these authors.

Finally, other approaches to analyzing terminologies include description logics [14]. These techniques may help to detect and fix semantic inconsistencies by automatically classifying the concepts (e.g., by comparing the expected classification to that proposed by the system). However, a significant amount of manual work is usually required for entering the terms into a description logics-based system. Moreover, lexical phenomena that do not influence the semantics of a term may still fail to be caught by such systems.

3. Material and methods

The method may be summarized as follows. Starting with a list of terms, a syntactic analysis of the terms supports the identification of adjectival modifiers. The analysis is restricted to simple terms constituted of one or more modifiers followed by a head noun. After a modifier is extracted from the term, the remaining head noun—along with the other modifiers, if any—forms the context of this term. Modifiers sharing the same context are clustered together and ranked by frequency. Pairs of frequently co-occurring modifiers (i.e., occurring in the same context) are established. For a given pair of modifiers (m_1, m_2), terms are created by associating each modifier with the context c in which either one was detected (m_1c, m_2c). The existence of m_1c and m_2c is checked in both the vocabulary studied and the entire UMLS Metathesaurus, as well as the existence of the term corresponding to the context alone. Relationships between m_1c and m_2c and between each of them and their context c are studied.

3.1. Material

The UMLS Metathesaurus[1] (12th edition, 2001) contains about 1.5 million unique

[1] umlsks.nlm.nih.gov.

English terms drawn from more than 50 medical vocabularies, and organized in some 800,000 concepts. A concept is defined as the set of synonymous terms corresponding to a single meaning. Conversely, terms are names for concepts [15]. In order to address the large size of the Metathesaurus, we limited our study to terms from SNOMED International[2] (version 3.5, 1998), one of the source vocabularies in the UMLS. We further selected from SNOMED terms from two major components of clinical medicine: diseases and procedures. We also removed from this set section headers, which often contain metadata. The notation "NOS", meaning "not otherwise specified", was removed from the terms. Finally, we excluded all terms containing a comma (10% of our original set). Commas usually signal a permuted form (e.g., glucose measurement, urine) or, more generally, a complex term (e.g., patient transfer, in-hospital, unit-to-unit) whose structure is usually not suitable for natural language processing tools. Our final list contains 65,124 terms (39,997 disease terms and 25,127 procedure terms), corresponding to 41,842 concepts in SNOMED and 43,627 concepts in the Metathesaurus.

3.2. Identifying adjectival modifiers

The study of adjectival modification in the SNOMED terms under consideration was based on an underspecified syntactic analysis [16] that draws on a stochastic tagger [17] as well as the SPECIALIST Lexicon, a large syntactic lexicon of both general and medical English that is distributed with UMLS. Although not perfect, this combination of resources effectively addresses the phenomenon of part-of-speech ambiguity in English, and, for example, correctly identifies *open* as an adjective (rather than a verb) in the term *open wound*. The resulting syntactic structure identifies the head and modifiers for the noun phrase analyzed. Each modifier is also labeled as being either adjectival, adverbial, or nominal. Although all types of modification in the simple English noun phrase were labeled, only adjectives were selected for further analysis in this study. For example, the term *male erectile disorder* was analyzed as:

[[modifier(*male*, adj)],
[modifier(*erectile*, adj)],
[head(*disorder*, noun)]].

This syntactic analysis was used to restrict the original set to terms consisting of at least one adjectival modifier followed by possibly other modifiers and a head noun. This specification excludes both simple terms (e.g., one isolated noun) and complex terms, not suitable for our analysis. 14,958 terms were considered for further analysis.

3.3. Establishing a list of adjectival modifiers and their contexts

For each adjectival modifier found in a term, we created a context made from the remainder of the term once the modifier was removed. Context words were lower cased and sorted by alphabetical order to help identify similar contexts. For example, from the term *primary lacrymal atrophy*, we identified the modifier *primary* associated with the context *atrophy lacrymal*, and the modifier *lacrymal* associated with the context *atrophy primary*. 20,176 (*modifier*, *context*) structures were created, corresponding to 3721 unique adjectival modifiers and 11,991 unique contexts.

[2] www.snomed.org.

3.4. Computing the co-occurrence of modifiers

The (*modifier*, *context*) structures were analyzed in order to identify pairs of modifiers frequently associated with the same context. From the two (m_1, c) and (m_2, c) structures created from the terms m_1c and m_2c where m_1 and m_2 represent two distinct modifiers and c represents their common context, the pair of co-occurring modifiers (m_1, m_2) is recorded. The frequency of co-occurrence for (m_1, m_2) is equal to the number of times m_1 and m_2 share a common context. For example, the context *atrophy lacrymal* is associated with the modifiers primary and *secondary*. Therefore, the pair of co-occurring modifiers (*primary*, *secondary*) is recorded for this context. The same pair of modifiers is associated with many other contexts, such as *amyloidosis* (in *primary amyloidosis* and *secondary amyloidosis*). The total frequency of co-occurrence for the pair of modifiers (primary, secondary) is 45. In other words, these two modifiers share 45 distinct contexts. 40,883 pairs of co-occurring modifiers (m_1, m_2) were recorded, with frequency of co-occurrence ranging from 1 to 208. Only 495 pairs have a frequency of five or more.

3.5. Transforming terms

The existence of a pair of co-occurring modifiers (m_1, m_2) means that the two modifiers share at least one common concept c. However, m_2 may not be systematically associated with all the contexts associated with m_1. For a given pair of co-occurring modifiers (m_1, m_2), we created possible terms by associating each modifier with all the contexts in which the other modifier from the pair was detected. For example, using the (*primary*, *secondary*) pair, contexts associated with *primary* include *ovarian failure* and *amyloidosis*, and contexts associated with *secondary* include *hyperprolactinemia* and also *amyloidosis*. The following six terms are created:

- *primary ovarian failure*,
- *secondary ovarian failure*,
- *primary amyloidosis*,
- *secondary amyloidosis*,
- *primary hyperprolactinemia*, and
- *secondary hyperprolactinemia*.

3.6. Looking-up transformed terms in SNOMED and the UMLS

The terms created in this way were mapped to UMLS (and therefore also to SNOMED) by first attempting an exact match between the input term and Metathesaurus concepts. If an exact match failed, normalization was then attempted. This process makes the input and target terms potentially compatible by eliminating such inessential differences as inflection, case and hyphen variation, as well as word-order variation. Moreover, the mapping is considered successful only if the concept mapped to is semantically compatible with the original term. Knowing that original terms are diseases (or procedures), mapping to concepts whose semantic type does not correspond to a disease (or a procedure) results in a failure.

3.7. Analyzing the relationships among terms associated with a pair of modifiers

Two terms m_1c and m_2c sharing the same context c and differing only by one adjectival modifier (m_1 or m_2) are expected to be represented as siblings and to be in direct hierarchical relationship with the context c. Such a representation is expected to be found in both the original vocabulary studied and UMLS. The hierarchical features of SNOMED codes were used to calculate the relationship between two SNOMED terms. For example, the terms *bilateral vasotomy*

(P1-7A124) and *unilateral vasotomy* (P1-7A122) were considered siblings because their codes share all digits but the last one. They can also be seen as descendants of *vasotomy* (P1-7A120), whose code ending with 0 denotes a higher level in the SNOMED hierarchy. In the UMLS Metathesaurus, two concepts were considered in direct hierarchical relationship if related by means of parent/child (PAR/CHD) and broader/narrower (RB/RN) relationships, and siblings if they shared at least one common first-generation ancestor.

4. Results

The list of the most frequent contexts and adjectival modifiers for disease and procedure terms is presented in Table 1 (contexts) and Table 2 (modifiers). The list of the most frequent pairs of co-occurring modifiers is presented in Table 3. Not surprisingly, the relation between adjectives in pairs of frequently co-occurring modifiers is often antonymy (e.g., *acute*, *chronic*). Other classes of frequently co-occurring modifiers include anatomical location (e.g., *cervical*, *thoracic*, *lumbar*, *sacral*), age of onset (e.g., *congenital*, *infantile*, *juvenile*, *adult*), and etiology (e.g., *bacterial*, *fungal*, *viral*).

From the most frequent pairs of co-occurring modifiers, we selected four pairs for further analysis: (*acute*, *chronic*), (*unilateral*, *bilateral*), (*primary*, *secondary*), and (*acquired*, *congenital*). For each pair (m_1, m_2), the number of contexts associated with at

Table 1
List of the most frequent contexts (term minus modifier) in SNOMED terms (diseases and procedures)

Frequency	Context	Example
103	Disease	Autoimmune disease
74	Syndrome	Paraneoplastic syndrome
73	Disorder	Bipolar disorder
72	Fistula	Lacrimal fistula
58	Hemorrhage	Subarachnoid hemorrhage
58	Haemorrhage	Subarachnoid haemorrhage
52	Abscess	Subphrenic abscess
49	Cyst	Nabothian cyst
47	Hernia	Diaphragmatic hernia
42	Ulcer	Duodenal ulcer
38	Dermatitis	Allergic dermatitis
35	Procedure	Cardiovascular procedure
35	Pneumonia	Pneumococcal pneumonia
31	Arthritis	Rheumatoid arthritis
29	Dysplasia	Diastrophic dysplasia
29	Colitis	Chronic colitis
28	Pelvis	Funnel-shaped pelvis
28	Infection	Acute infection
28	Infectious disease	Bacterial infectious disease
27	Anemia	Pernicious anemia
27	Anaemia	Pernicious anaemia
26	Oedema	Cerebral oedema
26	Fever	Recurrent fever
26	Edema	Cerebral edema
26	Cataract	Diabetic cataract

Table 2
List of the most frequent modifiers in SNOMED terms (diseases and procedures)

Frequency	Modifier	Example
874	Congenital	Congenital hydrocephalus
412	Acute	Acute pancreatitis
408	Chronic	Chronic interstitial cystitis
177	Pulmonary	Pulmonary inhalation study
171	Familial	Familial acholuric jaundice
167	Partial	Partial esophagectomy
152	Acquired	Acquired aplastic anemia
129	Renal	Renal transplant
125	Neonatal	Neonatal hyperbilirubinemia
123	Primary	Primary adrenal deficiency
121	Hereditary	Hereditary ataxia
113	Retinal	Diabetic retinal microaneurysm
109	Secondary	Secondary dilated cardiomyopathy
100	Idiopathic	Idiopathic cardiomyopathy
96	Cerebral	Cerebral hemorrhage
92	Infectious	Bacterial infectious disease
86	Small	Bilateral small kidney
75	Total	Total pancreaticoduodenectomy
74	Misshapen	Congenital misshapen clavicle
74	Hemorrhagic	Hemorrhagic gastritis
74	Haemorrhagic	Haemorrhagic gastritis
73	Viral	Viral conjunctivitis
73	Cervical	Cervical sympathectomy
72	Rheumatic	Rheumatic endocarditis
72	Gastric	Gastric lavage

Table 3
List of the most frequent pairs of co-occurring modifiers in SNOMED terms (diseases and procedures)

Frequency	Modifiers	Example
208	Acute/chronic	* Zinc deficiency
69	Haemorrhagic/hemorrhagic	* Gastritis
55	Misshapen/small	Congenital * adrenal gland
54	Fetal/foetal	* Biophysical profile
52	Acquired/congenital	* Sideroblastic anaemia
46	Fused/misshapen	Congenital * carpal bone
45	Primary/secondary	* Hyperparathyroidism
41	Acute/subacute	* Angle-closure glaucoma
39	Partial/total	* Gastrectomy
37	Bilateral/unilateral	* Nephrectomy
36	Malpositioned/misshapen	Congenital * femur
36	Chronic/subacute	* Angle-closure glaucoma
34	Fused/small	Congenital * frontal bone
30	Malpositioned/small	Congenital * liver
29	Cervical/thoracic	* Discography
29	Acquired/hereditary	* Factor VIII deficiency disease
27	Ischaemic/ischemic	* Heart disease
27	Gastrointestinal/gi	* Hemorrhage
27	Complete/partial	* Substernal thyroidectomy
25	Esophageal/oesophageal	* Varices
23	Cervical/lumbar	* Discography
21	Fused/malpositioned	Congenital * carpal bone
20	Lumbar/thoracic	* Discography
19	Anterior/posterior	* Uveitis
18	Sacral/thoracic	Supernumerary * vertebra

least one of the modifiers, the presence in the terminology of the modified terms (m_1c, m_2c) and of the context (c), and the nature of the relationship between the two modified terms (m_1c/m_2c) and between the modified terms and the context (m_1c/c, m_2c/c) are summarized in Table 4.

The pair (*acquired*, *congenital*) will be used to illustrate the results. Nine hundred and seventy-four contexts are associated with either modifier of the pair. Both modified terms are present in SNOMED in 52 cases, and in UMLS in 97 cases (e.g., *acquired spondylolisthesis*, *congenital spondylolisthesis*). Terms modified by *congenital* only (e.g., *congenital bronchiectasis*) are more frequent (822 in SNOMED) than those modified by *acquired* only (e.g., *acquired epidermolysis bullosa*, 100 in SNOMED). Their contexts (e.g., *epidermolysis bullosa*) are present in SNOMED in 306 cases and in UMLS in 418 cases. The terms modified by *acquired* and *congenital* are not frequently represented as siblings in SNOMED (10 cases). For example, *acquired keratoderma* (D0-22310) and *congenital keratoderma* (D4-40130) are represented in two separate branches of the disease hierarchy in SNOMED. Moreover, the relationships between modified terms and their context also contribute to the characterization of a pair of modifiers. Most terms modified by *acquired* and *congenital* do not have any paradigmatic relationship represented with their context. For example, although *keratoderma* exists as a concept in the Metathesaurus, there is no relationship between *acquired keratoderma* or *congenital keratoderma* and *keratoderma*. In 44 cases, the relationship is hierarchical (e.g., between *congenital porphyria* and *porphyria*). In 18 cases, the modified term and its context are siblings (e.g., congenital Addison's disease and Addison's disease). Finally, in 99 cases, they are considered synonyms in SNOMED (e.g., *acquired polycythemia* and *polycythemia*).

5. Discussion

Although a formal evaluation would be required, preliminary results suggest that the method is effective at automatically identifying potential inconsistencies in terminological resources. Examples of such inconsistencies, both structural and lexical, are analyzed in this section. Generalization issues are also addressed.

Table 4
Characteristics of four pairs of co-occurring modifiers (m_1, m_2)

		m_1: acquired, m_2: congenital ($N = 974$)				m_1: acute, m_2: chronic ($N = 608$)				m_1: primary, m_2: secondary ($N = 187$)				m_1: unilateral, m_2: bilateral ($N = 73$)			
		SNOMED		UMLS		SNOMED		UMLS		SNOMED		UMLS		SNOMED		UMLS	
Present	Both	52	5%	97	10%	208	34%	244	40%	45	24%	69	37%	37	51%	44	60%
	m_1c only	100	10%	76	8%	203	33%	190	31%	78	42%	67	36%	22	30%	18	25%
	m_2c only	822	84%	801	82%	197	32%	174	29%	64	34%	51	27%	14	19%	11	15%
	Context	306	31%	418	43%	324	53%	399	66%	119	64%	147	79%	41	56%	50	68%
m_1c and m_2c siblings		10	1%	51	5%	142	23%	225	37%	29	16%	64	34%	41	56%	41	56%
Relationship of m_1c or m_2c to c	Child	44	4%	181	17%	300	37%	294	35%	90	39%	101	39%	42	38%	41	35%
	Siblings	18	2%	93	9%	78	10%	239	28%	16	7%	65	25%	22	20%	32	27%
	Synonyms	99	10%	82	8%	38	5%	26	3%	29	13%	24	9%	2	2%	1	1%
	Child	865	84%	715	67%	400	49%	293	34%	97	42%	66	26%	44	40%	43	37%

5.1. Ontological perspective

Classically, in the Ogden–Richards triangle, there is a distinction between the symbol (here, the term), the concept named by the term, and the referent (“thing in the world”) referred to by the concept and for which the term stands [18]. In this study, the terms m_1c and m_2c modified by a pair of modifiers (m_1, m_2) and their context (c) are terms used to name concepts. In the simplest case, the three distinct terms m_1c, m_2c and c stand for three referents, referred to by three concepts. Indeed, we found many occurrences of this representation. In this case, the context represents generic knowledge, while the modified terms bear some kind of specification. The context c is in hierarchical relationship with both m_1c and m_2c, and therefore, m_1c and m_2c are siblings. In many cases, however, more than one symbol is available to name a concept (synonymy). Sometimes, the same symbol is used to name several concepts (polysemy). While synonymy and polysemy are well-known linguistic phenomena, other associations among terms, concepts and referents may be found as well. Namely, the following situations may occur and will be discussed: missing referent, missing concept and missing symbol.

5.1.1. Missing referent

In this experiment, we artificially created terms by associating modifiers with contexts, knowing that some of these associations may not actually stand for an existing referent. For example, a *congenital cleft hand* results from a developmental anomaly and no other circumstance later on in life can cause the same condition. In other words, there is no such referent as an *acquired cleft hand*. Therefore, the term acquired cleft hand and the concept it could name are purposely and correctly missing from medical terminologies. Moreover,

because the only possible circumstance for a cleft hand to occur is congenital, there is no need for a generic term *cleft hand*. The existence of only one specialized concept suppresses the need for a generic concept. *Acute copper deficiency* provides another example of a referent that does not exist. In this case, however, the generic term *copper deficiency* does exist, but symbolizes the same meaning as *chronic copper deficiency*.

Domain knowledge is needed to distinguish between a referent that does not exist and failure to represent an existing referent. The use of the UMLS partially supplies this knowledge. Modified terms and contexts are consistently more likely to be found in UMLS than in SNOMED, which is not surprising since, by design, UMLS is both broader in scope and more granular. Terms present in UMLS but not in SNOMED may indicate that the referent does actually exist while SNOMED lacks a term to name it. For example, the generic term *hearing disorder* is present in UMLS, but not in SNOMED, although *congenital hearing disorder* is present in SNOMED.

5.1.2. Missing concept

The absence of a concept in UMLS may also result from an incomplete representation of the world, and, once again, domain knowledge is needed to find out. For example, although *congenital pneumonia* and *pneumonia* are represented in both systems, there is no *acquired pneumonia* in SNOMED or in UMLS. In this case, *acquired pneumonia* is the most common form of the disease, and so common that the generic term *pneumonia* is used to represent the prototypical term. Many cases of this phenomenon can be found.

Incomplete knowledge representation may result in inaccurate reasoning. For example, the prototypical form of meningocele, *congenital meningocele*, is clustered together (in the same concept) with the generic term *meningocele*. As a consequence, *acquired meningocele* is correctly represented as a child of the generic term, but wrongly represented as a child of *congenital meningocele*, allowing properties of congenital meningoceles to be falsely inherited by acquired meningoceles. The term *congenital meningocele* symbolizes a concept distinct from that named by the generic term and should therefore be represented as a distinct concept.

5.1.3. Missing symbol

In some cases, the absence of a term in SNOMED simply results from a lack of synonymy being represented. For example, the term *primary polycythemia* does not exist in SNOMED, but the concept it symbolizes does, simply named by a different term (*polycythemia vera*). The synonymy between the two terms is recognized in UMLS.

5.2. Lexical inconsistency

Besides discrepancies in knowledge representation that could be detected by description logics-based analyses as well, this study also revealed lexical inconsistencies. For example, the two SNOMED terms *primary open-angle glaucoma* and *secondary open angle glaucoma* are hyphenated differently. Also, some but not all terms modified by *bilateral* exhibit a plural mark while the term modified by *unilateral* is often, but not always, singular. The systematic creation of synonyms based on spelling variants (e.g., *anemia*/*anaemia*) could have been tested as well. As shown in Table 1, for contexts differing only by spelling (*hemorrhage*/*haemorrhage*, *anemia*/*anaemia*, and

edema/*oedema*), the two spelling variants have exactly the same frequency. This suggests that terms were created consistently for each spelling variant.

5.3. Generalization

The method presented was voluntarily restricted to the domain of disorders and procedures, to adjectival modification, and to SNOMED, and used only pairs of adjectives sharing a given context. We believe, however, that these restrictions are simplifications made in the context of this study rather than limitations of the method.

Generalizing to other domains poses no problems as long as the relevant terminology is amenable to natural language-processing techniques and modification phenomena. This would include domains such as anatomy or physiology. However, in other domains such as molecular biology, with many genes and gene product names, and chemistry, with many chemical names, consistency would probably be more difficult to assess with this method.

Nominal modification is common in English and in principle can be addressed with a methodology similar to the one discussed here. Nominal modifiers often express a quality more closely related semantically to the head than do adjectives. Details in the methodology would be adjusted to accommodate this characteristic.

This method could be applied to virtually any terminology consisting of pre-coordinated terms, but not necessarily to compositional concept representation systems (e.g., GALEN) in which terms come into existence only as the byproduct of terminology services.

Finally, clusters of adjectives sharing a context need not be limited to pairs. Using the frequency of co-occurrence of the modifiers as the criterion to regroup them (as we did in this study), clusters of three modifiers or more could be identified instead of pairs of modifiers. Each modifier in the cluster would then be combined with the context as usual. Pairs of frequently co-occurring modifiers often correspond to binary descriptive adjectives (e.g., *congenital*, *acquired*). The interest of using larger clusters of frequently co-occurring modifiers is to extend this method to other kinds of adjectives, i.e., non-binary descriptive adjectives and relational adjectives [19]. Non-binary descriptive adjectives are often used in medical terms to express gradation (e.g., *congenital*, *infantile*, *juvenile*, *adult*). Relational adjectives may denote, for example, anatomical location (e.g., *cervical*, *thoracic*, *lumbar*, *sacral*) and etiology (e.g., *bacterial*, *fungal*, *viral*). However, grouping modifiers based on their sole frequency of co-occurrence may result in heterogeneous clusters mixing relational and descriptive adjectives (e.g., *bacterial*, *fungal*, *acute*) or mixing several kinds of relational adjectives (e.g., *bacterial*, *fungal*, *renal*, *hepatic*). While useful for maximizing the identification of potential inconsistencies, this extended method is likely to generate many false positives. Therefore, it should be followed by manual review or combined with other techniques.

6. Conclusion

In this study, we used lexical knowledge to assess the consistency of a biomedical terminology, not only from the perspective of knowledge representation, but also for checking the consistent use of linguistic phenomena in terms. This method alone is certainly not sufficient for ensuring consistency, and we reaffirmed the need for domain knowledge. However, we believe that it can be useful to limit and focus the effort of the human editors of biomedical terminological systems.

References

[1] C.G. Chute, S.P. Cohn, K.E. Campbell, D.E. Oliver, J.R. Campbell, The content coverage of clinical classifications for the computer-based Patient Record Institute's Work Group on codes and structures, J. Am. Med. Inform. Assoc. 3 (3) (1996) 224–233.

[2] J.J. Cimino, Auditing the Unified Medical Language System with semantic methods, J. Am. Med. Inform. Assoc. 5 (1) (1998) 41–51.

[3] O. Bodenreider, Circular hierarchical relationships in the UMLS: etiology, diagnosis, treatment, complications and prevention, in: Proceedings of the AMIA Symposium, 2001, pp. 57–61.

[4] S. Schulz, U. Hahn, Medical knowledge reengineering-converting major portions of the UMLS into a terminological knowledge base, Int. J. Med. Inf. 64 (2–3) (2001) 207–221.

[5] H. Gu, Y. Perl, J. Geller, M. Halper, L.M. Liu, J.J. Cimino, Representing the UMLS as an object-oriented database: modeling issues and advantages, J. Am. Med. Inform. Assoc. 7 (1) (2000) 66–80.

[6] A. McCray, O. Bodenreider, A conceptual framework for the biomedical domain, in: S. Myaeng, R. Green (Eds.), Semantics of Relationships, an interdisciplinary perspective. Boston: Kluwer Academic Publishers, Dordrecht, 2002, pp. 181–198.

[7] O. Bodenreider, A. Burgun, T.C. Rindflesch, Lexically suggested hyponymic relations among medical terms and their representation in the UMLS, in: Proceedings of the Terminology and Artificial Intelligence (TIA'2001), 2001, pp. 11–21.

[8] G. Grefenstette, Corpus-derived first-, second- and third-order word affinities, in: EURALEX 1994, Amsterdam, 1994.

[9] C. Jacquemin, Syntagmatic and paradigmatic representations of term variation, in: Proceedings of the ACL 1999, 1999, pp. 341–348.

[10] B. Daille, B. Habert, C. Jacquemin, J. Royauté, Empirical observation of term variations and principles for their description, Terminology 3 (2) (1996) 197–258.

[11] J. Pustejovsky, The Generative Lexicon, MIT Press, Cambridge, MA, 1995.

[12] B. Habert, C. Fabre, Elementary dependency trees for identifying corpus-specific semantic classes, Comput. Humanities 33 (3) (1999) 207–219.

[13] A. Nazarenko, P. Zweigenbaum, J. Bouaud, B. Habert, Corpus-based identification and refinement of semantic classes, in: Proceedings of the AMIA Annual Fall Symposium, 1997, pp. 585–589.

[14] A.L. Rector, S. Bechhofer, C.A. Goble, I. Horrocks, W.A. Nowlan, W.D. Solomon, The GRAIL concept modelling language for medical terminology, Artif. Intell. Med. 9 (2) (1997) 139–171.

[15] A.T. McCray, S.J. Nelson, The representation of meaning in the UMLS, Meth. Inf. Med. 34 (1-2) (1995) 193–201.

[16] T.C. Rindflesch, J.V. Rajan, L. Hunter, Extracting molecular binding relationships from biomedical text, in: Proceedings of the Sixth Applied Natural Language Processing Conference, Morgan Kaufmann Publishers, San Francisco, CA, 2000, pp. 188–195.

[17] D.R. Cutting, J. Kupiec, J.O. Pedersen, P. Sibun, A practical part-of-speech tagger, in: Proceedings of the Third Conference on Applied Natural Language Processing, 1992, pp. 133–140.

[18] C.K. Ogden, I.A. Richards, The Meaning of Meaning: A Study of the Influence of Language Upon Thought and of the Science of Symbolism, 8th ed., Harcourt Brace, New York, 1946.

[19] K.J. Miller, Modifiers in WordNet, in: C. Fellbaum (Ed.), WordNet: An Electronic Lexical Database, MIT Press, Cambridge, MA, 1998, pp. 47–67.