Estimating and Analysing Coordination in Medical Terminologies

Cornelia Hedeler, Bijan Parsia, and Sebastian Brandt

School of Computer Science, The University of Manchester, Manchester, UK

Abstract. Medical vocabulary is complex, expanding, and convoluted not least because of the large numbers of compound terms. Formalized medical terminologies such as SNOMED-CT and ICD-10 take one of two strategies when representing medical language: so-called pre-coordination where valid compound terms are included explicitly in the terminologies and so-called post-coordination where the terminology consists of a basis and a generative function from which the compound terms may be derived. However, these notions are not used with particular precision in the literature. In this paper, we provide a formalization of the notion of coordination, a technique for estimating the degree of coordination in a given system, and an examination, based on our technique, of the coordination level of a number of major existing terminologies.

Keywords: pre-coordination, post-coordination, medical vocabularies

1 Introduction

Controlled medical terminologies have long played an important role in the drive to improve patient care, e.g., in Electronic Health Records (EHR), decision support and expert systems, medical literature databases, and data exchange [19, 17, 27]. A large number of medical terminologies have been developed, including the Systemized Nomenclature of Medical Clinical Terms (SNOMED CT) [21, 6], the International Classification of Diseases and Related Health Problems, 9th Revision - ICD-9 and more recently its 10th Revision ICD-10 both with Clinical Modifications (CM) and Procedure Coding System (PCS) [10, 39, 11], and the National Cancer Institute (NCI) thesaurus [8, 16].

Term composition [24–26] or as it is more recently (in particular in the context of SNOMED CT) called, (pre-/post-) coordination [23, 35] in medical terminologies has been a concern for developers of standard terminologies [41] and end-users [22] alike. For example, end users can struggle with post-coordination [22, 30, 12, 33], but pre-coordination brings its challenges during development and maintenance and can lead to an exponential explosion of the terminology size [26, 32], which in turn can make it hard for end-users to determine the appropriate term to use. In addition, different levels of coordination can make data exchange between systems and mapping of different terminologies harder, requiring additional approaches to handle the differences in coordination [42].

These discussions have led to repeated suggestions of terminologies to consist of a base of atomic concepts, with a function, e.g., in form of a grammar or a description logic ontology, specifying how to compose the atomic terms to form more complex terms [24, 26, 32, 38, 27]. Despite the longevity of the discussion on composition of terminologies, there is no consensus on when a style is to be preferred or even exactly what constitutes a pre- or post- coordinated vocabulary. There are current exemplars of each style, e.g., the heavily pre-coordinated ICD-10 [11] and, SNOMED [40, 32], the poster child for post-coordination.

Examples of pre-coordinated terms include 'Removal of External Fixation Device from Left Upper Femur, Open Approach', a procedure in ICD-10-PCS, "Removal of implanted devices from bone, femur", a procedure from ICD-9 procedures, which contains less detail (no detail on the location of the femur or the kind of approach used for the removal of the implanted device) than the procedure in ICD-10, i.e., could be considered to be less coordinated. Similarly, the procedure 'Removal of internal fixation device of femur' in SNOMED CT, even though less coordinated than the procedure in ICD-10, would still be considered to be coordinated as there is potential for further decomposition, e.g., into the procedure itself, i.e., 'removal', the object that is being removed, i.e., 'internal fixation device', and the location from where it is being removed, i.e., 'femur'. SNOMED CT also contains atomic terms to be used for post-coordination, for example, 'laterality' along with 'left' and 'right', but also pre-coordinated terms such as 'X-ray of left foot'.

However, so far no computational method to determine the level of coordination of terminologies has been available. In this paper, we propose a method for estimating the level of coordination of a given terminology and determine the coordination level of a number of standard terminologies.

2 Background

Consider a simple clinical terminology, T_1, that contains a single term, `Back_Pain`, that is:

$$T_1 = \{\texttt{Back\_Pain}\} \tag{1}$$

While `Back_Pain` is *atomic*, that is, a single term with no substructure, it is clear from our perspective that it is a compound term. That is, it is composed of two sub-terms which themselves are clinically relevant, to wit `Back` and `Pain` which gives us an alternative terminology:

$$T_2 = \{\texttt{Back}, \texttt{Pain}\} \tag{2}$$

We, of course, can generate T_1 from T_2 via a suitable function. The simplest *correct* function would be a map from a subset of T_2 (i.e., $\{\texttt{Back}, \texttt{Pain}\}$) to the corresponding term in T_1 (i.e., `Back_Pain`). If we extend T_1 and T_2 we need to extend our mapping function as well:

$$T_1' = \{\texttt{Back\_Pain, Heart\_Pain, Stomach\_Pain}\} \tag{3}$$

$$T_2' = \{\texttt{Back, Heart, Stomach, Pain}\} \tag{4}$$

$$\begin{aligned} Map = \\ \{\{\texttt{Back,Pain}\} &\rightarrow \texttt{Back\_Pain} \\ \{\texttt{Heart,Pain}\} &\rightarrow \texttt{Heart\_Pain} \\ \{\texttt{Stomach,Pain}\} &\rightarrow \texttt{Stomach\_Pain}\} \end{aligned}$$

We can also encode Map by means of a simple grammar:

$$\begin{aligned} Map2 = \\ \textbf{Term} &::= \textbf{Location}, \_, \texttt{Pain} \\ \textbf{Location} &::= \texttt{Back | Heart | Stomach} \end{aligned}$$

While $Map2$ is rather more illuminating than Map or T_1' and contains more information than T_2' (T_2' is not, by itself, sufficient to generate exactly T_1'), it is conceptually more complicated, i.e., we have to understand a BNF style formalism. While $Map2$ is very simple, it is easy to imagine situations that are inherently more complicated:

$$
\begin{aligned}
Map3 = & \\
\textbf{Term} &::= \textbf{Location}, \_, \texttt{Pain} \\
\textbf{Location} &::= \textbf{LocationWithLaterality} \mid \texttt{Back} \mid \texttt{Heart} \mid \texttt{Stomach} \\
\textbf{LocationWithLaterality} &::= \textbf{Laterality}, \_, \texttt{Leg} \\
\textbf{Laterality} &::= \texttt{Right} \mid \texttt{Left}
\end{aligned}
$$

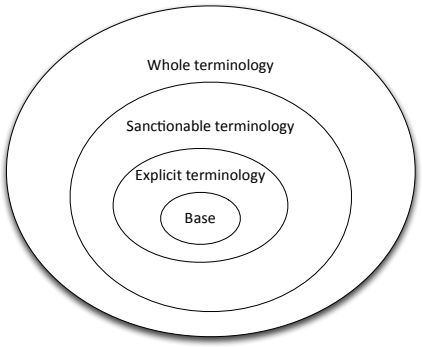

Fig. 1. Diagram of size of base in comparison to explicitly modelled, sanctionable and whole terminology.

The more exceptions and irregularities, the more difficult it is to come up with an exact and yet intelligible description of the mapping, or, for that matter, a terser description. In addition, ideally the map should only allow the construction of complex terms that are sensible for the domain. If that is not possible, additional constraints, also known as sanctions, have to be added to specify which compositions that can be built following the map, are meaningful, and thus are allowed or sanctioned [20, 37, 13].

If we consider the set of compound terms generated from $Map3$, there are only 5. If we include the "intermediate" terms (e.g., `Right_Leg`), there are only 7, and if we include the base as well, there are a mere 12. Thus, the enumeration of the terms is shorter than the grammar itself, in this simple case. However, extending the set of terms with laterality causes the generated terms to grow more quickly than the base set or the grammar. And it is easy to see that terminologies with more dimensions can have even more dramatic gaps between the enumerative and the descriptive presentations of the terminology.

In general, a terminology (even idealised as complete) will not be fully enumerative or fully descriptive. Instead, given a set of base terms that are truly atomic, we can project a structure between the basis and the total terminological space generated from that basis. (Note, that if we allow for arbitrary iteration of base terms in a compound term, then the full space will be infinite.) Figure 1 illustrates a slightly idealised structure of a representation of a vocabulary. The "real" vocabulary, that is the set of terms which are meaningful for a domain, corresponds

to the sanctioned terminology. A terminology might include a degree of pre-coordination in the form of an explicit set of coordinated terms without fully covering the vocabulary. This might be because the terminology is incomplete or that there are some compound terms which are so common or significant that the terminology designer wanted to make them more salient. Finally, there may be terms in the basis or explicit terminology which are not domain significant, at least, in isolation.

Note that this sort of phenomenon is not limited to terminologies per se. For example, we might have a hierarchical terminology where the hierarchical relations are themselves aspects of the coordination. (It is easy to see that we could generate a term plus a list of parent terms via a mapping function.) Similarly, we might be generating forms from a base "vocabulary" of form elements, and so on.

Furthermore, the map could be described in any number of formalisms, for example, by a description logic ontology.

With this picture, it is straightforward to articulate the classic trade offs between pre- and post-coordinated terminologies. A post-coordinated vocabulary has the possibility of 1) being more "compressed" in size but also 2) it can be expressed in a more perspicuous way. Instead of having a huge pile of terms, we can isolate important subgroups of terms and characterise them by their contribution to a mapping function. This can potentially support more effective review for correctness and completeness of the terminology itself and of systems that make use of it, as well as just lowering the cost of maintenance (since we have "fewer things" to look at or connect to other parts of a system).

Contrariwise, while we might have fewer things to look at, each thing might be significantly more complex. The correctness of a coordination function with respect to the sanctioned vocabulary might be quite difficult to determine. Furthermore, the execution of the coordination function to generate or validate a term might be expensive in computational resources, and it might not be obvious "how much" of some section of the vocabulary to manifest for a particular situation.

3 Methods

3.1 Formalization

Intuitively, coordination is a process of inductive composition, that is, a coordination of a (base) terminology is a recursively defined function over that terminology. Since we can define such functions arbitrarily over any given basis, just identifying coordinations with recursively defined sets is not particularly illuminating. In our case, the point of coordination is to capture some *specific* set of terms (i.e., exactly the meaningful ones).

Definition 1. *A* vocabulary, $\mathcal{V}$, *is a set of terms such that each term is meaningful wrt to a domain and the set is exhaustive of such terms.*

"Meaningfulness" is domain and application specific.

Definition 2. *A terminology T is a* coordination *of another terminology T' just in case there is a function, f, which is a recursive definition of T' with the base cases all lying in T and* $T' \subseteq \mathcal{V}$. *We call f the* coordinating *or* coordination function. *We call* T' *the* basis *of the coordination or the* coordinated terminology.

In the most natural instance, a term is a string and a canonical coordinating function constructs a new term by concatenating the old ones. For our current purposes, it suffices to be this narrow.

Clearly, a coordination can, but does not have to, contain its basis. It may also only contain part of its basis. For example, a terminology might not contain left on its own even though left is part of its basis. For the target application, left might not be meaningful on its own. Furthermore, we can arrange coordinations in a partial order and there is always the possibility of exactly capturing $\mathcal{V}$.

Definition 3. *T is more coordinated than T' ($T' \preceq T$) if T is a coordination of T' and T' is not a coordination of T.*

Definition 4. *A coordination, T, is saturated or fully coordinated if $T = \mathcal{V}$.*

A saturated coordination might still be extensible, i.e., there are still desired terms not in the coordination, but any extension requires an extension to the basis. It is highly likely that few terminologies are saturated, especially if we allow for all clinically meaningful, i.e., sanctionable terms.

Definition 5. *The coordination factor (relative to a basis) of T is the size of T / the size of the basis.*

It might be the case that there are several possible bases or that the true basis is unknown. In the first case, a given coordination factor might not be determinate. In the second case, it might not be precise.

It is possible that we have the following situation:

1. There is a basis, B
2. a terminology, T, which is what is physically manifested and delivered,
3. another terminology T_{total}, which is the complete set of terms for the application or domain (i.e., T_{total}, is saturated)

where $B \preceq T \preceq T_{total}$. In this case, T is partially pre-coordinated (since it is less coordinated than its saturation). The pre-coordinated terms might be the most frequently used, or the most convenient for use, or they might form a more intuitive basis of the full coordination. Speculatively, we might expect that the coordination factor (relative to B) of T_{total} to be an order of magnitude greater than that of T in a nicely designed system.

3.2 Approach for estimating level of coordination in medical terminologies

It is not immediate how to determine the amount of coordination in a given terminology, especially if we are aiming at the *inherent* amount of coordination. This requires determining the smallest basis for the terminology and even highly factored systems may not be minimal. In more typical cases or in highly pre-coordinated terminologies, the basis may never have been made explicit. Estimates for the size of the basis of medical terminologies have varied widely between 20,000 and 1,000,000 [24, 34].

However, we can attempt to estimate the amount of coordination in existing terminologies by parsing and analysing the terms explicitly present in the terminology. To obtain loose lower and upper bounds of the size of the basis for each terminology, we have analysed the full descriptions of the terminologies listed in Table 1 using two different approaches, described in the following. For the purpose of this analysis we were only interested in the level of coordination of the core terms of each terminology, therefore chose to ignore synonyms that are available for some of the terminologies analysed.

Approach 1: We have tokenized the descriptions by breaking them up into separate tokens using the Apache Lucene[1] standard tokenizer. On the one hand, this approach results in single tokens to be considered as part of the basis that domain experts would potentially not consider to be part of the basis (e.g., stop words, such as 'of', 'the', 'in', and words that do not make sense on their own in the context of the terminology, e.g., 'due', 'using', or 'object'). On the other hand, concepts that domain experts might consider to be part of the base, but that consist of multiple tokens are missed (e.g., 'blood group', 'cardiac arrest', or 'foreign body').

Approach 2: Some stop words, such as 'of', 'and', 'with', or 'by'[2] and punctuations, such as ',', can be considered as indicators for pre-coordination of a term. Following this intuition, we have broken up the descriptions into the parts that are separated by any of these stop words and commas while excluding these words from the basis. Following this approach, for example, the term 'entire superior segment of left lower lobe of lung', a term specifying a body structure in SNOMED CT, is broken up into 'entire superior segment', 'left lower lobe', and 'lung'. However, if laterality, such as 'left' and 'right' is considered to be part of the basis, 'left lower lobe' could further be broken up into 'left' and 'lower lobe', the latter of which could potentially be further broken up into 'lower' and 'lobe'. Examples where this approach is successful in identifying what could be considered atomic concepts include 'fracture of pelvis' (i.e., 'fracture', 'pelvis') in NCIt, 'incision and drainage of perianal abscess' (i.e., 'incision', 'drainage', 'perianal abscess') procedure in SNOMED CT, or 'removal of drainage device from lower back, percutaneous endoscopic approach' (i.e., 'removal', 'drainage device', 'lower back', 'percutaneous endoscopic approach') procedure in ICD-10-PCS.

For each terminology, we then identified the set of unique terms or atomic concepts, i.e., the basis B of the terminology produced by each of the two approaches. Once we have the size of the basis, it is easy to determine the coordination factor of the terminology (see Definition 5).

In addition to the coordination factor itself, there are other measures that can provide indications of the level of coordination of the terminologies. These include the length of the descriptions with respect to the number of tokens or atomic concepts of the base, and the usage of the tokens or atomic concepts of the base, i.e., how many of the concepts in the base are used in the pre-coordinated terms and how often they are used. In case of the former, longer terms suggest higher levels of coordination, and for the latter, more frequent usage of a greater number of concepts of the base also suggest higher levels of coordination. In contrast, in a predominantly post-coordinated terminology, the average length of the descriptions would be expected to be fairly short and none or very little re-use of the concepts of the base would be expected.

To evaluate the length of the descriptions, we have calculated the distribution of terms with a particular number of tokens as identified using the first approach as well as the maximum, median and average number of atomic concepts in terms as determined by approach 2. To determine the usage of the concepts in the basis, we have calculated the maximum, median and average occurrence of tokens and atomic concepts determined by both approaches, as well as the number and percentage of tokens that appear only in a single term, along with the average length of those terms with respect to its number of tokens. We expect that not all members of the base contribute equally to the pre-coordination of a terminology, i.e., some tokens might appear more frequently than others and a number of them might only be used once. In the case of terminologies that support post-coordination, some of the tokens that appear only once might be atomic concepts to be used for post-coordination, but others might be tokens that are only used very rarely. To

[1] http://lucene.apache.org

[2] complete list: {"and", "at", "by", "for", "from", "in", "into", "no", "not", "non", "of", "on", "or", "with", "within", "without"}

gain an insight into the change in level of coordination of a terminology during its life cycle, we have analysed the coordination factor and the sizes of the base (using approach 1) and the whole explicit terminology of 115 versions of the NCI thesaurus released almost every month between late 2003 and July 2013.

Unfortunately, determining the saturated coordination is not possible without out of band knowledge of the coordinating function. This might be given by a set of rules or by examining the pattern of use in a large application. It is not clear whether we can give an interesting loose estimation: If we allow arbitrary repetition of base terms in coordinated terms, then we have an infinite number of possible terms. If we disallow repetition, the total possible set of terms is exponential in the size of the basis. It is highly unlikely that the true saturated coordination is remotely close in size to the potential term space, as most combinations of terms from the basis will be clinically nonsensical [36, 34].

Table 1. Terminologies analysed.

Terminology	Description
SNOMED CT	Concepts of the Systematic Nomenclature of Medicine Clinical Terms, release 07/2013 [6].
NCI Thesaurus	Reference terminology covering clinical care, translational and basic research and administrative activities (v.13.07e) [8].
SNOMED CT disorder	Subset of SNOMED CT describing disorders.
SNOMED CT finding	Subset of SNOMED CT describing findings.
SNOMED CT CORE problem list	Subset of SNOMED CT containing information most useful for capturing clinical information at a summary level [9].
SNOMED CT CORE disorder	Subset of SNOMED CT CORE describing disorders.
SNOMED CT CORE finding	Subset of SNOMED CT CORE describing findings.
ICD-9-CM diagnosis codes	International Classification of Diseases, Clinical Modifications; Long descriptions of version 30 of the diagnosis codes, released in 2012 [5].
ICD-10-CM diagnosis codes	International Classification of Diseases, Clinical Modifications; Long code descriptions of 2014 release of the diagnosis coding system [3].
SNOMED CT procedure	Subset of SNOMED CT describing procedures.
SNOMED CT CORE procedure	Subset of SNOMED CT CORE describing procedures.
ICD-9-CM procedure codes	International Classification of Diseases, Clinical Modifications; Long descriptions of version 30 of the procedure codes[5].
ICD-10-PCS procedure codes	International Classification of Diseases, Procedure Coding System; Long code descriptions of the 2014 release of the ICD-10-PCS procedure coding system [4].
SNOMED CT body structure	Subset of SNOMED CT covering anatomy.
FMA	Foundational Model of Anatomy ontology (v3.2.1) [1].
LOINC	Logical Observation Identifiers Names and Code (LOINC 2.44), which includes laboratory tests and other clinical observations [7].
LOINC top 2000 SI lab results	Subset of LOINC of most frequently used lab observations (SI version).
LOINC top 300 lab orders US	Subset of LOINC covering most frequently used lab order codes in US.
Gene Ontology (GO)	Controlled vocabulary of gene and gene product attributes, covering biological process, cellular components and molecular functions; weekly releases, downloaded August 20, 2013 [2].
GO biological process	Subset of GO describing biological processes.
GO cellular component	Subset of GO describing cellular components.
GO molecular function	Subset of GO describing molecular functions.

3.3 Source Data

Where available, we collected the OWL representation of ontologies (SNOMED CT, NCIt, FMA, and Gene Ontology), in all other cases we collected the source files from the terminologies listed in Table 1. The terminologies are grouped according to their contents, with the first group containing SNOMED CT and NCI thesaurus (NCIt) representing general terminologies. The second group covers diagnosis descriptions and contains ICD-9-CM, ICD-10-CM and the corresponding subsets disorder and findings of SNOMED CT as well as the Clinical Observations Recording and Encoding (CORE) subset [28] of SNOMED CT, which contains terms frequently used in problem lists and was placed into this group based on the analysis of its content in comparison to ICD carried out in [40]. The third group covers procedures and contains the ICD-9 procedures, ICD-10-PCS and the corresponding subset of SNOMED CT and SNOMED CT CORE describing procedures. The fourth group covers anatomy and contains the Foundational Model of Anatomy ontology and the corresponding body structure subset of SNOMED CT. The remaining groups are LOINC along with its subsets of the top 2000 most frequently used lab results and the top 300 most frequently used lab orders, and the whole of the Gene Ontology (GO) along with its separate parts covering biological processes, cellular components and molecular functions. Subsets of the most frequently used terms, i.e., CORE problem subset of SNOMED CT and the subsets of LOINC, were analysed separately in addition to the whole terminology to evaluate whether a difference in coordination level of the most frequently used terms in comparison to the whole terminology can be observed.

4 Results

The distributions of terms with a certain number of tokens as determined using approach 1 are shown in Figure 2. The increase in the length of the terms in ICD suggest an increase in the level of coordination in both cases, ICD procedure and diagnosis between the last release of ICD-9 in 2012 and the most recent release of ICD-10. Differences can also be observed between the CORE subset of SNOMED CT and the corresponding whole set of terms as illustrated in the case of SNOMED CT and CORE SNOMED CT, as well as the corresponding subsets for disorder and finding, with the terms in the CORE subset tending to be shorter, and potentially less coordinated. A comparison of the distributions of tokens of a particular length between SNOMED CT in Figure 2 and NCIt in Figure 4, NCIt appears to be less coordinated.

The results of the analysis of the size of each terminology as well as the size of the corresponding basis, the median and maximum length of the terms and the coordination factor with respect to the basis determined using both approaches 1 and 2 are shown in Table 2. Table 3 presents the results of the usage analysis of the terminologies, i.e., how many members of the base occur in how many terms and how many members of the basis occur only once.

Considering that SNOMED CT supports and encourages post-coordination, its coordination factor is perhaps higher than expected, in particular when compared to pre-coordinated terminologies such as LOINC. However, evaluations of SNOMED CT suggest that it contains a mixture of pre-coordinated terms and atomic terms that can be post-coordinated (e.g., [29]). The analysis of various of the categories of SNOMED CT covering, e.g., disorder, finding, procedure, and body structure, to mention only those included here, shows significant differences in their coordination factors, e.g., 6.33 for body structure and 2.97 for finding (for approach 1). Similar differences have been observed for other categories not included here. This would suggest that the coordination factor of the whole of SNOMED CT might not be representative and further analyses would best be carried out on the subsets covering individual categories. Similar differences between the whole terminology and its specific subsets can also be observed for the Gene Ontology.

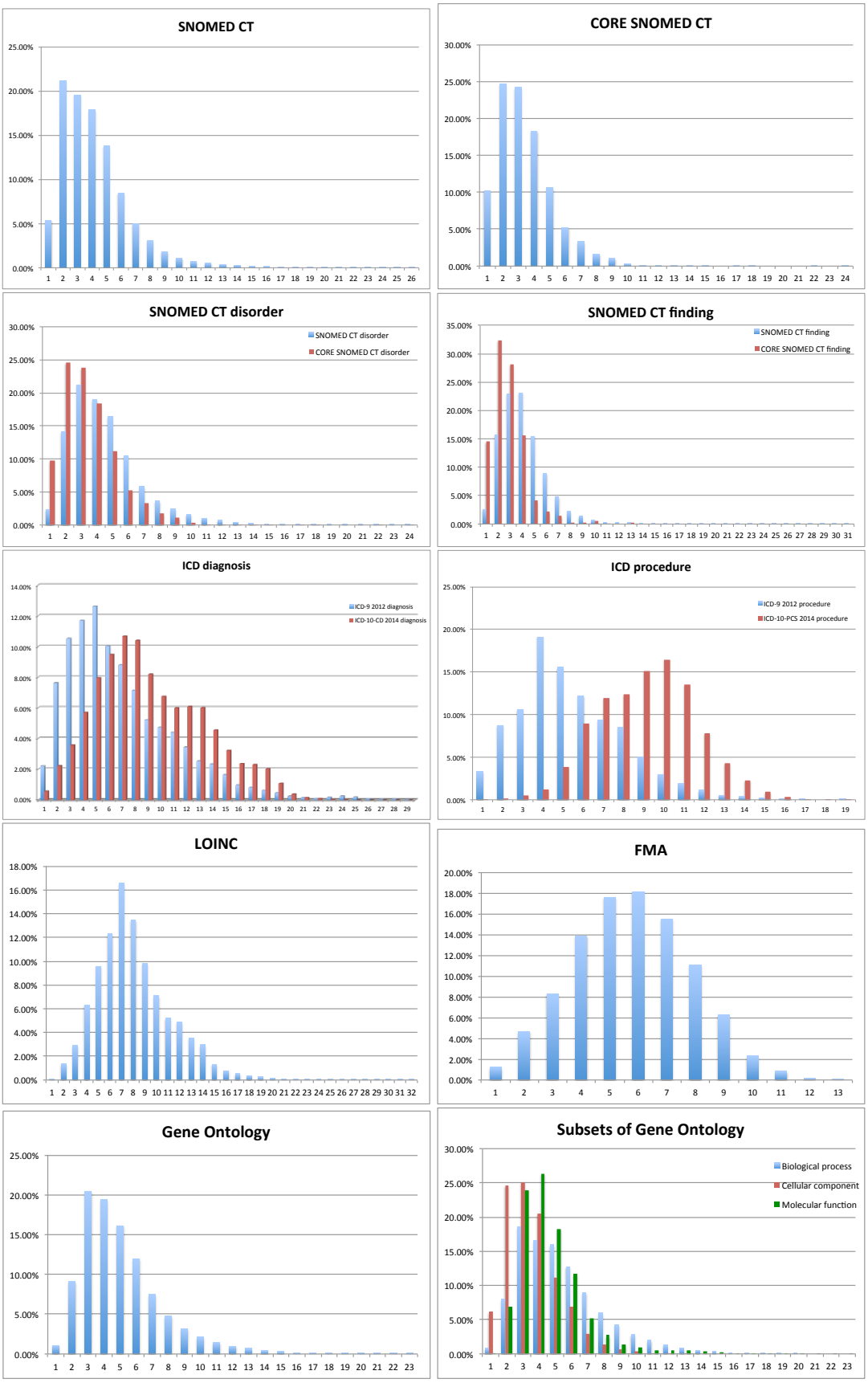

Fig. 2. Distribution of terms with a certain number of tokens, x-axis: number of tokens, y-axis: percentage of terms in the terminology with number of tokens

Table 2. Size of the terminologies, size of their basis and their coordination factor.

Terminology	# of entries	Size of basis		Median/max no of tokens per term		Coordination factor	
		Appr. 1	Appr. 2	Appr. 1	Appr. 2	Appr. 1	Appr. 2
SNOMED CT	347,427	73,349	279,346	4/33	1/16	4.74	1.24
NCI thesaurus 13.07e	98,865	40,988	98,858	3/29	1/11	2.41	1.00
SNOMED CT disorder	82,551	18,448	61,535	4/26	2/14	4.47	1.33
SNOMED CT finding	36,205	12,185	35,145	4/33	2/16	2.97	1.03
SNOMED CT CORE problem list	6,305	3,975	5,988	3/24	1/7	1.59	1.05
SNOMED CT CORE disorder	4,457	2,953	4,150	3/24	1/7	1.51	1.07
SNOMED CT CORE finding	954	1,128	1,052	3/13	1/4	0.85	0.91
ICD-9-CM diagnosis codes 2012	14,567	5,875	10,370	6/29	2/15	2.48	1.40
ICD-10-CM diagnosis codes 2014	91,737	7,170	17,601	8/29	4/14	12.79	5.17
SNOMED CT procedure	58,170	14,554	43,275	5/21	1/13	4.00	1.34
SNOMED CT CORE procedure	547	768	670	3/15	1/5	0.71	0.82
ICD-9-CM procedure codes 2012	3,878	2,159	3,097	5/19	2/13	1.80	1.25
ICD-10-PCS procedure codes 2014	72,769	1,360	9,094	9/19	4/8	53.51	8.00
SNOMED CT body structure	26,816	4,236	20,175	5/15	2/6	6.33	1.32
FMA 3.2	84,453	4,633	33,104	6/13	2/6	18.23	2.55
LOINC 2.44	64,820	13,863	44,628	8/33	2/11	4.68	1.45
LOINC top 2000 SI lab results	2,012	1,502	1,663	8/26	8/16	1.34	1.22
LOINC top 300 lab orders US	329	491	349	7/26	8/14	0.67	0.94
Gene Ontology (GO)	38,010	9,586	31,688	4/23	1/13	3.97	1.20
GO biological process	25,181	5,340	19,245	5/20	1/8	4.72	1.31
GO cellular component	3,227	2,251	3,249	3/13	1/4	1.43	0.99
GO molecular function	9,599	5,222	9,306	4/23	1/13	1.84	1.03

An observation that applies to both terminologies for which subsets of frequently used terms could be obtained (SNOMED CT and LOINC) is that the coordination factor of the frequently used terms are significantly lower than those of the whole terminology, even when the corresponding subsets for the categories in SNOMED (disorder and finding) are considered. In the case of SNOMED this could suggest that the most frequently used terms are atomic concepts that are being post-coordinated, however, as LOINC does not support post-coordination, this does not apply to LOINC.

A significant increase in the coordination factor can be observed for both, the diagnosis and the procedure codes between ICD-9 and ICD-10, with the former increasing from 2.48 to 12.79 and the latter from 1.80 to 53.51 (for approach 1) with the base decreasing in size in the case of the latter. ICD-10 aims to contain an as complete as possible list of variations of diagnoses and procedures, with, in particular the procedures being descriptions of the procedures rather than their names [39], which could explain the high coordination factors. The observation of a high level of pre-coordination is further supported by the low percentage of tokens of the base obtained using approach 1 that are not re-used in multiple terms, i.e., that occur only once. As can be seen in Table 3, the percentages for ICD-10 are lower than the percentages observed for the other terminologies, with a significantly lower percentage of tokens not used multiple times for the procedures in ICD-10 (only about 5%), which means that 95% of the basis of ICD-10 is used multiple times.

In contrast, the percentage of those that occur only in a single term is surprisingly high across the majority of the other terminologies (around 50%) with the median length of the terms in

Table 3. Usage of the base of the terminologies, i.e., the maximum and average number of terms a member of the base (excl. stop words) occurs in, as well as the number of members of the base that appear in only a single term, i.e., are not re-used.

Terminology	Max no of terms a token occurs in		Avg no of terms a token occurs in		No of tokens occurring once	% of tokens occurring once	Median length of terms with tokens occurring once
	Appr. 1	Appr. 2	Appr. 1	Appr. 2	Appr. 1	Appr. 1	Appr. 1
SNOMED CT	10,232	17,241	20.03	1.62	35,073	47.82	2
NCI thesaurus 13.07e	5,021	1,107	8.49	1.17	22,179	54.11	2
SNOMED CT disorder	4,742	4,377	20.67	1.78	7,940	48.48	2
SNOMED CT finding	3,932	3,288	12.83	1.58	5,051	41.45	2
SNOMED CT CORE problems	310	155	5.47	1.52	1,927	48.48	2
SNOMED CT CORE disorder	300	86	5.27	1.51	1,463	49.54	2
SNOMED CT CORE finding	106	22	2.41	1.30	726	64.36	2
ICD-9-CM diag. codes 2012	3,014	7,443	17.14	3.08	2,457	41.82	4
ICD-10-CM diag. codes 2014	35,646	26,136	116.79	3.96	2,256	31.46	3
SNOMED CT procedure	5,189	1,682	19.53	1.96	6,422	44.13	3
SNOMED CT CORE procedure	36	19	2.64	1.65	475	61.85	2.5
ICD-9-CM proc. codes 2012	242	842	9.73	2.58	938	43.45	4
ICD-10-PCS proc. codes 2014	58,736	19,606	490.04	3.44	72	5.29	8
SNOMED CT body structure	9,638	6,946	29.04	1.85	1,227	28.97	1
FMA 3.2	84,453	4,084	105.00	2.20	1,370	29.57	2
LOINC 2.44	64,820	17,532	37.83	2.31	4,317	31.14	8
LOINC top 2000 SI lab results	855	860	10.53	2.60	668	44.47	8
LOINC top 300 lab orders US	177	161	5.25	2.70	279	56.82	8
Gene Ontology (GO)	8,703	3,034	20.06	1.52	3,945	41.15	4
GO biological process	8,301	3,034	25.46	1.71	1,384	25,92	3
GO cellular component	1,477	15	5.07	1.08	1,249	55.49	3
GO molecular function	8,032	588	8.60	1.18	2,888	55.30	4

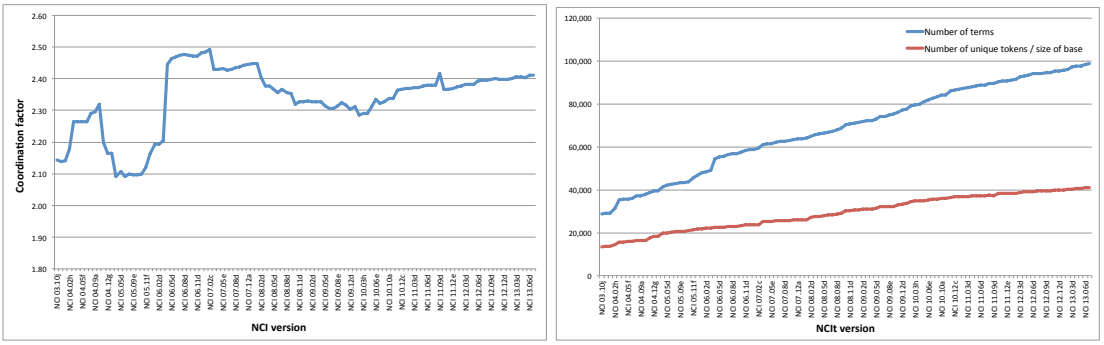

Fig. 3. On the left: Change of the coordination factor of NCIt, x-axis: NCIt version, y-axis: coordination factor. On the right: Change of the size of NCIt, x-axis: NCIt version, y-axis: size of base/whole ontology in number of tokens/terms.

which these rarely used tokens appear suggesting that these are not only atomic concepts to be used for post-coordination, which is only supported by SNOMED CT.

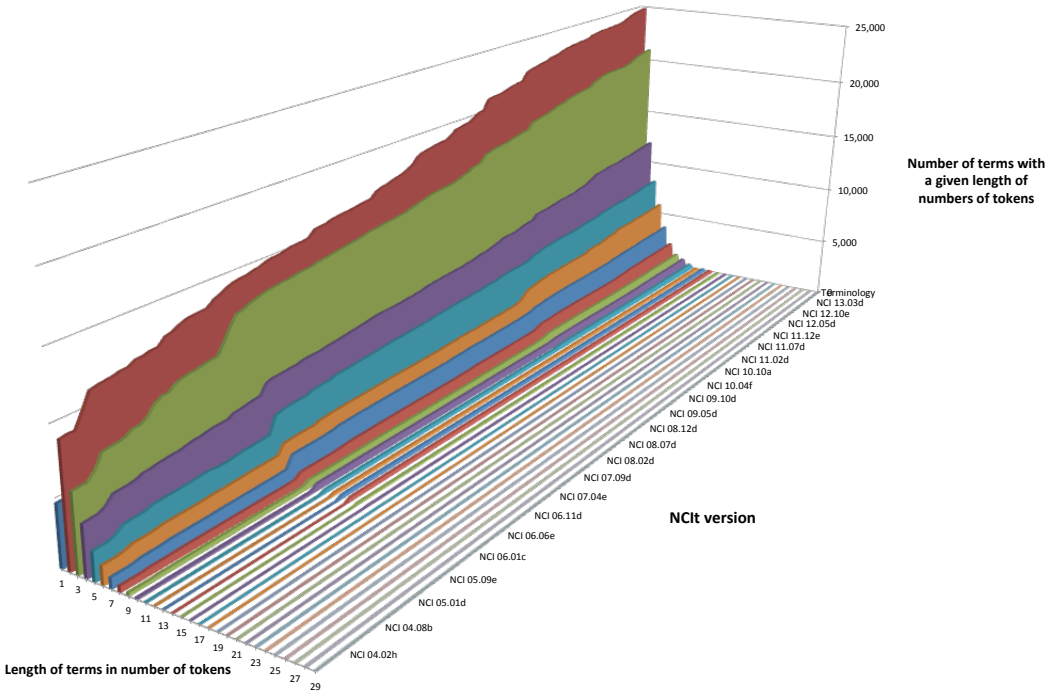

Fig. 4. Change of the distribution of terms with a specific length in number of tokens over time.

The results of the analysis of the NCI thesaurus, namely the observed change in coordination factor, increase in size of the base as well as the whole terminology and the changes in the distribution of terms with a particular number of tokens are shown in Figures 3 and 4. In comparison to the change in coordination factor observed between ICD-9 and ICD-10, the coordination factor of NCIt does not change significantly over time. However, the plot of the coordination factor over time on the left hand side of Figure 3 suggests that efforts are being undertaken to reduce the level of coordination of NCIt, which is followed by an increase of the co-ordination factor, most likely due to new terms being added to the NCI thesaurus. The reductions in the level of coordination do not correlate with significant reductions in the size of the terminology or the length of the terms, as can be seen by the steady increase of both (see on the right hand side of Figure 3 and Figure 4).

5 Discussion

The key advantage of our technique for determining the coordination factor along with supporting measures providing indications for the level of coordination is that it is done completely analytically and computationally. No domain expertise or manual inspection of a sample of terms (a la [18]) is required. Furthermore, we do not need access to the often implicit, inchoate, or incomplete actual coordination function in order to gain some insight into the coordination strategy of the terminology. The price of these conveniences is that we currently have no method for estimating the saturated variant (and thus estimate the amount of post-coordination needed) and the process generates no material level insight into the structure of the terminology (for the

simple reason that it does not examine the content per se). While the lack of content insight is inherent in any analytical approach, the inability to estimate the true term space is unfortunate and can affect the actual coordination factor, as can be seen by the results produced using the two different approaches (see Table 2). We hope to address this lack with extrapolation methods based on 1) the distribution pattern of terms, 2) a semantic and frequency based categorization of the basis, and 3) token combination pattern learning, and more domain aware approaches.

The coordination level has previously been suggested as a quality measure for terminologies [31]. For example, it is easy to see the dramatic increase in pre-coordination between ICD-9 and ICD-10. Perhaps more surprising is the relatively high coordination factor for SNOMED-CT, although it is well known that while SNOMED-CT is geared toward post-coordination in general, it contains a substantial number of pre-coordinated terms (e.g., [29]). We intend to extend such analyses to other aspects of terminologies, such as hierarchical relations.

Different levels of coordination in different terminologies have also been shown to cause issues or require additional work when mapping between terminologies (e.g., [15, 42]). Having an easily computable measure of the level of coordination can help assess the effort required to map between different terminologies. Further analysis, such as overlap of the bases could further help with the assessment of the effort required.

In case of terminologies that are being adapted from pre-coordinated to post-coordinated terminologies, the measures presented here enable a continuous assessment of the progress made in the conversion process.

As the analysis presented here suggests that most commonly used terms have a lower coordination factor than the terminology, highly coordinated terms might deserve some analysis to determine the reasons for their limited usage and whether they could be removed from the terminology or whether they are pre-coordinated terms of post-coordinated most commonly used terms.

In addition to having a measure for assessing the level of coordination in terminologies, the basis that is obtained as part of the process presented here, can also be utilised further. The analysis of the usage of the basis can highlight tokens or atomic concepts that are rarely used, and perhaps deserve further analysis to decide whether they should be included in the basis, or whether some pre-coordinated terms have been missed. The basis itself could prove useful for the maintenance and development of the terminology. For example, an agreed canonical basis of the preferred names for the atomic concepts to be used with synonyms could be utilised to ensure that the preferred term is used, and not one of the multiple synonyms. For example, inconsistencies in term usage, such as in the findings in SNOMED CT 'Amputated big toe' and 'Cock-up deformity of great toe' where the big toe is called big toe in one finding and great toe in the other finding. The same inconsistency can be observed in FMA, e.g., 'Dorsal surface of great toe' and 'Eponychium of big toe'. Other examples include the representation of ordinal numbers, which are written out in some terms and written as numbers in other terms within the same terminology. This inconsistent usage could be one reason for the observed high percentage of rarely used tokens in the basis. In NCIt, the following terms can be found: 'Cardiac Arrest' and 'CTCAE Grade 5 Asystole'. Using the Unified Medical Language System (UMLS) [14] to lookup synonyms for 'cardiac arrest' suggests that 'asystole' is a synonym. Ideally, the same name for a concept would be used consistently throughout the whole terminology. This, however, is very hard to ensure without the knowledge of the basis of the terminology.

In addition to the basis, a coordination function could furthermore help to ensure that the order in which the pre-coordinated complex terms are created is consistent, e.g., in ICD-10, a diagnosis called 'Abrasion, left great toe' can be found, with the laterality associated with the great toe, without the explicit mention of the foot to which the great toe belongs. In contrast, the diagnosis called 'Fused toes, left foot' lists the corresponding foot and assigned the laterality to

the foot rather than the toes. These inconsistencies make maintenance and usage of terminologies harder, and could be avoided with the identification and then usage of coordination functions.

References

1. FMA, http://sig.biostr.washington.edu/projects/fm/
2. GO, http://www.geneontology.org
3. ICD-10-CM, http://www.cms.gov/Medicare/Coding/ICD10/2014-ICD-10-CM-and-GEMs.html
4. ICD-10-PCS, http://www.cms.gov/Medicare/Coding/ICD10/2014-ICD-10-PCS.html
5. ICD-9, http://www.cms.gov/Medicare/Coding/ICD9ProviderDiagnosticCodes/codes.html
6. International Health Terminology Standards Development Organisation (IHTSDO), http://www.ihtsdo.org/
7. LOINC, http://loinc.org
8. NCIthesaurus, http://ncit.nci.nih.gov
9. SNOMED, http://www.nlm.nih.gov/research/umls/Snomed/core_subset.html
10. WHO — International Classification of Diseases (ICD), http://www.who.int/classifications/icd/en/
11. Averill, R.F.R., Mullin, R.L.R., Steinbeck, B.A.B., Goldfield, N.I.N., Grant, T.M.T.: Development of the ICD-10 Procedure Coding System (ICD-10-PCS). Journal of AHIMA / American Health Information Management Association 69(5), 65–72 (Apr 1998)
12. Bakhshi-Raiez, F.F., de Keizer, N.F.N., Cornet, R.R., Dorrepaal, M.M., Dongelmans, D.D., Jaspers, M.W.M.M.: A usability evaluation of a SNOMED CT based compositional interface terminology for intensive care. International Journal of Medical Informatics 81(5), 351–362 (Apr 2012)
13. Bechhofer, S., Stevens, R., Ng, G., Jacoby, A., Goble, C.: Guiding the user: an ontology driven interface. In: Proceedings User Interfaces to Data Intensive Systems. pp. 158–161. IEEE (1999)
14. Bodenreider, O.: The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acicds Research 32(Database issue), D267–70 (Jan 2004)
15. Bodenreider, O.: Issues in mapping LOINC laboratory tests to SNOMED CT. AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium pp. 51–55 (2008)
16. Ceusters, W.W., Smith, B.B., Goldberg, L.L.: A terminological and ontological analysis of the NCI Thesaurus. Methods of information in medicine 44(4), 498–507 (Jan 2005)
17. Chute, C.G., Cohn, S.P., Campbell, J.R., for the ANSI Healthcare Informatics Standards Board Vocabulary Working Group and the Computer-based Patient Records Institute Working Group on Codes and Structures: A Framework for Comprehensive Health Terminology Systems in the United States: Development Guidelines, Criteria for Selection, and Public Policy Implications. Journal of the American Medical Informatics Association 5(6), 503–510 (Nov 1998)
18. Chute, C.G., Cohn, S.P., Campbell, K.E., Oliver, D.E., Campbell, J.R.: The Content Coverage of Clinical classifications. Journal of the American Medical Informatics Association (1996)
19. Cimino, J.J.J., Clayton, P.D.P., Hripcsak, G.G., Johnson, S.B.S.: Knowledge-based approaches to the maintenance of a large controlled medical terminology. Journal of the American Medical Informatics Association : JAMIA 1(1), 35–50 (Jan 1994)
20. Cornet, R.: Definitions and qualifiers in SNOMED CT. Methods of information in medicine 48(2), 178–183 (Jan 2009)
21. Cornet, R., de Keizer, N.: Forty years of SNOMED: a literature review. BMC Medical Informatics and Decision Making 8(Suppl 1), S2 (2008)
22. Cornet, R., Nyström, M., Karlsson, D.: User-Directed Coordination in SNOMED CT. Studies in health technology and informatics 192, 72–76 (2013)
23. De Silva, T.S., MacDonald, D., Paterson, G., Sikdar, K.C., Cochrane, B.: Systematized nomenclature of medicine clinical terms (SNOMED CT) to represent computed tomography procedures. Computer methods and programs in biomedicine 101(3), 6–6 (Mar 2011)
24. Elkin, P.L., Bailey, K.R., Chute, C.G.: A randomized controlled trial of automated term composition. In: Proceedings of the AMIA Symposium. p. 765. American Medical Informatics Association (1998)

25. Elkin, P.L., Brown, S.H.: Compositionality: An Implementation Guide. In: Terminology and Terminological Systems, pp. 71–94. Springer London, London (Mar 2012)
26. Elkin, P.L., Brown, S.H., Lincoln, M.J., Hogarth, M., Rector, A.: A formal representation for messages containing compositional expressions. International Journal of Medical Informatics 71(2-3), 89–102 (Sep 2003)
27. Evans, D.A.D., Cimino, J.J.J., Hersh, W.R.W., Huff, S.M.S., Bell, D.S.D.: Toward a medical-concept representation language. The Canon Group. Journal of the American Medical Informatics Association : JAMIA 1(3), 207–217 (Apr 1994)
28. Fung, K.W.K., McDonald, C.C., Srinivasan, S.S.: The UMLS-CORE project: a study of the problem list terminologies used in large healthcare institutions. Journal of the American Medical Informatics Association : JAMIA 17(6), 675–680 (Nov 2010)
29. Goss, F.R., Zhou, L., Plasek, J.M., Broverman, C., Robinson, G., Middleton, B., Rocha, R.A.: Evaluating standard terminologies for encoding allergy information. Journal of the American Medical Informatics Association : JAMIA pp. – (Feb 2013)
30. de Keizer, N.F., Bakhshi-Raiez, F., de Jonge, E., Cornet, R.: Post-coordination in practice: evaluating compositional terminological system-based registration of ICU reasons for admission. International Journal of Medical Informatics 77(12), 828–835 (Dec 2008)
31. Kless, D., Milton, S.: Towards quality measures for evaluating thesauri. Metadata and Semantic Research pp. 312–319 (2010)
32. MacIsaac, P., Walker, D.: Essential SNOMED: Simplifying SNOMED CT and Supporting Integration with Health Information Models. In: Proceedings of KR-MED 2008 (2008)
33. McKnight, L.K., Elkin, P.L., Ogren, P.V., Chute, C.G.: Barriers to the clinical implementation of compositionality. Proceedings / AMIA ... Annual Symposium. AMIA Symposium pp. 320–324 (1999)
34. Rassinoux, A.M., Miller, R.A., Baud, R.H., Scherrer, J.R.: Modeling Just the Important and Relevant Concepts in Medicine for Medical Language Understanding: A Survey of the Issues. In: Proceedings of the IMIA WG6 Working Conference, Jacksonville, FL (1997)
35. Rector, Iannone: Lexically suggest, logically define: Quality assurance of the use of qualifiers and expected results of post-coordination in SNOMED CT. Journal of Biomedical Informatics 45(2), 11–11 (Mar 2012)
36. Rector, A.L.A., Bechhofer, S.S., Goble, C.A.C., Horrocks, I.I., Nowlan, W.A.W., Solomon, W.D.W.: The GRAIL concept modelling language for medical terminology. Artificial Intelligence in Medicine 9(2), 139–171 (Feb 1997)
37. Rogers, J.E.: Development of a methodology and an ontological schema for medical terminology. Ph.D. thesis, School of Computer Science (May 2005)
38. Spackman, K.A., Campbell, K.E.: Compositional concept representation using SNOMED: towards further convergence of clinical terminologies. Proceedings / AMIA ... Annual Symposium. AMIA Symposium pp. 740–744 (Jan 1998)
39. Steindel, S.J.: International classification of diseases, 10th edition, clinical modification and procedure coding system: descriptive overview of the next generation HIPAA code sets. Journal of the American Medical Informatics Association : JAMIA 17(3), 274–282 (Apr 2010)
40. Steindel, S.J.S.: A comparison between a SNOMED CT problem list and the ICD-10-CM/PCS HIPAA code sets. Perspectives in Health Information Management / AHIMA, American Health Information Management Association 9, 1b–1b (Jan 2012)
41. Wade, G., Rosenbloom, S.T.: The impact of SNOMED CT revisions on a mapped interface terminology: Terminology development and implementation issues. Journal of Biomedical Informatics 42(3), 4–4 (May 2009)
42. Wang, Y., Patrick, J., Miller, G., O'Hallaran, J.: A computational linguistics motivated mapping of ICPC-2 PLUS to SNOMED CT. BMC Medical Informatics and Decision Making 8 Suppl 1(suppl 1), S5–S5 (Jan 2008)