from Domenico M. Pisanelli (ed.): *Ontologies in Medicine: Proceedings of the Workshop on Medical Ontologies*, Rome, October 2003. IOS Press, Amsterdam.

Mistakes in Medical Ontologies: Where Do They Come From and How Can They Be Detected?

Werner Ceusters[a], Barry Smith[b], Anand Kumar[b], Christoffel Dhaen[a]
[a] *Language & Computing nv, Hazenakkerstraat 20a, B-9520 Zonnegem, Belgium*
[b] *Institute for Formal Ontology and Medical Information Science, Härtelstrasse 16-18, 04107-Leipzig, Germany*

Abstract. We present the details of a methodology for quality assurance in large medical terminologies and describe three algorithms that can help terminology developers and users to identify potential mistakes. The methodology is based in part on linguistic criteria and in part on logical and ontological principles governing sound classifications. We conclude by outlining the results of applying the methodology in the form of a taxonomy different types of errors and potential errors detected in SNOMED-CT®

1 Introduction

The main goal of Language and Computing nv (L&C) is to deliver advanced natural language understanding applications directed primarily towards the management of terminologies and also towards the coding, semantic indexing, and retrieval and extraction of information, primarily in the healthcare and biomedical domains. Natural language understanding requires knowledge about both reality (i.e. about what is described by language) and about language itself (so that one can assess a language user's current perspective on reality by understanding how he is using language to describe it). To achieve these ends L&C has developed LinKBase®, a realist ontology for the healthcare domain, and LinKFactory®, a multi-purpose ontology authoring and management system. Since 2002 LinKBase® has been developed in close collaboration with IFOMIS, the Institute for Formal Ontology and Medical Information Science of the University of Leipzig, drawing on the IFOMIS methodology for ontology construction and alignment in the biomedical domain based on rigorous formal definitions and axioms [1].

Like L&C, IFOMIS starts out from the idea that we need to understand the general structure and organization of a given domain (its ontology) before we start building software models. This means above all that the basic categories and relations should as far as possible be formally defined in a logically rigorous way from the very start. Such definitions and associated axioms of basic ontology should then serve as constraints on coding and on manual curation of associated ontologies and terminology-systems.

In part as a result of the collaboration with IFOMIS, both LinKBase® and LinKFactory® have now reached a level of maturity that enables them to be used to assess the quality of external systems. This document describes the results of using LinKBase® and two specific algorithms implemented in LinKFactory® to carry out a still on-going review of the January 2003 and July 2003 versions of SNOMED-CT® for possible mistakes and inconsistencies. It explains the basic mechanisms of the approach, as well as the various

types of mistakes that can be detected, and draws conclusions for the methodology of quality assurance in large terminologies in the future.

2 Materials and Methods

In this section we describe the various components that have been used in the error detection process. These are:

- LinkFactory®: a multi-user formal ontology authoring and managing system.
- LinkBase®: a large multi-lingual healthcare ontology and terminology system developed using LinkFactory®.
- Three quality assurance algorithms that are available in LinkFactory® to map ontologies or terminologies onto each other or onto LinkBase®:
    - conservative lexical matching
    - the TermModelling algorithm
    - the Description Logic based classification algorithm.

We also provide a brief introduction to SNOMED-CT®, concentrating on the issues relevant for this document.

2.1 LinkFactory®

Since none of the available ontology management tools were adequate to the task of building the formal ontology required for L&C's natural language understanding applications, L&C developed its own state-of-the-art ontology management tool known as LinKFactory®, which is designed for the purpose of building, managing and maintaining large, complex, language independent ontologies [2]. LinKFactory® provides an effective and user-friendly way to create, maintain and extend extensive multilingual terminology systems and ontologies in a distributed, multi-author environment. Among its key features are:

- The ability to fully represent universals and particulars with all their relevant relationships by means of classes and instances.
- The ability to connect terms in several languages to all entities in the representation in a way which supports natural language processing.
- The ability to connect the resulting association of terms and entities with third party terminology systems such as SNOMED, Read, ICD-9-CM, ICD-10, MedDRA, etc., and to other classifications in neighboring domains.
- The ability to cope with on-going changes in the ontology via a versioning which ensures that the information contained in references to older versions is not lost over time.

2.2 LinKBase®

LinKBase® is a large-scale medical ontology developed by L&C using the ontology authoring environment provided by LinKFactory®. LinKBase® contains over 1.5 million language-independent medical and general-purpose *domain-entities*, representing universals and particulars in the Aristotelian sense. As such domain-entities abstract away from the specific features of natural language representations, fullfilling to that end the same function as *concepts* in other terminologies or ontologies. They are however not equivalent to concepts in the sense that they do not represent abstractions from how humans think about real world entities, but rather the entities themselves, to which such thoughts are directed. Concepts in people's minds, in order that they be clearly separated from the ontology proper, are represented in LinkBase® as *meta-entities*, where they are included in

order to allow mappings to third party terminologies and ontologies. Domain-entities are associated with more than 4 million terms in several natural languages [3]. A *term* consists of one or more *words*, and the latter may be associated with other domain-entities in their turn.

Domain-entities are linked together into a semantic network in which some 480 different link types are used to express formal relationships. The latter are derived both from the specific requirements of semantics-driven natural language understanding [7, 8] and also from formal-ontological theories of mereology and topology [4, 5], of time and causality [6, 18] and of the rules governing classification [16, 17]. A good classification will satisfy rules for example to the effect that each class on any given level in the hierarchy will subsume a plurality of classes lower down in the hierarchy; and that classes which share a subclass in common are either identical or one is subsumed by the other. In addition a good classification will respect the ontological dichotomy between entities in reality and our knowledge about and our concepts of such entities [19].

Link types form a multi-parented hierarchy in their own right. At the heart of this network is the formal subsumption (*is-a*) relationship, which in LinKBase® covers only some 15% of the total number of relationships involved. As such, LinKBase® has a much richer structure than do Description Logic-based terminological ontologies in which the ontology builder is limited effectively to relationships such as: *is strictly narrower than* and *is strictly broader than*. LinKBase® is a living ontology, in which entries are changed at a rate of some 2000 to 4000 modifications a day and in such a way that domain-entities can be added even before they have been completely defined [9]. In addition, the set of available relationships has been periodically expanded to accommodate new demands and finer ambiguity resolution in ways which have necessitated thorough revision of its existing domain-entity definitions. Currently, the system is being re-engineered in conformity with the theories of Granular Partitions [10] and Basic Formal Ontology [11].

2.3 LinkFactory®'s Conservative Lexical Matching algorithm

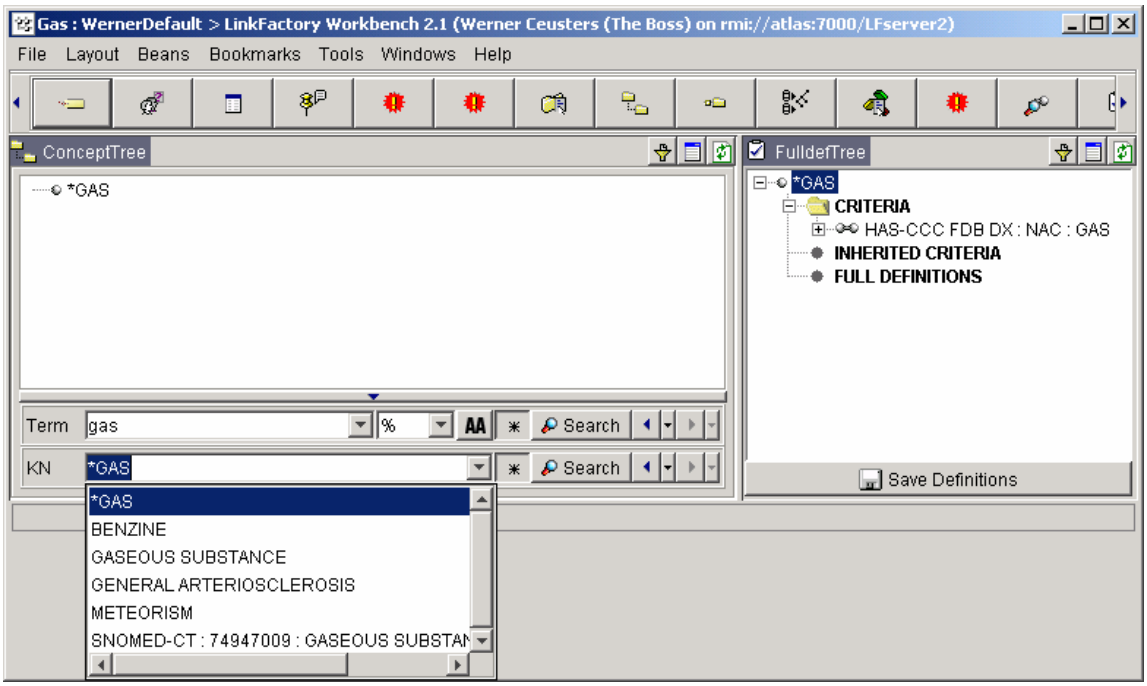

Figure 1. Conservative lexical matching

LinkBase® maintains relationships among domain-entities and terms as n-to-m correspondences, so that both homonymy and synonymy are adequately taken care of. Querying LinkBase® for the term "gas", for example, gives six results (see the list box in the lower left corner of Figure 1). The first five results are LinkBase® domain-entities, while the last is a SNOMED-CT® concept. The first result, "*GAS", is a domain-entity that was created in the course of mapping a new terminology (FDB DX) to LinkBase®. Since that terminology does not contain any sort of computer-understandable meaning representation for the terms it contains, LinkFactory® cannot calculate what the FDB DX term "gas" actually means; hence it flags this term for further review. The second to fifth results are domain-entities for which the term "gas" is a homonym.

2.4 LinKFactory®'s TermModelling Algorithm

LinKBase's TermModelling algorithm uses ontological and linguistic information to seek out missing relationships. Input is in the form of terms in a given language. The algorithm then works by attempting to find domain-entities which enjoy the closest (where possible an exact) match to these terms. To achieve this, the algorithm makes use not only of terms already stored but also of linguistic variants generated on the fly, as also of the ontological definitions of the corresponding domain-entities, whether or not these are complete. In the following paragraphs we describe the algorithm in its simplest form (which is to say: without the optimizations that had to be implemented in order to enable efficient searches over a huge ontology such as LinKBase®). We first describe the algorithm FRVP (for: Find-Relation-Via-Path) that takes not terms but domain-entities as input. We then explain the mechanisms by which an extended algorithm decides what domain-entities to present as input to FRVP on the basis of input terms.

2.4.1 Base algorithm

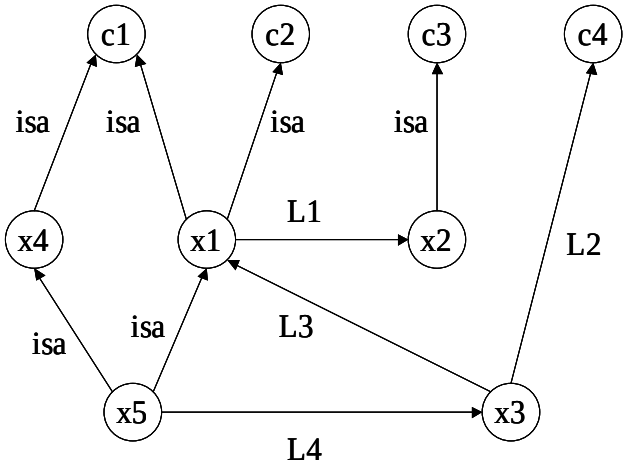

Figure 2: Find relation via path algorithm

All domain-entities in LinKBase® are represented in a directed graph, the links representing either subsumption (*is-a*) or associative relations (L1, …) between the domain-entities (c1, c2, …, see Figure 2). FRVP uses these links to build paths through the graph starting from each given input domain-entity (c1, …, c4) with the goal of finding all domain-entities lying on the intersections of these paths – i.e. at the places where the paths cross. Such an intersection may be *partial*, i.e. it may involve only paths stemming from some of the domain-entities with which we begin, as in cases as x1, x2, and x4. Or it may be *complete*, which means that it lies on paths stemming from all of the initial domain-

entities, as in cases x3 and x5. Thus for example when LinkBase® itself is queried for the domain-entities “tumor”, “stomach” and “removal”, then FRVP will retrieve “stomach tumor” as a partial intersection, and “gastric neoplasm removal” as a complete intersection, these results being in conformity with what we should expect from a well-structured ontology.

In figure 2, both x3 and x5 are complete intersections. In order to establish which of these is closer in meaning to the input an edge-based *cost* calculation needs to be performed. LinkBase® is built in such a way that the smaller the cost related to a path, the more closely are the domain-entities at each side of the path are semantically related. As such, it is easy to verify that x5 provides a closer match to the input than x3. The algorithm is implemented in such a way that, when no complete intersections can be found, partial results are proposed.

The basic TermModelling algorithm is a naïve variant of the FRVP algorithm in the sense that searches start not from domain-entities but from terms. Given a search term T1 made out of words W1, ..., Wn, the simplest way to find the needed domain-entities would be to find all the domain-entities that have as term any substring composed of W1, …, Wn. Note that wherever polysemous words are involved it is already possible to find more domain-entities than words with which one begins. This is shown in Figure 3. The picture shows a search term T1 consisting of two words W1 and W2, where W1 is triply polysemous in a way which yields three distinct LinKBase® domain-entities. To adjust for this problem, the FRVP algorithm was modified to find the intersections of the paths between groups, S1 to Sn, of one or more domain-entities. (Such groups are called *sections* in what follows). This modification can be viewed as if we would be applying the FRVP algorithm to each of the possible combinations of domain-entities associated with given words. For the example in figure 3 we would need to apply the FRVP algorithm three times to find the three complete intersections which exist for the domain-entities (c1, c4), (c2, c4) and (c3, c4), namely: x6, x7, x3 and x8. The complete intersection x6 will be the one best ranked, since all other complete intersections are reached by using incoming links from x6, regardless of the type of links involved.

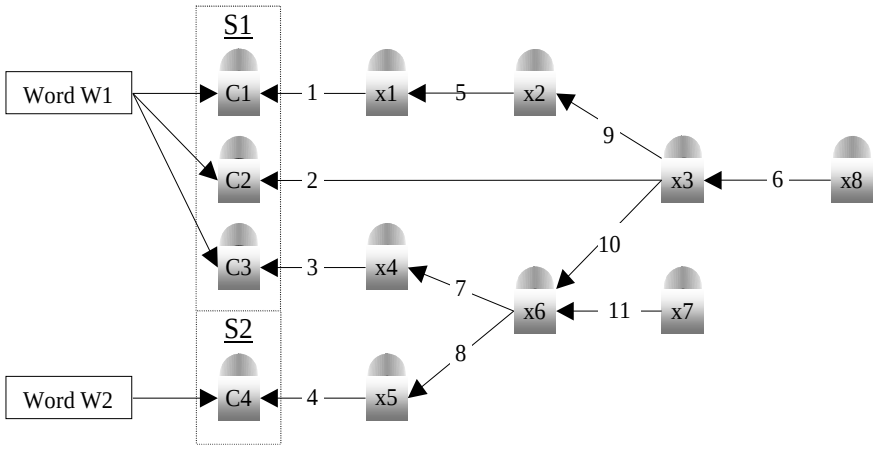

Figure 3 : The FRVP algorithm redesigned to start search from terms instead of domain-entities

2.4.2 Extended algorithm

Several extensions were required to make the TermModelling algorithm find all domain-entities in the ontology that are relevant to a particular query. Most of them are implementations of ideas described in [7]. Problems are posed by terms containing words that are not themselves associated with LinKBase® domain-entities (such as the word “mellitus”

that is never used in isolation but exclusively in combination with “diabetes”) and by the verbal overspecifications of concepts involved in terms such as “dorsal back pain”, or “knee joint arthropathy”. To accommodate these problems, the algorithm was modified in such a way that it also picks up concepts associated with terms containing a subset of the words from the query term. The FRVP algorithm is then used to find complete intersections using the resultant larger set of concepts. As an example, whereas on the basis of the input term “diabetes mellitus treatment” the base algorithm would start its search exclusively with the words “diabetes”, “mellitus” and “treatment” and find no domain-entity related to “mellitus”, the extended algorithm starts also with “diabetes mellitus” for which a domain-entity does indeed exist. The ontology structure of LinKBase® can thus be used as a means for determining the degree to which given input term make any sense at all. If words are combined that do not make sense, then intersecting paths will either not be found, or, if they are found, then the paths involved will be very long and thus have a high cost.

A second extension involves the generation of new terms via the implementation of a language-specific term generator that takes as input the query term, and then generates additional terms based on inflection-, derivation-, and clause-generation rules. As an example, the term “bacterial infection” would generate terms also consisting of the words “bacterium”, “bacteria”, “infections”, “infected”, “infecting”, etc. As with subset generation as described in the previous paragraph, overgeneration could be tamed by checking whether such constructed combinations of words qualify as terms for existing domain-entities in LinKBase®.

Most important for our purposes here is the third extension, which generates larger sections (see figure 3) for a given word by checking the LinKBase® ontology also for translations and/or possible synonyms of the word and its generated words in other languages. Suppose for example that there is a domain-entity with which the terms “pulmonary infarction” and “lung infarction” are associated, but that “pulmonary” is not yet known as a synonym for “lung”. The extended algorithm helps out by finding a path between the term “pulmonary embolism” and the domain-entity for which the term “lung embolism” exists. The cross-language version of this extention goes even further. If there is in LinkBase® a domain-entity annotated with the English term “lung embolism” and in French with the term “embolie pulmonaire”, and if there is also a domain-entity annotated in French with the annotations “infarction pulmonaire” and “infarctus du poumon” (but in this case without any English annotation), and if “lung” and “poumon” are terms for the same domain-entity in English and French respectively, then the algorithm will also find the correct domain-entity for the term “pulmonary embolism”. This method frees us from using additional external systems such as EuroWordNet (which has very poor medical coverage) or the UMLS (which has a minimal coverage of languages other than English). A complete analysis of the algorithm with respect to the capabilities of this third extension for finding underspecification in large ontologies is given in [12]. Here the same capabilities are used to find evidence for a broader range of mistakes in large ontologies.

To process large volumes of terms, typically deriving from third party terminologies, the TermModelling algorithm with its three extensions is embedded in a special component of LinkFactory® called OntologyMatcher that uses a blackboard process control system to allow analyses to be performed in background mode.

2.4.3 Using the TermModelling algorithm for quality assurance

When applied to the task of quality assurance for terminologies, the TermModelling algorithm is used in two different settings, both taking as input the terms that come with the terminologies.

In a first setting the ranking of the semantic distances of the various retrieved entities with respect to each given input term is assessed manually for accuracy. When retrieved entities that at first sight are judged by a human reviewer to be more closely related to the input term receive a higher value than entities that appear less closely related, then this is taken as an indication of some problem in the source terminology.

As an example (Figure 4), the semantic distance 0 for the retrieved SNOMED-CT® concept "387842002: neoplasm of heart" with respect to the query term "heart tumor" (inside SNOMED-CT® a synonym for the concept "92132009: benign neoplasm of heart") is 6. The semantic distance for the SNOMED-CT® concept "387840005: neoplasm of heart and/or pericardium" with respect to the same query term is 26. The former is thus much larger than one would expect, as the second should subsume the first. One can see however in Figure 4 that these two concepts are assigned quite different positions in the SNOMED-CT® concept hierarchy.

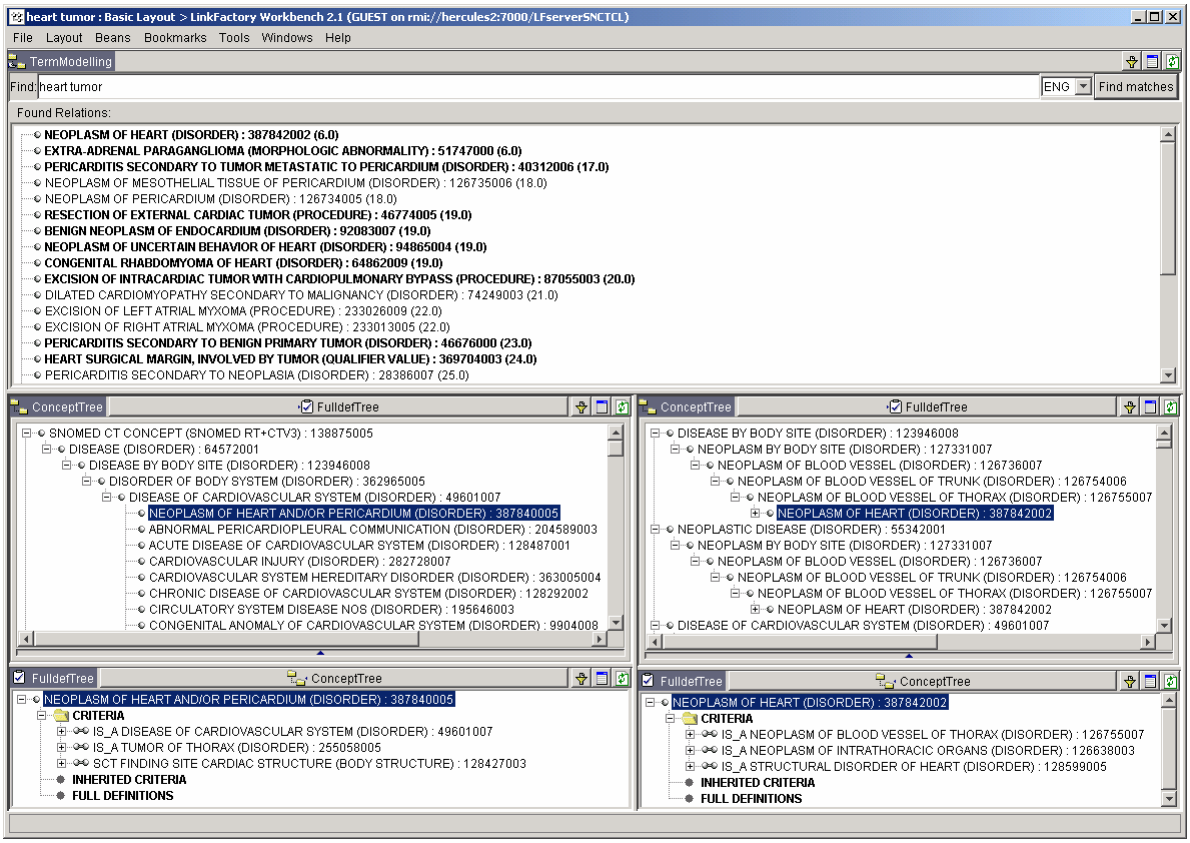

Figure 4: Unexpected differences in semantic distances between the retrieved SNOMED-CT®-concepts "387842002: neoplasm of heart" and "387840005: neoplasm of heart and/or pericardium" (top list) in relation to the term "heart tumor".

One can also wonder why the semantic distance of the concept "51747000: extra-adrenal paraganglioma" from the query term "heart tumor" should be equal to that of "387842002: neoplasm of heart". As one can see from Figure 5, this is due to the fact that SNOMED-CT considers the term "heart base tumor" to be a synonym of "51747000: extra-adrenal paraganglioma" while it is only used in veterinary medicine.

The drawback of this method is its need for manual verification of the results. However, statistical methods can also be used to scan for unusual distributions of semantic distances. For example, if we input a number of terms from the terminology to be evaluated, then we can have them automatically ranked on the basis of their scores in terms of semantic distance from a specific query term. We can then flag those cases in which the

difference in semantic distance between the N-th and (N+1)-th ranked entities is larger than the mean difference over all entities. In addition, we can flag those cases where the semantic distance of the highest ranked retrieved entity for a specific query term is higher than the mean semantic distance of all the highest ranked entities over all query terms. This method does not guarantee to find all mistakes, nor does it guarantee to find only mistakes. Empirical studies are being conducted to find a metric that maximizes recall and precision.

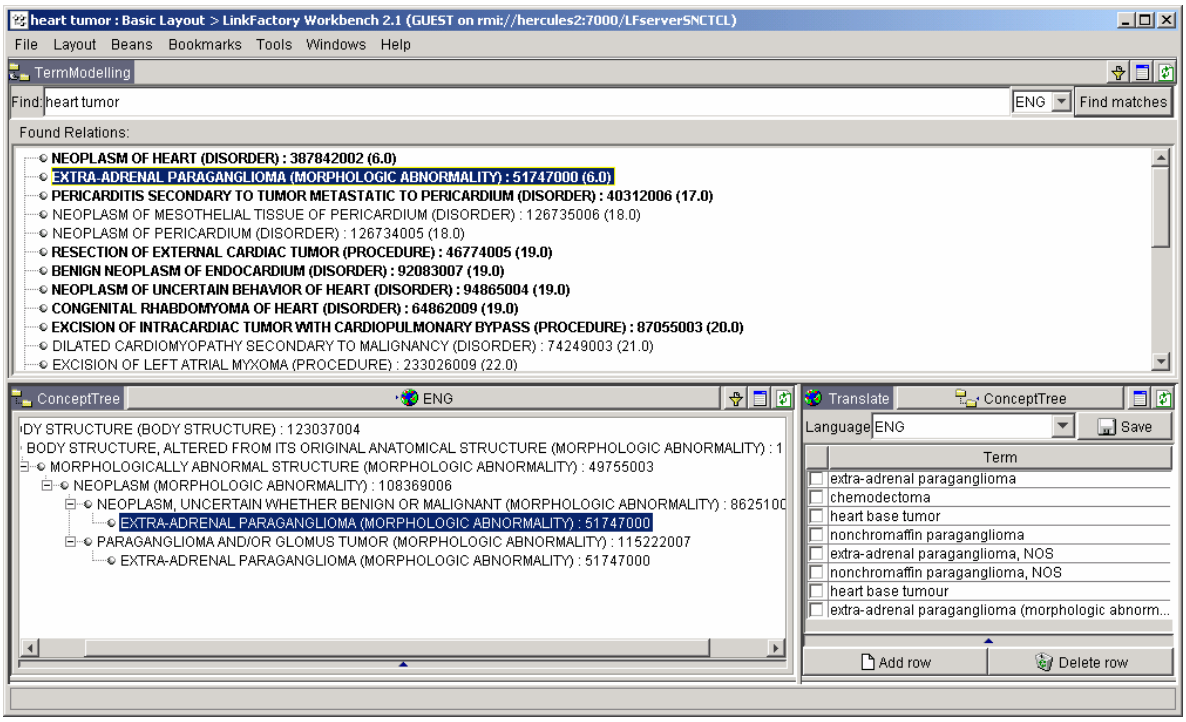

Figure 5: Unexpected low semantic distance from the query term “heart tumor” to the SNOMED-CT® concept “51747000: extra-adrenal paraganglioma”.

A second setting is only applicable to multi-word terms, though these do after all constitute the bulk of every terminology. Here a term is decomposed into its constituent words and the TermModelling algorithm is then applied to both the original term and each of the constituents to find the best fitting domain-entities. Figure 6 shows the results of applying this process with the term “heart tumor”.

The middle part of Figure 6 tells us that the best-fitting SNOMED-CT® concept for the term “heart” is “80891009: heart structure”, while for “tumor” it is “108369006: neoplasm”. Both have the semantic distance 0 with respect to the search terms. The result of using them as input for the FRVP-algorithm is shown in the lower part of the same Figure. The procedure results in two lists that should be identical on the assumption that all SNOMED concepts are perfectly represented both ontologically and linguistically. A different ranking of retrieved entities, more specifically the appearance of concept pairs that are ranked in opposite orders in the two lists – a phenomenon which can be flagged automatically by software – is here a strong indication of inconsistencies. As an example, one can see in Figure 6 that the TermModelling algorithm using as input the term “heart tumor” ranks the concept “387842002: neoplasm of heart” higher than “94865004: neoplasm of uncertain behaviour of heart”. The FRVP-algorithm, in contrast, retrieves the two concepts in the opposite order. In the SNOMED-CT® hierarchy itself, however, one can easily see (Figures 7 and 8) that the concept “94865004: neoplasm of uncertain behaviour of heart” (Figure 7) is inadequately represented, since it is not recorded that it is subsumed by “387842002: neoplasm of heart”.

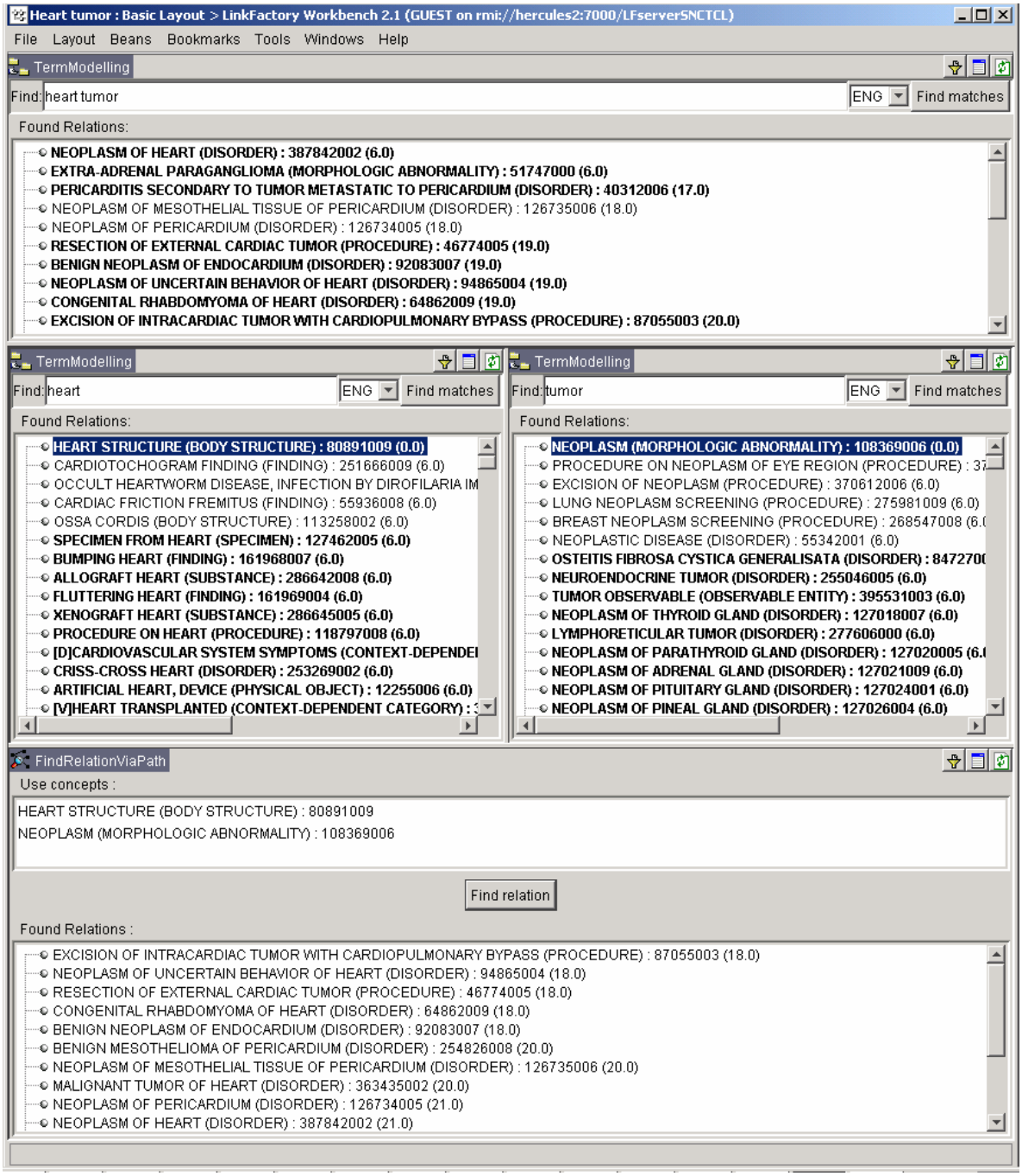

Figure 6: Finding mistakes by comparing the results of pure ontology-based querying with those obtained by adding also linguistics-based querying.

2.5 LinKFactory®'s Classifier Algorithm

The third quality control algorithm in LinKFactory is a Description Logic-based classifier optimised for working with extremely large terminology systems. This algorithm not only computes subsumption relations on the basis of necessary and sufficient conditions for entities defined by the external terminology, it also proposes new entities to be added, based on the distribution of entity-characteristics as computed during the analysis. Parameters can be set for the types of entities generated according to various principles [13]. One simple (but useful) example of such a principle is: if there is a type of object that

causes a specific type of infection then there are infections necessarily caused by objects of that type. Figure 9 (where the concepts starting with "XXX" are those generated by this algorithm) shows part of the generated reclassification of the concept "387842002: neoplasm of heart".

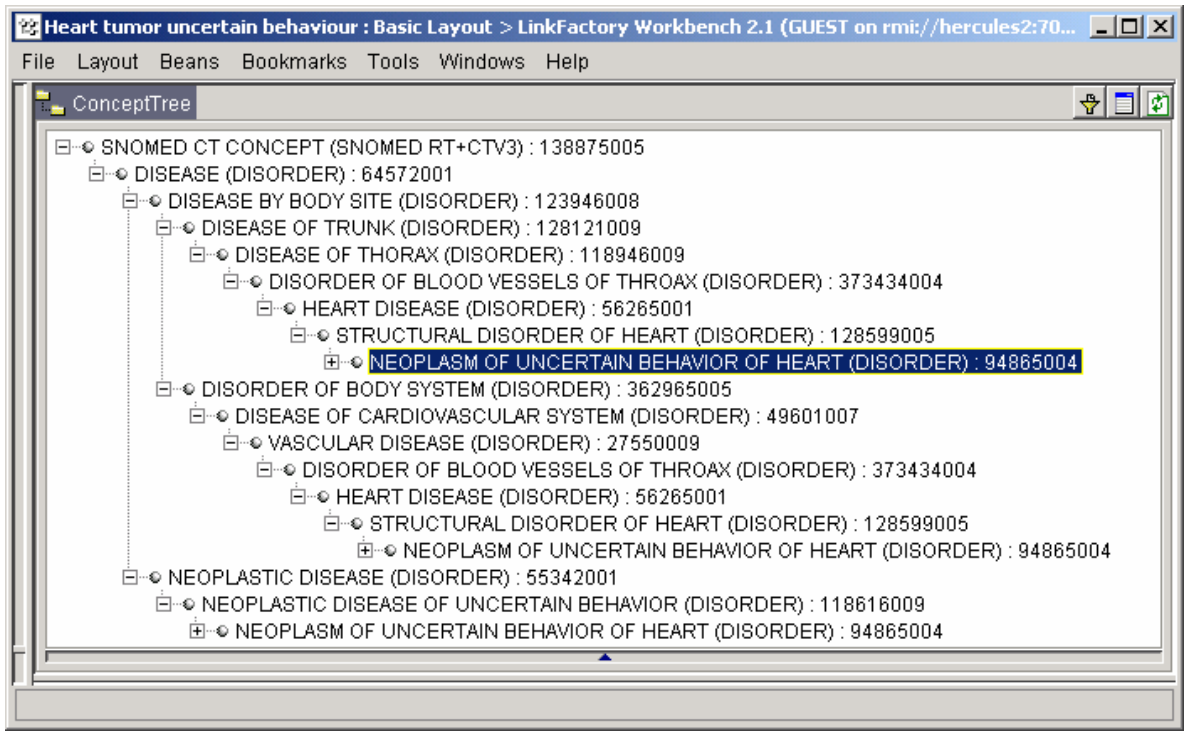

Figure 7: Inaccurate representation of the concept "94865004: neoplasm of uncertain behaviour of heart" in SNOMED-CT®.

Manually inspecting such a generated hierarchy, although it is a good way to detect inconsistencies, is a tedious task. (In examining, for example, the criteria for the generated concept "XXX-10005797XXX", one can ruminate at length on the question of what way it is different from the existing SNOMED-CT® concept "255058005: tumor of thorax".) But given the high number of generated concepts (see Table 1) this methodology is unfeasible. Rather, we sought to establish empirically which patterns in the generated hierarchies are strongly indicative of errors in such a way as to speed up this process. Examples of such patterns are:

- the presence of only one generated concept in a list of the concepts subsumed by a given concept;
- the presence of only one existing subsumed concept next to a list of generated concepts for the same subsumer;
- the presence of a pre-existing (non-generated) concept that is subsumed by a generated concept without any other additional relationships from the pre-existing concept to another one.

Formal proof of these findings, which correspond broadly to the expectations associated with ontologically well-constructed classification systems [17], has still to be given.

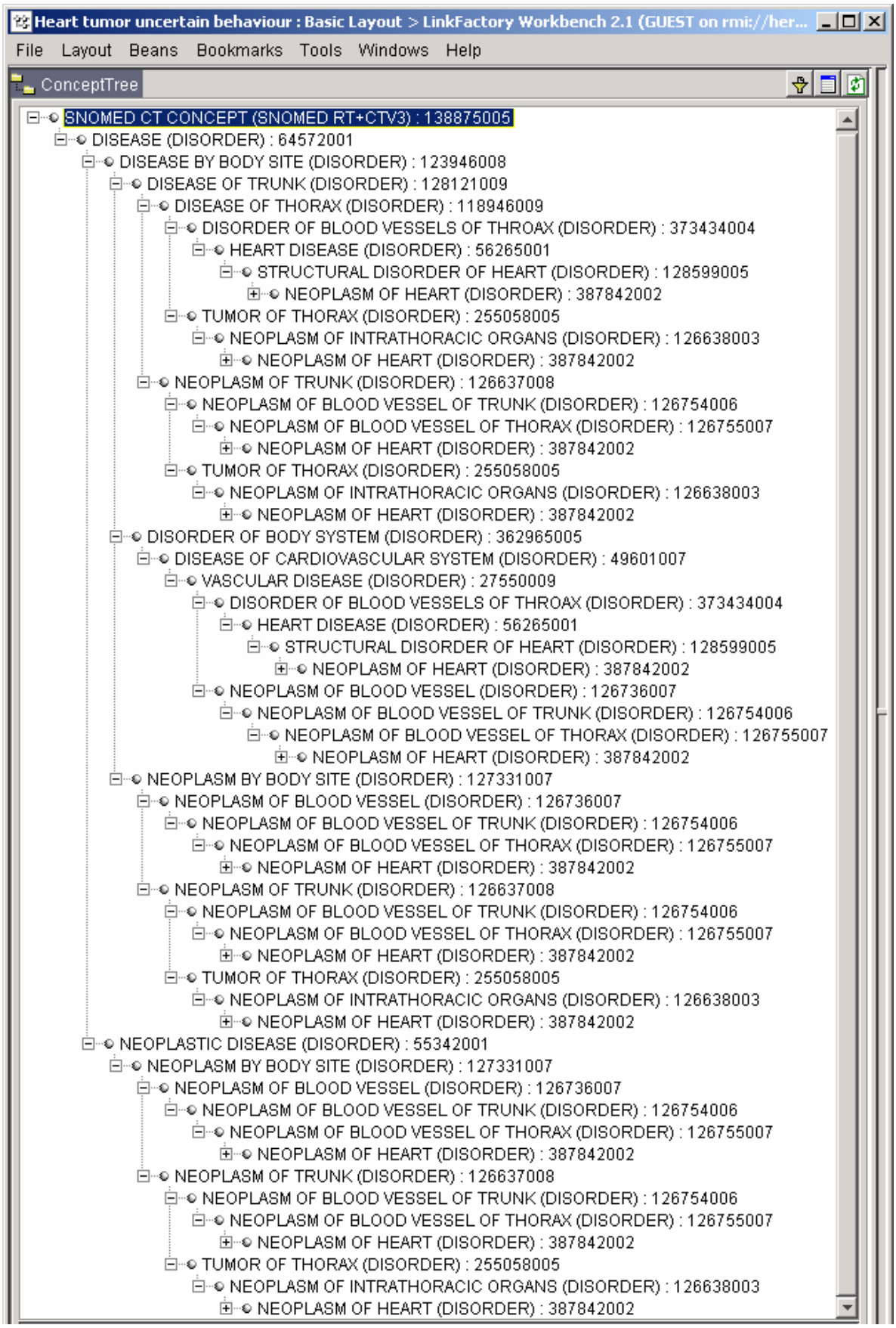

Figure 8: representation.of the concept “94865004: neoplasm of uncertain behaviour of heart” in SNOMED-CT®.

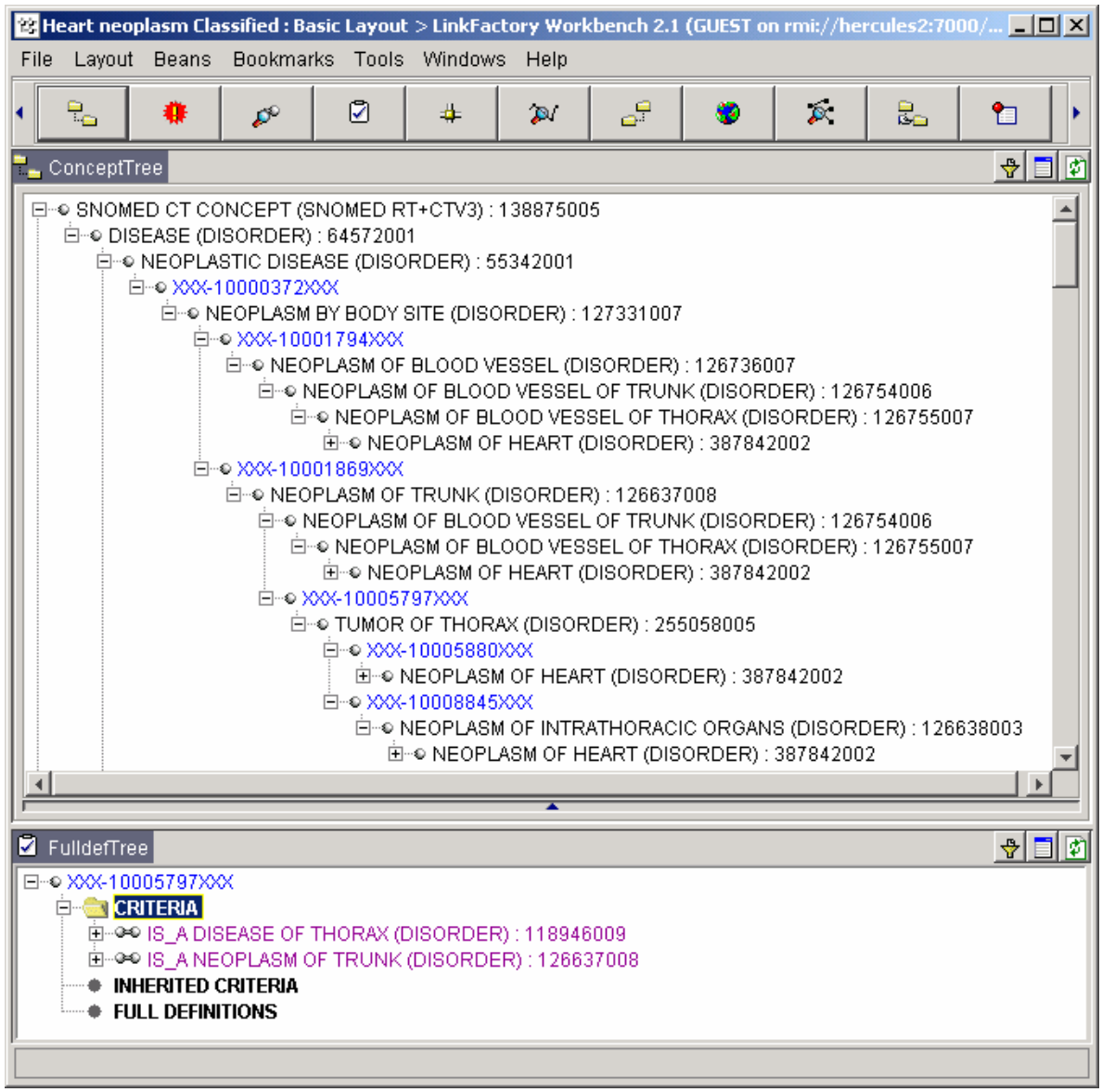

Figure 9: LinKFactory's reclassification of the SNOMED-CT® concept "387842002: neoplasm of heart" based on the relationships present in SNOMED-CT®.

2.6 SNOMED-CT®

SNOMED-CT® is a terminology system developed by the College of American Pathologists. It contains over 344,000 concepts and was formed by the merger, expansion, and restructuring of SNOMED RT® (Reference Terminology) and the United Kingdom National Health Service Clinical Terms (also known as the Read Codes).

2.7 Approach

LinKFactory®'s Ontologymatcher component used the terms of the January 2003-version of SNOMED-CT® to find related concepts in LinKBase®. The generated lists were examined manually to find superficial indicators for mistakes. Those SNOMED-CT® concepts deemed most liable to error were then subjected to a process of detailed examination that is still ongoing.

In addition, the July 2003 version of SNOMED-CT® was processed by the LinKFactory®'s classifier algorithm to find missing pre-coordinated concepts such as "abscess

of central nervous system". This was done in such a way as to utilize exclusively what is contained in that version, i.e. without taking advantage of any LinKBase® information.

Note that although the experiment involved elements of manual checking, neither the system nor the manual checkers were instructed as to the types of inconsistencies which might be detected. Thus none of the types of inconsistencies reported here was sought out a priori. Rather their detection is in each case an incidental by-product of the approach to mapping external terminology systems such as SNOMED-CT® into the LinKBase® environment.

3 Results

What follows is a very brief analysis of output generated by the TermModelling algorithm when applied to the January and July 2003 versions of SNOMED-CT®.

The analysis is by no means complete, and more work is required to yield an exhaustive list of possible inconsistencies. In what follows we assign for purposes of further reference in the discussion section of this paper an identifying number of the form "Ja-#", "Ju-#", or "Jau-#" to each reported mistake or inconsistency, indicating presence in the January, July or in both versions of the system, respectively.

3.1 Human error

Some mistakes must have their origin in inattentiveness on the part of human beings during the manual phases of the process of creating and error-checking SNOMED-CT®. The following are some of the types of errors we found under this heading:

3.1.1 Improper assignment of is-a relationships

The concept "265047004: diagnostic endoscopic examination of mediastinum NOS" is subsumed by "309830003: mediastinoscope". Thus a procedure is classified as an instrument (Jau-1). The former concept is marked as "limited", meaning that it is of limited clinical value as it is based on a classification concept or an administrative definition. Yet SNOMED-CT® still considers concepts with this status as valid for current use and as active. Another example has a procedure wrongly subsumed by a disease: thus the concept "275240008: Lichtenstien repair of inguinal hernia" is directly subsumed by "inguinal hernia" (Jau-2).

Mistakes of this type can be further divided into:

- *Improper treatment of negation*: the concept "203046000: Dupuytren's disease of palm, nodules with no contracture" is subsumed by the concept "51370006: contracture of palmar fascia" (Jau-3).
- *Improper treatment of the partial/complete distinction.* We found 9 concepts being qualified as "complete" having together 17 subsumers qualified as "partial", and respectively 6 and 11 the other way round. As an example, the concept "359940006: partial breech extraction" is subsumed by the concept "177151002: breech extraction" which in turn is subsumed by "237311001: complete breech delivery" (Jau-4). In many cases, it is the assignment of a term of the form "complete X" to a SNOMED-CT® concept with the preferred name "X", "X" to subsume "partial X" (see Figure 10 for an example).

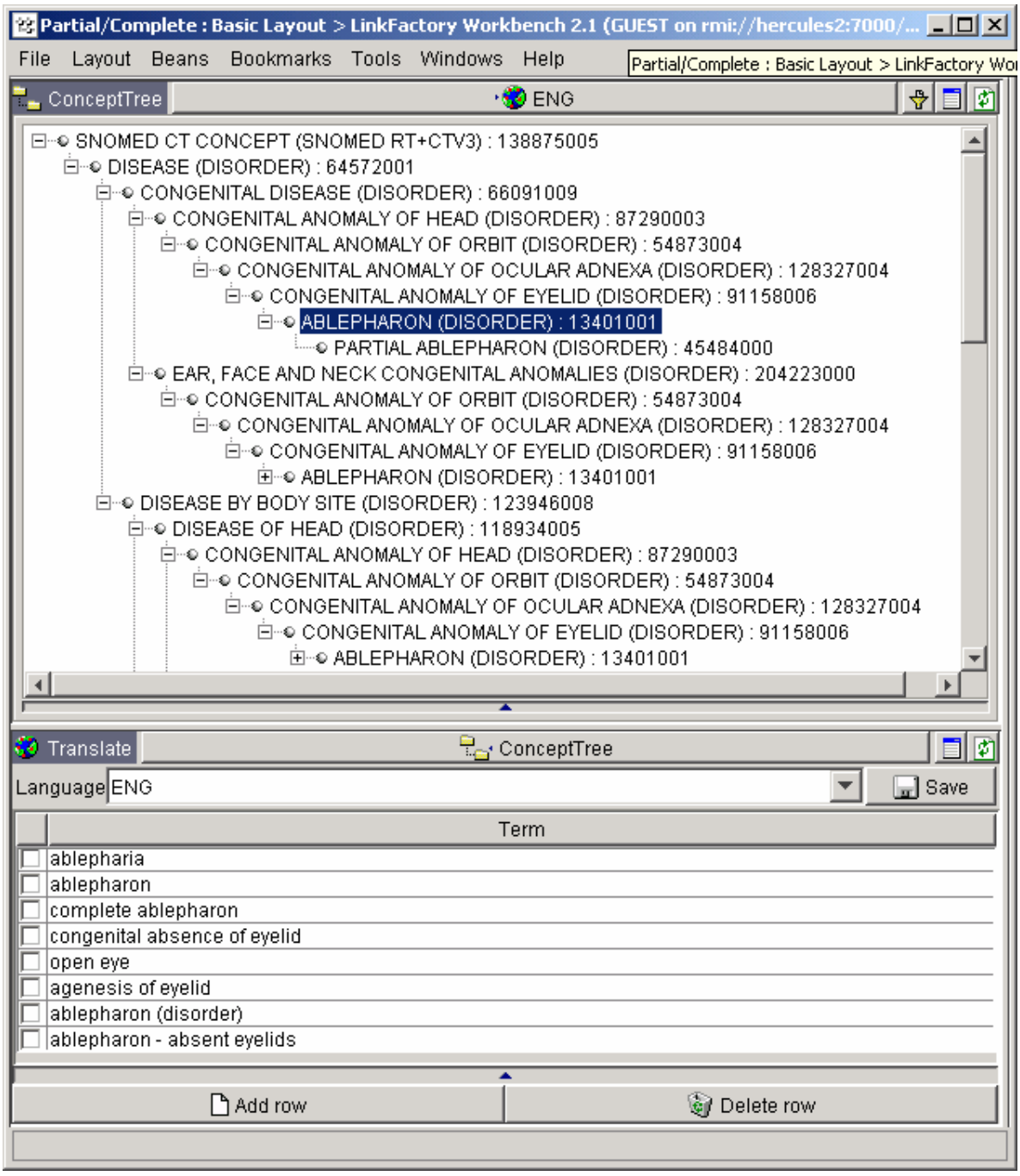

Figure 10: The SNOMED-CT® concept “13401001: Ablepharon” has as synonym the term “complete ablepharon”, while it subsumes the concept “45484000: Partial ablepharon”.

3.1.2 Improper assignment of non-is-a relationships

The concept “51370006: contracture of palmar fascia” is linked by SNOMED’s *Finding Site* relationship to the concept “plantar aponeurosis structure”. Probably as a consequence of automated classification, the concept is wrongly subsumed by “disease of foot” since “plantar aponeurosis structure” is subsumed by “structure of foot” (Jau-5). A similar phenomenon is observed in the concept “314668006: wedge fracture of vertebra”, which is subsumed by “308758008: collapse of lumbar vertebra” (Ja-6). Although the wrong subsumption is no longer present in the July version, the wrong association via *Finding Site*: “bone structure of lumbar vertebra” is still present (Jau-7). Equally the concept

“30459002: unilateral traumatic amputation of leg with complication” is classified as an “open wound of upper limb with complications” due to an erroneous association with *Finding Site*: “upper limb structure” (Jau-8).

3.2 Technology induced mistakes

A first example of this type has been referred to already above (Jau-5): wrong subsumption because of inappropriately assigned relationships. Other errors are probably induced by tools that perform lexical or string matching. We can hardly imagine that a human modeller would allow the concept “9305001: structure of labial vein” to be directly subsumed by both “vulval vein” and “structure of vein of head”. The error probably comes from an unresolved disambiguation of the word “labia” that is used for both lip (of the mouth) and vulval labia (Jau-9).

3.3 Shifts in meaning from SNOMED-RT® to -CT®

The meanings of some SNOMED-CT® concepts have been changed with respect to the corresponding SNOMED-RT© codes that have the same concept identifier and concept name. Above all, the adoption of [14]’s idea of SEP-triplets (structure-entire-part) led to a large shift in the meanings of nearly all anatomical concepts. One might argue that in RT anatomical terms such as “heart” were never supposed to mean “entire heart”, but rather always: “heart or any part thereof”; in CT this distinction has been made explicit.

Many other concepts with the same unique ID in RT and CT appear also to have changed in meaning. A notable example is the concept “45689001: femoral flebography” that in RT relates only to ultrasound and in CT involves in addition the use of a contrast medium (Jau-10). The meaning of “leg” has changed. In RT lower leg was invariably intended; in CT the situation is unclear. The concept “34939000: amputation of leg” means in RT: “amputation of lower leg” and in CT: “amputation of any part of the lower limb, including complete amputation” (Jau-11). We observed also numerous examples of inconsistent use of “leg” within CT itself: “119675006: leg repair” refers explicitly to “lower leg structure”, while “119673004: leg reconstruction” refers explicitly to “lower limb structure” (Jau-12). OntologyMatcher was able to identify these problems easily because LinKBase®, thanks to homonym processing and its mappings to UK systems such as OPCS4, is aware of differences between American and British English with respect to the meanings of “leg” and certain other words.

3.4 Redundant concepts

The TermModelling algorithm identified immediately 8746 concepts that are the seat of redundancies, that is to say cases where no apparent difference in meaning can be detected between one concept and another. (This is in reality a severe underestimation because candidate matching parameters were set very conservatively, sacrificing recall for precision.) These are all pairs or larger pluralities of terms among which differences in meaning could not be identified either conceptually or linguistically. Many of them, we believe, are the result of incomplete or inadequate integration of the Read terms into SNOMED-CT®.

An astonishing example is “210750005: traumatic unilateral amputation of foot with complication”, which co-exists in SNOMED-CT® with “63132009: unilateral traumatic amputation of foot with complication”. It seems that an incomplete modelling of the latter is at the origin of this mistake (Jau-13).

Of the same nature is the co-existence of the concepts “41191003: open fracture of head of femur” and “208539002: open fracture head, femur” (Jau-14), concepts which are modelled entirely differently but in such a way that the technology used in the development of SNOMED-CT® was not able to find the redundancy involved: the former was modeled as directly subsumed by “fracture of femur”, the latter by “fracture of neck of femur”.

Some redundancies become overt only when a larger part of the subsumption hierarchy is examined. Thus one can question to what extent “172044000: subcutaneous mastectomy for gynecomastia” is different from its immediate subsumer “59620004: mastectomy for gynecomastia” when the latter is itself immediately subsumed by “70183006: subcutaneous mastectomy” (Jau-15).

3.5 Mistakes due to lack of an ontological theory

3.5.1 Lack of sound mereotopology

It is difficult to imagine that a single object can be a proper part of two regions that are topologically disconnected. Despite this, “45684006: structure of tibial nerve” is directly subsumed by both “thigh part” and “lower leg structure”, which explicitly refer to the upper and lower parts of the lower limb, respectively (Jau-16).

3.5.2 Omission of obvious relationships

Certainly no large terminology can be expected to be complete. However, one can wonder why the concept “248182008: cracked lips” is a “301346001: finding of appearance of lip” but “80281008: cleft lip” is a “disease” and has no relation to “finding of appearance of lip”. (Jau-17) Such omissions have the consequence that many sound inferences cannot be made. As another example: “181452004: entire uterus” is *part-of* “362235009: entire female internal genitalia”, which itself is *part-of* “362236005: entire female genitourinary system”. This means, however, that SNOMED-CT® does not allow the inference to “181452004: entire uterus” *part-of* “181440006: female genital tract” since this concept has no relationships with “female internal genitalia”, and nor will it allow inferences to the effect that pregnancy involves the uterus. (Jau-18).

3.5.3 Generation of additional concepts by the classification algorithm

Table 1 shows the number of generated pre-coordinations using the LinKFactory®-classifier algorithm under the most conservative setting of minimal generation [13].

6,352 of the 17,551 newly generated pre-coordinations appear to be parents of concepts which they exclusively subsume, a phenomenon which, as we pointed out, is suggestive of mistakes in the neighbourhood of the concept in question. An example is shown in Figure 10, where one would expect the concept “exploration of disk space” to be subsumed by “exploration of spine”.

4 Discussion

SNOMED-CT®’s technical reference [15] describes the quality assurance process used in developing SNOMED-CT®. Both manual and automated procedures play a role. The mistakes discovered by the algorithms described above suggest that there is room for improvement.

Table 1: Number of generated intermediate concepts per SNOMED-CT® category

SNOMED CT Concept	original number	number added	% added
ORGANISM	24768	221	0.89
PHYSICAL OBJECT	3336	69	2.07
SPECIAL CONCEPT	130	0	0.00
CONTEXT-DEPENDENT CATEGORIES	6172	233	3.78
OBSERVABLE ENTITY	6430	33	0.51
PHYSICAL FORCE	199	3	1.51
SOCIAL CONTEXT	5120	191	3.73
SPECIMEN	936	148	15.81
EVENTS	75	0	0.00
ENVIRONMENTS AND GEOGRAPHICAL LOCATIONS	1631	5	0.31
STAGING AND SCALES	1118	0	0.00
PROCEDURE	50107	4339	8.66
BODY STRUCTURE	30737	2817	9.16
PHARMACEUTICAL / BIOLOGIC PRODUCT	13623	751	5.51
FINDING	39105	2349	6.01
ATTRIBUTE	975	2	0.21
SUBSTANCE	22062	599	2.72
DISEASE	70286	5688	8.09
QUALIFIER VALUE	7963	103	1.29
Total	284773	17551	6.16

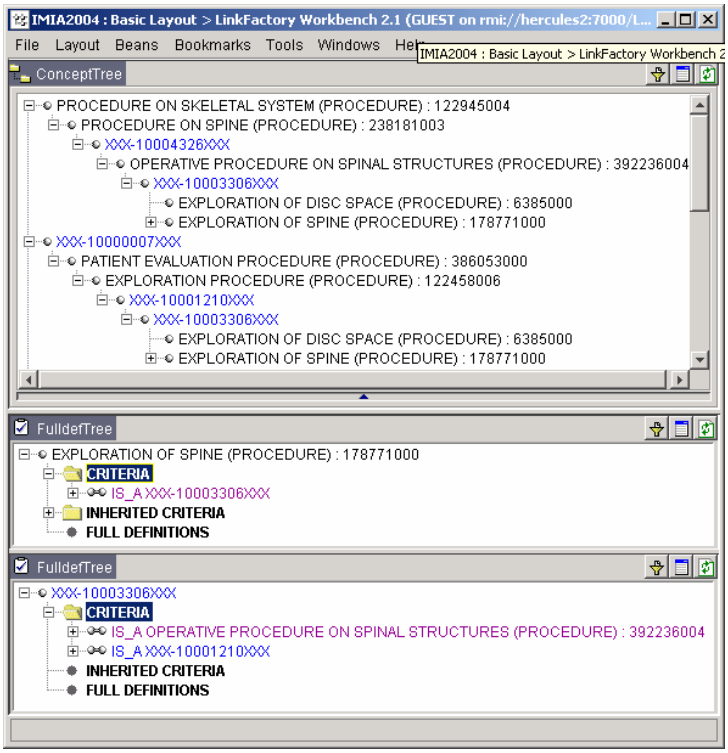

Figure 10: Algorithmically generated pre-coordinations (marked XXX) as indicators for erroneous modelling in SNOMED-CT®.

We noticed some quality improvements in the July versus January version, as the examples Jau-7 and Ja-6 demonstrate: the wrong subsumption relation with “308758008:

collapse of lumbar vertebra" has been removed, though the basic human-introduced mistake was not touched upon.

Certain mistakes could be prevented by using better logical and ontological theories implemented in more powerful ontology authoring tools. Imposing restrictions to the effect that entities of disjoint top-level categories should not stand in subsumption relations would prevent mistakes like Jau-1 and Jau-2. Enforcement of logical relations would prevent cases like Jau-3. Enforcement of mereotopological relations in accordance with an RCC-type system [4] would prevent Jau-4 and Jau-16 and lead to the flagging of cases like Jau-8 and Jau-9 for possible error. Enforcement of clear distinctions between entities and knowledge of entities would prevent cases like (Jau-17).

Features such as these are the main difference between systems such as SNOMED-CT® and LinKBase®. The latter incorporates strict ontological distinctions, for example between continuant and occurrent entities (i.e. between those entities, such as objects, conditions, functions, which continue to exist identically through time, and those entities, such as processes and events, which unfold themselves in successive temporal phases). When procedures are classified as instruments or as diseases then this reflects a conflation of high-level ontological categories which an adequate terminology system should have ways to prevent. LinKBase® also incorporates formal-ontological theories of mereology and topology (theories of completeness and incompleteness, connectedness, fiat and bona fide boundaries, etc.), and of other basic ontological notions in whose terms relations (link types) between general concepts can be rigorously defined. The presence of such theories results in a more accurate treatment of foundational relations such as *is-a* and *part-of* than is possible when such relations are left formally unanalyzed. Finally it incorporates a clear opposition between ontological notions such as object, process, organism function, and epistemological notions such as concept, finding, test result, etc.

As is argued in [16] the resultant ontologically clarified approach can be used as the basis for more rigorous but also more intuitive and thus more reliably applicable principles of manual curation than those employed in systems like SNOMED-CT® thus far.

5 Conclusion

Without doubt, a tremendous effort went into developing SNOMED-CT®. But one can wonder whether or not the appropriate tools have been used to build and subsequently test the system.

A tremendous effort is still being invested in developing the LinKBase® ontology. Thanks to the ways the LinKBase® and LinKFactory® systems have been built, however, L&C already has a powerful tool to detect inconsistencies not only in external systems but also in its own ontology.

6 References

[1] Smith B, Ceusters W. Towards industrial strength philosophy: how analytical ontology can help medical informatics. Interdisciplinary Science Reviews, 2003; 28: 106-11.

[2] Ceusters W, Martens P, Dhaen C, Terzic B. LinKBase: an Advanced Formal Ontology Management System. Interactive Tools for Knowledge Capture Workshop, KCAP-2001, October 2001, Victoria B.C., Canada (http://sern.ucalgary.ca/ksi/K-CAP/K-CAP2001/).

[3] Montyne F. The importance of formal ontologies: a case study in occupational health. OES-SEO2001 International Workshop on Open Enterprise Solutions: Systems, Experiences, and Organizations, Rome, September 2001 (http://cersi.luiss.it/oesseo2001/papers/28.pdf).

[4] Smith B. Mereotopology: a theory of parts and boundaries, Data and Knowledge Engineering 1996; 20: 287-301.

[5] Smith B, Varzi AC. Fiat and bona fide boundaries, Proc COSIT-97, Berlin: Springer. 1997: 103-119.

[6] Buekens F, Ceusters W, De Moor G. The explanatory role of events in causal and temporal reasoning in medicine. Met Inform Med 1993; 32: 274-278.
[7] Ceusters W, Buekens F, De Moor G, Waagmeester A. The distinction between linguistic and conceptual semantics in medical terminology and its implications for NLP-based knowledge acquisition. Met Inform Med 1998; 37(4/5): 327-33.
[8] Bateman JA. Ontology construction and natural language. Proc Int Workshop on Formal Ontology. Padua, Italy, 1993: 83-93.
[9] Flett A, Casella dos Santos M, Ceusters W. *Some Ontology Engineering Processes and their Supporting Technologies*, in: Gomez-Perez A, Benjamins VR (eds.) Ontologies and the Semantic Web, EKAW2002, Springer 2002, 154-165.
[10] Bittner T, Smith B. A theory of granular partitions. Foundations of Geographic Information Science, Duckham M, Goodchild MF and Worboys MF, eds., London: Taylor & Francis Books, 2003: 117-151.
[11] Grenon P, Smith B. SNAP and SPAN: Towards dynamic spatial ontology. Spatial Cognition and Computation. In press.
[12] Ceusters W, Smith B. Ontology and medical terminology: Why descriptions logics are not enough. TEPR 2003 (electronic publication): http://ontology.buffalo.edu/medo/TEPR2003.pdf
[13] Dhaen C, Ceusters W. A novel algorithm for subsumption and classification in formal ontologies. (forthcoming).
[14] Hahn U, Schulz S, Romacker M: Part-whole reasoning: a case study in medical ontology engineering. IEEE Intelligent Systems & Their Applications vol 14 nr 5, 1999: 59-67.
[15] College of American Pathologists. Snomed Clinical Terms® Technical Reference Guide, July 2003 release.
[16] Smith B, Rosse C. The role of foundational relations in the alignment of biomedical ontologies. (http://ontology.buffalo.edu/medo/isa.doc)
[17] Smith B. The logic of biological classification and the foundations of biomedical ontology, forthcoming in D Westerstahl (ed.), Invited Papers from the 10th International Conference in Logic Methodology and Philosophy of Science, Oviedo, Spain, 2003
[18] Grenon P, Smith, B. SNAP and SPAN: Towards dynamic spatial ontology", forthcoming in Spatial Cognition and Computation.
[19] Schulze-Kremer S, Smith B and Kumar A. Revising the UMLS Semantic Network. http://ontology.buffalo.edu/medo/UMLS_SN.pdf.