

The Language of Use and the Language of Knowledge in Health Language Processing

Jon D. Patrick, Copyright, 2008.

The structured use of language in medical practice has been developed over the last one hundred driven the need to systematise the description of health phenomena and to collect statistics about the health of the population. Initially the structure imposed was the consistent use of lexical items but as that grew into a very large set classifications were developed that grouped lexical items into categories following the notions of Linnaeus schemes but without the scientific rigour of definitions. The classifications served the needs of epidemiologists and students of population profiles of health needed for health administration and planning. The 1980s and 90s have seen the emergence of a new generation of usage of medical terminology from a number of directions driven by the emergence of : the use of computers to process medical records, the large scale demand for better access to case records by medical practitioners, and speedy logic processing systems.

These developments in turn have driven further strategies to deliver more sophisticated computer processing to the health desktop and in terms of the medical lexicon it has bred the formalisation of medical knowledge for logic processing in the form of transforming medical classifications into medical ontologies. However whilst this could be a good thing for supplying many productive new services it has also left a few important issues in its wake that need to be deliberated on and used to moderate the direction of current technology development and deployment.

SNOMED CT is the largest medical ontology currently available. Historically it has been developed over 40 years by the College of American Pathologists. Designed as more than a classification system it has emerged as a repository of a formal description of medical knowledge. In its latter years it has moved from a description of pathology knowledge to representing a wider range of biological and veterinary knowledge, so much so that any particular medical speciality only needs to use a very small percentage of its structures to ringfence the whole of its (SNOMED)knowledge, e.g. ICU would use about 2% of SNOMED CT. At the same time it is been bound into the logic processing technology of computer scientists at Stanford University so that it became an ontology in the computing sense of the word.

SNOMED no doubt started its life as a project to record medical terminology unlike a dictionary which aims to describe all lexical items. In time it moved into the expanded role of being a record of medical knowledge and the relationships of elements of that knowledge. The current version, SNOMED CT, is a blend of SNOMED RT and the Read Codes of the United Kingdom's National Health Service. The merging of these two systems has lead to many inconsistencies that now occupy the minds of many logicians and medical ontologists.

However one aspect of the history of SNOMED has escaped the attention of scholars. The development of SNOMED RT had the objective of recording "all" relevant medical knowledge. In this belief the system editors took the position that all the different conceptual positions that clinicians would perceive at the point of care of patients needed to be included in the terminology. This lead to a process of

capturing a large amount of the *language of use* of clinicians. In fact their process became one of converting the language of use into the *language of knowledge*. This is essentially a misguided task and has led to much of the difficulties of using SNOMED CT with today's computer technology.

USING THE LANGUAGE OF KNOWLEDGE

The language of knowledge is a description of a human knowledge that is identified by a number of features if to be used by modern computer processing such as the description logics that are used on SCT:

Statements of knowledge have canonical forms, usually defined by the description logic through which they are processed,

Atomic elements are conceptually indivisible for the purpose of the task and in the context of the rest of the knowledge in the ontology, this does not mean such items are indivisible in any other system of knowledge, or ontology. (?)

Indivisible or atomic elements have a grounding in the common meaning of the language they are rendered in,

Indivisible or atomic elements are presented in a morphologically canonical form as in dictionaries which present the lemmatised form,

Elements created by combinations of elements are not canonical but nevertheless must be composed by systematic methods,

Elements used in combinations of elements should not require morphological variations that are meaning changing and are not backwards computable to their lemmatised form,

...

THE LANGUAGE OF USE

The language of use is that form of the natural language that is used by the language community for whom the knowledge system is being prepared. Whilst the greater ambition for SCT is to include the whole world as the language community, this is not going to happen anytime soon. However there are useful intermediate goals to be achieved.

Importantly the language of use is the routine usage the community puts their language expression to work on. It is idiosyncratic, it is highly diverse at expressing the same concept, it has many local shortcuts and abbreviations and these strategies are applied preferentially to serve the community. That is, it is specifically not designed for the greater mass of humanity but rather for the members of the community who participate in the community. The knowledge the community shares is embedded within the language in use but it is not the same as the language in use. Importantly it changes in three cascading ways. Due to the productive nature of language the morphology is changed while in use, the meaning changes with the change in context, and changes in context produce changes in importance and value of knowledge elements. This is true even between medical specialities let alone between more distant professional groups and geographical groups.

SCT has been used very heavily to draw in the language of use when its original purpose as a terminology was extended to be interpreted as a description of all unity phenomena encountered at the point of care. This in essence has led to the wholesale development of pre-coordinated terms to express the great variety of

phenomena met in the daily work at the point of care, with little consideration for the nature and variety of language usage in a language community. This has led to some undesirable outcomes, now that there has been a move to use SCT for a wider range of functions.

One of the new pressures on SCT is to use its ontological form for automatic processing. This has led to serious investigations into the structure of pre-co-ordinated terms which challenges the consistency of their composition as is manifest by computational attempts to reproduce identical concept compositions by post-co-ordination. If one cannot correctly compute a pre-co-ordinated term from a set of atomic elements then there is an inconsistency in the assembly of the pre-co-ordinated term or the editorial governance of the process of creating pre-co-ordinated terms. The evidence from IHTSDO email lists indicates active discussions on how to perform these tasks and the difficulties of completing them successfully.

This essay argues that the development of pre-co-ordinated terms in SCT is the result of not distinguishing between Language in Use and Language of Knowledge, and the resultant mistaken belief that all that can be expressed at the coal face of clinical care should be included in the knowledge base. Rather than understanding the complexity of Language in Use and realising that it needs to be filtered through expert linguistic eyes to separate the meaning of terms from the expression of terms.

An example of the extreme version of the problem can be seen with the children of the class : 216177002 - Hit by aircraft, without accident to aircraft (finding)

SCTID for Child Concept	FSN for Child
216178007	Hit by aircraft, without accident to aircraft, occupant of spacecraft injured (finding)
216179004	Hit by aircraft, without accident to aircraft, occupant of military aircraft injured (finding)
216180001	Hit by aircraft, without accident to aircraft, member of crew of commercial aircraft in surface to surface transport injured (finding)
216182009	Hit by aircraft, without accident to aircraft, occupant of commercial aircraft in surface to air transport injured (finding)
216184005	Hit by aircraft, without accident to aircraft, occupant of unpowered aircraft, except parachutist, injured (finding)
216185006	Hit by aircraft, without accident to aircraft, parachutist injured (finding)
216186007	Hit by aircraft, without accident to aircraft, member of ground crew or airline employee injured (finding)
216181002	Hit by aircraft, without accident to aircraft, other occupant of commercial aircraft in surface to surface transport injured (finding)

216183004 Hit by aircraft, without accident to aircraft,
occupant of other powered aircraft injured (finding)
216187003 Hit by aircraft, without accident to aircraft,
other person injured (finding)

Some the issues with this class and its children.

1. The definition of the class is also presented in all of its children "Hit by aircraft, without accident to aircraft", then each sub-class is defined by a type of person who is injured.

2. The negated concept with the definition is no available for any ontological processing without complementary language processing, that is "without accident to aircraft".

3. Two sub-classes are differentiated by a negation clause in the text description "except parachutist".

4. One sub-class although conceptually valid is somewhat fantastical: "occupant of spacecraft injured", and so hardly warrants inclusion in a general purpose ontology.

5. The definition of these phenomena as "findings" are also an interesting categorisation. They seem intuitively to be a list of administration categories suitable for describing the circumstances of an accident. It is hard to understand how a clinician would arrive at reporting such phenomena in a clinical record apart from the secondhand reporting of the patient or their associates/family compared to findings about physiological such as :

165264009 Walks in 1 minute 0-29 metres (finding)
43005009 Shuffling gait (finding)
85763007 Blister of thigh without infection (finding)

IMPROVING THE SITUATION

The current state of SCT cannot be abandoned readily and so it is necessary to move incrementally to a system founded on an elemental representation of all composite concepts. This requires in the first instance the annotation of all elements considered to be atomic. This process does not compromise the existing pre-co-ordinated concepts but at the same time allows the technologists to more readily develop effective methods of representing the knowledge content needed at the clinical point of care. At the same time it will be necessary to develop natural language processing systems to convert the language in use at the point of care into the language of knowledge need at the point of care. This work will need to advance in tandem with the needs of particular language communities, particularly medical specialities as the costs of completing this work is intensive with the needs of high expertise.

At the same time work needs to be directed at improving the SCT ontology of its weaknesses as testified by many commentators over the last ten years. Additional work as planned needs to ensure that no more pre-co-ordinated terms are added so

the situation is not made worse.

Language technology systems have now grown to a point where they can do a lot of useful work analysing text and extracting needed content. In terms of processing clinical progress notes to automatically generate SCT codings the work is only just in its beginnings. But it is this very work that brings forward in rich contact the importance between Language in Use and Language in Knowledge as that is its exact function in this work. The role of the language technology is to translate the language in use into the language of knowledge.