From a Time Standard for Medical Informatics to a Controlled Language for Health

W. Ceusters [1, 3], F. Steurs [3], P. Zanstra [4], E. Van Der Haring [4], J. Rogers [5]

(1) Office Line Engineering NV, Het Moorhof, Hazenakkerstraat 20, B-9520 Zonnegem, Belgium. Tel: +32 53 62 95 45.

(2) Department of Informatics, Katholieke Vlaamse Hogeschool, St-Andriesstraat 2, B-2000 Antwerpen, Belgium.

(3) Department of Medical Informatics, University Hospital 5K3, De Pintelaan 185, B-9000 Gent, Belgium. E-mail: werner.ceusters@rug.ac.be.

(4) Department of Medical Informatics and Epidemiology, University of Nijmegen, PO Box 9101, 6500 HB NIJMEGEN, The Netherlands

(5) Medical Informatics Group, Department of Computer Science, University of Manchester, Oxford Rd, Manchester M13 9PL, UK

Keywords: controlled language
natural language processing
coding and classification
formal terminological systems

Summary

CEN ENV 12381 is a European Prestandard focusing on formal representation and explicit reference of temporal information in healthcare informatics and telematics. One of its merits is not just the possibility to represent natural language expressions containing time-related information in a structured way, but also to give some mechanisms on how clinical language itself can be used to convey meaning unambiguously. As such, CEN ENV 12381 introduces the notion of "controlled language use" in the domain of healthcare.

In this paper the principles behind controlled language design and use are explained. Through a detailed study of the inconsistencies and ambiguities that arise when interpreting Snomed procedure terms in the framework of the Galen-In-Use project, it is shown that most of them can be explained as a violation of sound term-formation principles. A proposal is made to develop a controlled language for health and to use it in subsequent versions of coding and classification systems. It is expected that such an endeavour will lead to a more effective application of linguistic engineering in areas such as automatic knowledge acquisition, automatic translation, and terminology validation in the domain of healthcare informatics.

1.   Time standard for Healthcare specific problems

1   Purpose and applicability of ENV 12381

In 1993, Working Group 2 of CEN/TC251[1], dealing specifically with issues such as terminology, knowledge bases semantics in healthcare informatics and telematics, recognised the need for the development of a system of concepts for time-related information. This led to the formation of Project Team PT2-017[2] of CEN/TC251 that, covered by the European Commission and the European Free Trade Association, was given the task to prepare a European Prestandard for a standard representation of time-related expressions in healthcare. The standard, accepted by the Member States as ENV 12381, had to go further than the existing time-standards [i, ii] in which only numeric date and time elements were covered.

Preceding studies of the literature carried out by the PT2-017 [iii, iv], showed that the standard should allow, as a minimum requirement, to order temporal facts in three major ways, independent of any specific ontology of time itself:

- by relating situations to a calendar,

- by relating situations to "reference" situations

- by relating events together in "before- and after-" chains.

The main reason for this threefold organisation is that our everyday temporal discourse (not necessarily limited to the domain of healthcare) contains a variety of expressions that only with a certain artificiality can be regimented into a uniform style of analysis.

The standard provides a set of principles for syntactic and semantic representation that allow the comparability of specific ontologies on time, and the exchange of time-related information that is expressed explicitly. As such, ENV 12381 provides a standardised way of representing time-related expressions, such that all kinds of questions about the temporal organisation of events (or whatever similar constructs are called or stand for in specific ontologies) can be answered on the basis of the information available. However, ENV 12381 does not provide a means to interpret implicit time-related information. In an expression such as "diabetes since childhood", "since childhood" is an explicit temporal reference for the diabetes, but the implicit information what "childhood" might mean (e.g. starting at the age of 2 years ?), is not addressed. Interpretation of the source information is the task of the provider of information itself.

---

1 CEN/TC251 is the Technical Committee of the European Committee for Standardisation (CEN), dealing with Medical Informatics.

2 CEN/TC251/PT2-017 was composed of an international pluridisciplinar team of experts in the domain of medical informatics, medicine, logic, philosophy and computational linguistics and had the following members: Ceusters W (PT-leader, Belgium), Buekens F (Belgium), Bernauer J (Germany), De Keyser L (Belgium), Surján G (Hungary), Rossi-Mori A. (Italy), Olesen H. (observer, Denmark), Rector A (reviewer, UK),

Of course, the language provided by ENV 12381 has enough expressive power to allow a specific provider of information to state explicitly what his understanding is of "childhood".

2   Making time-related information explicit and unambiguous

ENV 12381 allows information providers to express time-related information in such a way that the intended meaning can be unambiguously understood by a receiver. This of course requires the use of a "restricted", regimented model or language, allowing the disambiguation of many time-related expressions uttered in natural language. The model (language) presented in ENV 12381 is restricted enough to allow such disambiguation for time-related expressions in "traditional" medical language, but is not expressive enough to account for all time-related linguistic phenomena that can be encountered in natural language.

Figure 1 shows the representation of the sentence "*sudden loss of smell following head injury around 3 o'clock*" according to the provisions of ENV 12381.

(**EVENT**("sudden loss of smell")
    (*has-occurrence* AFTER
        TP(**EVENT**("head injury")
           (*has-occurrence* AT **TP**("around 3 o'clock"))))))

Figure 1: ENV 12381 representation of the sentence "*sudden loss of smell following head injury around 3 o'clock*".

This representation makes clear **by explicit indication of the sender of the information,** and not as a post hoc interpretation by the receiver (see [v] for a more thorough description and Table 1 for the ENV 12381 definitions of the underlined terms):

1. that the sentence "*sudden loss of smell following head injury around 3 o'clock*" is a predication containing the propositional clauses "*sudden loss of smell*" and "*head injury*", both representing an event, and the absolute temporal expression "*around 3 o'clock*"[3], representing a time point, hence categorised as time point expression.

2. that the predication contains implicitly a basic temporal link and a temporal comparator referring to the fact that the *head injury* occurred at the specified time point *around 3 o'clock*. This implicit link is made explicit in the representation of the predication according to the provisions of ENV 12381 by using "**has_occurrence AT**".

---

[3] ENV 12381 recommends that existing standards on time shall be followed. As such, "3 o'clock" will in legacy information systems have to be written down in the standard ISO format for date-time values.

3. that the head injury event occurring at the specified time (and grammaticalised in the sentence as *head injury around 3 o'clock*) is used itself as a <u>time point expression</u> from which there is a <u>basic temporal link</u> with the <u>propositional clause</u> *sudden loss of smell*, grammaticalised in the sentence by the preposition *following*, and made explicit in the representation by using the notational convention "**has_occurrence AFTER**".

4. that there is a non-specified amount of time (possibly even considerable !) between the head injury and the loss of smell. If the loss of smell would have appeared immediately after the injury, the sender had to use **has_occurrence *AT*** instead of **has_occurrence *AFTER***.

5. that the duration of the injury, as well as the duration of the loss of smell are considered irrelevant, otherwise these <u>situations</u> would have to be annotated as being <u>time intervals</u> instead of <u>time points</u>.

Term	ENV 12381 definition
absolute temporal expression	: <u>temporal expression</u> whose exact meaning in a given context can directly be derived from the <u>temporal expression</u> itself
basic temporal link	: <u>temporal link</u> specifying purely time-related information
event	: <u>situation</u> considered to occur at a <u>time point</u>
predication	: representation of a <u>situation</u> in a language
propositional clause	: component of a <u>predication</u> to which <u>temporal references</u> implicitly or explicitly refer
situation	: phenomenon occurring (or having the potential to occur) at or over a time in a given world context
temporal comparator	: specifier of the temporal relation expressed by the <u>temporal link</u> between the <u>propositional clause</u> and a <u>temporal expression</u>
temporal expression	: component of a <u>temporal reference</u> specifying a <u>time point</u>, a <u>time interval</u> or any allowed combination of <u>time points</u> and <u>time intervals.</u>
temporal link	: component of a <u>temporal reference</u> capturing the semantic relation in a <u>predication</u> between the <u>propositional clause</u> and the <u>temporal expression</u>
temporal reference	: component of a <u>predication</u> representing information related to time
time interval	: portion of time of which the duration in a given context is considered to be significant and relevant
time point	: portion of time of which the duration in a given context is considered to be insignificant or irrelevant
time point expression	: <u>temporal expression</u> denoting a <u>time point</u>

Table 1: ENV 12381 definitions for the terms used in the descriptive analysis of Figure 1. (Underlined words in the definitions are used in a precise meaning and are defined themselves in this table.)

3   From natural language to a "restricted", regimented language

It is obvious that when only the "raw" sentence of Figure 1 is used to transmit the *loss of smell* finding, a lot of the information that is made explicit through the ENV 12381 annotations would be less self-explanatory. At the other hand, it would not be acceptable for nurses or physicians to write clinical findings or other healthcare information in the format specified by ENV 12381. The required artificiality to regiment time-related expressions into a uniform style of analysis is an unacceptable

feature in man-man communication, but is perfectly acceptable for machine-machine communication, and to a certain extend also for man-machine communication. Indeed, ENV 12381 is supposed to be used only in a perspective of machine-machine and man-machine communication. However, the same ideas behind this regimentation can be used with respect to natural language. This brings us in the domain of controlled language.

2. Controlled language

1 Definitions

A <u>controlled language</u> is a precisely **defined** subset of a natural language, on the one hand constrained in its lexicon, grammar and style, and on the other hand possibly extended by domain-specific terminology and grammatical constructions. As such it differs from a <u>sublanguage</u> which is a **natural** language, be it used in a particular semantic domain and for a specific purpose. Both controlled languages and sublanguages have in common that they differ from "general" natural languages by being restrictive, deviant and preferential with respect to vocabulary, syntax, semantics and pragmatics [vi, vii, viii, ix]. The main difference is however that sublanguages evolve naturally within a community while controlled languages are artificial adaptations of a language that are tried to be kept as natural as possible. In this respect they differ also from true artificial languages such as Esperanto (created for true communication over linguistic borders [x]) or Klingon (alien language in the Star Trek tv series [xi, xii]). And finally, controlled languages are not to be mixed up with <u>controlled vocabularies</u> that are (possibly hierarchically) structured sets of certified terms that are verbal canonical representations of concepts. The aspect of control in a controlled vocabulary is related to the position of a specific term in the vocabulary as a whole, the choice of a particular term as canonical form, and the requirement that only terms from within the vocabulary are to be used in an application. The terms themselves are however not written in a controlled language.

The motivation for the use of controlled language is that it makes all aspects of text manipulation (both human and computational) easier. By eliminating sources of ambiguity and by prescribing stylistic rules, controlled languages aim for improved readability, understandability, maintainability and easier computational processing such as for information retrieval, automated translation or language understanding [xiii, xiv]. The history of controlled languages goes back to 1930 when "Basic English" was created by Charles Kay Ogden [xv]. The main idea was to create a variant of English that could easily by learned, and that would allow the writing of legal documents that are easy to understand. The breakthrough of controlled languages was however the birth of *Simplified English* that now is generally used in aircraft documentation [xvi] and led to the development of other controlled English variants in various industries [xvii, xviii].

2 ENV 12381 as the basis for a controlled language for time-related expressions in healthcare

In this paragraph we merely describe what it would mean to develop a controlled language for time-related information according to the ENV 12381 provisions without (yet) taking any position on whether this is advisable or not. Motivations for such an approach in a different context is to be found in the section 2.

ENV 12381 lends itself perfectly to such an endeavour due to the principle of explicitness that has been maintained throughout the entire document. What needs to be done is to bring the proposed representation (see Figure 1 for an example) back from the level of a formalism to the level of a language, however without loosing any of the information, or without introducing ambiguities.

There are many possibilities to achieve this goal. As Dodd said: "*Somewhere between ridiculous pedantry and erroneous formulation there presumably exists a reasonable precise way of specifying a problem in English*" [xix]. The sentence of Figure 1 could be rephrased (not exclusively) as :

(ex. 1) *event of sudden loss of smell after event of head injury at around 3 o'clock*

or

(ex. 2) *event of a sudden loss of smell after event of a head injury at around 3 o'clock*

For both possibilities, some principles of controlled languages have been applied in order to claim the unambiguous interpretation of the sentence.

First, there is the restrictive mode of the lexicon. Words or word groups such as "*event of*", "*at*" and "*after*" are used with a precise meaning. A requirement might be that these words may be used in only one meaning (in this case as defined by ENV 12381), excluding f.i. the use of "at" and "after" for locatives. This (severe) restriction reduces the expressiveness of the controlled language, but at the other hand facilitates automatic parsing afterwards by keeping it context-free. The sentence of (ex. 2) follows the Simplified English writing rule that noun phrases should have an article unless the intended meaning is altered by doing so [xvi]. This rule exemplifies the deviant mode of the controlled language at grammatical level. The possible syntactic ambiguity in both sentences - i.e. whether the head injury occurred at around 3 o'clock or the sudden loss of smell - is resolved by requiring that in absence of preferences for certain prepositions to be attached to possible head clauses, the principle of right association ("late closure") must be applied in the controlled language [xx]. This is an example of the restrictive mode of the controlled language with respect to syntax.

3. Would medical nomenclatures and thesauri benefit from controlled language use ?

In order to answer this question we conducted a detailed study of the language used in SNOMED International (V3.2), more precisely of the procedure axis [xxi]. The study is carried out within the Galen-In-Use project to find out whether natural language texts can be used to (semi-)automatically populate formal terminological systems.

1   The Galen-In-Use project

The purpose of the GALEN project is to develop language independent concept representation systems as the foundations for the next generation of multilingual coding systems [xxii]. At the heart of the project is the development of a common reference model for medical concepts (CRM) supported by a formal language for medical concept representation (GRAIL) [xxiii]. A particular characteristic of the approach is the clear separation of the pure conceptual knowledge from other types of knowledge, including linguistic knowledge [xxiv], in order to arrive in the future to application-independent medical terminologies [xxv].

In the GALEN-IN-USE project, various centres are collaborating to build an exhaustive model for surgical procedures [xxvi]. An initial hypothesis was that this modelling work could be speeded up by semi-automatic processes relying on natural language processing techniques. The MultiTALE syntactic semantic tagger was used for this purpose. It was originally designed to analyse full text neurosurgical procedure reports, and to extract all the surgical deeds in the format of the CEN ENV 1828:1995 standard "Structure for the classification of surgical procedures." [xxvii, xxviii]. First attempts gave acceptable results with respect to the generation of formal representations, be it however with considerable re-engineering efforts, and not an immediate gain in time[4] [xxix]. Next it was decided to take more advantage of the manual modelling work being carried out by building a machine learning system for natural language analysis. The first step in this approach was the development of the CASSANDRA tagging technique in order to re-introduce in an explicit and formal way links between the semantic model and the surface language [xxx]. At the same time, the technique is used to annotate parallel corpora of medical texts in different languages for marking similarities independent of a specific grammar formalism [xxxi].

---

[4] What is the advantage obtained when building a system to perform a task automatically takes as long as doing the original task manually ?

2 What makes SNOMED International difficult to use for automated knowledge acquisition ?

The following analysis is based on the general principles of controlled language development. In particular (some of) the "violation" of these controlled language principles by the language used in Snomed International will be discussed [5].

.1 Inappropriate use of synonymy

Many controlled languages require that no synonyms should be used at all as a first step to reduce the number of words in the language. For Snomed as such, it would not be acceptable that there wouldn't be any synonyms because one of its objectives is precisely to bring terms found in clinical narrative back to a canonical form expressed in language. One could however argue that internally in Snomed, terms that are characterised by means of the "02"-class field as being synonyms, should not appear in other terms that have the "01" class field status as in (ex. 3).

(ex. 3)   P1-AC902   "01"   closure of fistula of <u>ear drum</u>
                P1-AC902   "02"   closure of fistula of <u>tympanic membrane</u>
      where:
                T-AB320   "01"   <u>tympanic membrane</u>, NOS
                T-AB320   "02"   <u>ear drum</u>, NOS

.2 Misleading use of homonymy

Controlled languages tend to reduce homonymy as much as possible. Regulations with respect to this do not cover only pure semantic issues but take into account the syntactic ambiguities related to the various parts of speech that a token can have. The word "round" e.g. can be assigned to 5 different categories (noun, verb, preposition, adjective and adverb) and in total to 40 different meanings. Controlled languages might allow less meanings and less possible parts of speech, e.g. that "round" as preposition should be replaced by "around". Snomed contains a lot of terms where these principles are violated.

(ex. 4)   P1-91262   "01"   drainage of <u>ventricle</u> by aspiration
                P1-31884   "01"   implantation of mammary artery into <u>ventricle</u>
      where:
                T-32400   "01"   <u>ventricle</u>, NOS
                T-32400   "02"   cardiac ventricle
                T-A1600   "01"   cerebral ventricle, NOS

---

[5] a) We are perfectly aware that SNOMED International has not been designed to be used in such a setting, but will argue that it (and any other similar system) would benefit in going from a sublanguage to a controlled sublanguage.

    b) We concentrate here on problems related to language use and term formation in Snomed, and not on the structural inconsistencies of the concept system behind the terms, a topic already addressed in the literature.

The fact that in the term "drainage of ventricle by aspiration" a <u>cerebral</u> ventricle is referred to and not a <u>cardiac</u> ventricle cannot be deduced from the language itself, but only from the codes. Even more surprisingly, if the coding conventions for clinical narrative are used as prescribed by Snomed itself [6], we end up with an erroneous result as in "drainage of ventricle by aspiration" the term "ventricle" should be coded as T-32400.

Here is another example of ambiguous use of homonymy:

(ex 5)   P1-171A0   "01"   drainage of finger abscess, <u>simple</u>
              P1-171A1   "01"   drainage of finger abscess, *complicated*
       whereas:
              P1-65110   "01"   <u>simple</u> drainage of lymph node abscess
              P1-65112   "01"   *extensive* drainage of lymph node abscess
       and: G-A537   "01"   <u>simple</u>

The modifier "simple" is used in two different meanings in the procedure axis while there is only one entry (with unspecified meaning !) in the modifier axis. There is nothing in each of the terms in which appears the word "simple" that can tell us what meaning is understood. Only contrastive studies as done in this example, can help us figuring it out. Hence this is a serious violation of general term formation principles that require terms to be understandable independent of context [xxxii].

Very often additional complexity is introduced by using homonyms as quasi-synonyms:

- "hip" for "hip region" or "hip joint"
- "anastomosis" for both the "act of making an anastomosis" and the "result" of the act. There are numerous other examples of this kind
- "graft" for the act of grafting and the "instrument participant role" (see further) in such an act
- "bone" for a real bone, such as the humerus, or for the material of which bones are made

Though the notion of homonymy is commonly used with respect to verbs, nouns or adjectives, the same phenomenon can be encountered with prepositions. Table 3 and Table 2 give a (non-exhaustive !) overview of the various meanings (where possible expressed as conventional *thematic roles* or θ-*roles* [xxxiii, xxxiv] for predications involving events [7], otherwise by using an *object-relation-system* more close to healthcare) of the preposition "of" encountered in the Snomed International procedure axis[8].

---

[6] Parse a text by taking the longest possible terms within SNOMED.

[7] The words "event" and "predication" are used here in linguistic sense and not in the literal CEN ENV 12381 meaning.

[8] Whether or not prepositions such as "of" carry meaning, is in pure linguistic environments a matter of discussion, while in computational linguistics it depends on the linguistic theor(y)(ies) underlying specific applications. This discussion falls outside the scope of this paper.

1	P1-10880	change of length **of** tendon
	→	initiates the *object-relation* "*as property of*" whereas the actual property (in this case the "length") precedes the preposition.
2	P1-1830A	debridement of open fracture **of** leg
	→	initiates a specialisation of the "pathologic undergoer" object-relation in the sense that not the leg is fractured, but a bone <u>inside</u> the leg.
3	P1-16997	open reduction of separated epiphysis **of** humerus
	→	initiates the *partitative object-relation*
4	P1-52820	suture of laceration **of** tongue
	→	initiates the *object-relation* of "*pathologic undergoer*" (having embedded in it both notions of location and undergoer)
5	P1-10935	repair of fascia with graft **of** muscle
	→	initiates the *object-relation* of "<u>*ingrediency*</u>"
6	P1-18B02	suture of ligament **of** lower extremity, NOS
	→	initiates the *object-relation* of "<u>*internal location*</u>"
7	P1-10508	implantation of prosthesis or prosthetic device **of** joint, NOS
	→	initiates the *participant role "internal benefactive"* (as contrasted with the real "benefactive" which is the patient).

Table 2: Thematic role or object-relation initiation by the preposition "of" connected to non-events in the Snomed International procedure axis.

1	P1-0C010	lysis **of** adhesions, NOS
	→	initiates the *participant recipient role* of "<u>*undergoer*</u>" [9] in a procedure without the notion of directed movement
2	P1-05020	injection **of** prophylactic substance, NOS
	→	initiates the *participant spatial role* of "<u>*theme*</u>" in a procedure with notion of directed movement
3	P1-10540	injection **of** ganglion cyst
	→	initiates the *participant spatial role* of "<u>*goal*</u>" in a procedure with notion of directed movement (compare with 2)
4	P1-19321	synovectomy **of** ankle
	→	initiates the *participant spatial role* of "<u>*source*</u>"
5	P1-16A00	arthroplasty **of** elbow, NOS
	→	initiates a specialisation of the participant role of undergoer. It is not the elbow that undergoes something, but the joint within the elbow. One could argue to assign here the thematic participant role of "experiencer" though this usually is reserved for animate things.
6	P1-11110	osteotomy **of** mandible, NOS
	→	initiates the participant role of undergoer, but can only be used when a property of the undergoer is expressed in the event. In this case : that the mandible "is-a" bone allows <u>osteotomy</u> to be used in this configuration (pleonastic undergoer).

Table 3: Thematic role or object-relation initiation by the preposition "of" connected to events in the Snomed International procedure axis.

Many of the role or relation assignments in Table 2 and Table 3 are debatable as a consequence of the semantic ambiguity in other constituents of the terms taken as examples.

---

[9] In traditional thematic role terminology, "patient" is used instead of "undergoer". The "undergoer" as stated here is not necessarily the "undergoer" macro-role as defined by Foley and van Valin.

Example 5 of Table 2 presents an extremely ambiguous Snomed term as the ingredient relation can only be opted for if the word "with" is considered to be a preposition that initiates "graft" as the participant role of instrument for the repair event. Another interpretation is to see "with" as a conjunctor, the word "graft" being a nominalisation of grafting, in which case "of" initiates the participant role of "undergoer". In example 6 of Table 2, it is possible to assign also the object-relation *part-of* that can be seen as a specialisation of the internal location for non-events.

.3 Complexity of noun groups or noun clusters

A typical example of this problem is (ex. 6) where without an extensive amount of pragmatic knowledge a large number of possible interpretations could be generated.

(ex. 6)   P1-17A26   Tenodesis for proximal interphalangeal finger joint stabilization.

In the light of machine learning for natural language processing, there is nothing in this term that can tell us that "finger joint" is to be seen as a compound and hence "proximal" and "interphalangeal" refer to "finger joint". One could presume erroneously that there are "proximal" and "distal" fingers, that fingers could be located "interphalangeal" or that the stabilization was only carried out on the proximal part of the joint. Breaking up such noun clusters would make understanding far more easier. Simplified English f.i. requires noun-clusters to be limited to three units. In addition parts of units that modify each other should be hyphenated [xvi]. As such, the Snomed term would have to be rewritten as "*tenodesis for stabilization of proximal interphalangeal finger-joint*".

It is strictly forbidden to use prepositional phrases inside compound structures such as in:

(ex. 7) P1-17440   lateral fasciotomy with annular ligament of finger resection.

These examples show that with respect to nominalisations, multilingual comparative studies in medical terminology should be conducted. Extensive nominalisation is a typical characteristic of technical English, but (at least in the medical domain) it tends to introduce ambiguities. It is interesting to notice that in Dutch such long nominalisations would never occur, and hence, less ambiguities are present.

.4 Various co-ordinated constructions

Controlled languages tend to reduce grammatical complexity by disallowing many of the co-ordinated constructions that are found in general language [xiv]. Indeed, the scope of co-ordination and how it relates to other attachments, in particular prepositional phrase attachment, can be difficult to determine (see ex. 8 and 9 [10]).

---

[10] We only give the "medically sensible" readings and not all possible syntactic combinations.

(ex. 8) P1-188A7   epiphyseal arrest by stapling, combined, proximal and distal tibia and fibula and distal femur

which can read: ... ((proximal and distal) tibia) and fibula and distal femur
or: ... ((proximal and distal) (tibia and fibula)) and distal femur

(ex. 9) P1-18932   primary repair of torn ligament and capsule of knee, collateral
which can read: ... repair of ((torn ligament) and (capsule)) of knee ...
                       ... repair of (torn (ligament and capsule)) of knee ...

Additional complexity is introduced by the different "meanings" of the word "and" (Table 4).

In example 1 of Table 4, "and" is used instead of "or" or "and/or", a direct implication of the Snomed procedure axis as a concept system with hierarchical structure. In example 2, "add" functions as "additive" in the sense that both radius and ulna are involved in the procedure. There is not much in the term itself that can tell us so, but rather the existence of two other codes : P1-16026 for "diagnostic procedure on radius, NOS" and P1-16028 for "diagnostic procedure on ulna, NOS". If in example 2 also "and/or" is to be understood, then we would expect the two other terms being classified as children of P1-16024 and not as siblings.

In example 3 of Table 4 "and" is used in the connective-restrictive sense: there are not 2 tumours removed (one from the pelvis and one from the hip area), but only one, namely from the area of pelvis and hip.

In example 4 of Table 4, "and" is again used in additive sense, but with an additional notion of temporal relativity. Whereas in the previous examples the nouns connected by "and" could switch place without altering the meaning of the term, this is not the case for example 4 (otherwise one would aspirate what previously has been injected). "And" is here used for "and then".

1	P1-10500	musculoskeletal system : injections **and** implantations
2	P1-16024	diagnostic procedure on radius **and** ulna, NOS
3	P1-14314	excision of tumor of pelvis **and** hip area, subcutaneous
4	P1-10558	aspiration **and** injection for treatment of bone cyst

Table 4: Various uses of "and" in Snomed

The correct interpretations of the examples in Table 4 is not too problematic for a human reader, and can with the appropriate computational linguistic techniques also be discovered by a natural language analyser. Frequently however, terms are constructed in an extremely misleading way, especially when commas[11] are involved, as in (ex. 10).

---

[11] Commas are in Snomed used as "synonym" for "and", "or" or "and/or", as well as for readability (comparable with first-level bracketing) and in post-modification by adjectives, very often with long-distance dependency.

(ex. 10) P1-17112   decompression fasciotomy of wrist, flexor and extensor compartment

In this example, the comma is not used for "and" as usually is done when several entities are summed up in a co-ordination, but as a readability marker in post-modification, to indicate that not the wrist as such is operated upon, but a specific part of it, more precisely the flexor and extensor compartment**s**. What is problematic in this sentence, is that it is not easy to determine whether or not "and" is used in additive or connective sense. Only with a fair knowledge of anatomy, one can know that there is a flexor compartment and an extensor compartment. It would have been less misleading if in this term a plural was used, or even better [12], no ellipsis at all.

.5   Long-distance dependency and cross-modification

There are numerous examples in Snomed where (adjectival) modification of a sentence constituent is done further down in the sentence, often at long-distance. In (ex. 11), '"collateral" modifies "ligament", and not "knee" [13]. In (ex. 12), "single" modifies (presumably [14]) "dislocation", while "with uncomplicated soft tissue closure" is to be attached to "reduction".

(ex. 11) P1-18933   primary repair of torn ligament of knee, collateral

(ex. 12) P1-17834   reduction of open carpometacarpal dislocation, except Bennett fracture, single with uncomplicated soft tissue closure

4.   Recommendations for controlled language usage in healthcare

The many examples in section 2 show that the clarity of terms in Snomed International can and should be improved dramatically [15]. A first step would be to follow the relevant standards in the field [xxxv, xxxvi, xxxvii, xxxviii], an interesting and harmonised view of some of them being given in [xxxix]. None of these standards address however the issue of structural term formation in a sufficient detailed way, and certainly not in the scope of natural language understanding as terminologies so far have only been designed to be used by humans and not by machines.

The way in which term formation currently is handled in systems such as Snomed makes them difficult to translate, difficult to understand by novice users or medical students that don't have the necessary pragmatic knowledge to resolve linguistic ambiguities, and certainly nearly inadequate as machine readable knowledge

---

[12] Otherwise one could still assume that there are several extensor compartments and several flexor compartments in the wrist.

[13] One could argue that the sentence is even completely wrong as "collateral" is part of the compound noun "collateral ligament" and should not be detached from "ligament".

[14] It could even be a postmodification of "reduction".

[15] This holds for most other thesauri and nomenclatures in healthcare as well.

repositories. We are convinced that some simple term writing conventions as outlined below can improve the overall usability with only a small cost [16].

1) <u>Avoid using the same word in different meanings and with different parts of speech</u>. Use f.i. "suture" only for the wire and not for the deed of suturing or the anatomical sutures [17].

      P1-B1305   removal of **suture** of thorax
      * P1-91272   Ventricular puncture through **suture** without injection
            → ventricular puncture through *anatomic-suture* without injection
      * P1-91870   **Suture** of cerebral meninges
            → *suturing* of cerebral meninges

2) <u>Use prepositions in such a way that they (preferably uniquely) identify the thematic role or object-relation</u> (see .2):

      P1-67394   aspiration **of** bone marrow from donor for transplant
      * P1-10120   aspiration **of** joint, NOS
            → aspiration *from* joint, NOS

      P1-38C14   single injection **of** sclerosing solution for spider veins of face
      * P1-10542   injection ***of*** ligament, NOS
            → injection *into* ligament, NOS

Whereas the previous examples highlight the ambiguous use of the preposition "of", the next examples show the same effect for the preposition "for" [18]. At the same time, recommendation 3 is applied to remove the ambiguity.

      * P1-75144   External urethrotomy **for** perineal urethra
            → External urethrotomy *for reason of* perineal urethra
      * P1-75144   Poncet operation **for** perineal urethrostomy
            → Poncet operation *with purpose of* perineal urethrostomy

3) <u>Use double or triple prepositions for</u> [19] <u>expressing meaning with greater precision</u>

      * P1-7A510   insertion of valve **in** vas deferens
            → insertion of valve *into* vas deferens

      * P1-03177   incision and removal **by** magnet
            → incision and removal *(by) using* (a) magnet [20]

---

[16] a) In the examples given below, terms that should be replaced or rewritten are marked with an asterisk.
  b) The recommendations are not yet to be seen as exhaustive.

[17] Of course, it is possible to make another choice of what should be kept, and what should be changed.

[18] Notice that - anecdotally - the double use of "for" occurs in synonymous terms !

[19] with purpose of

[20] Various alternatives can be proposed whether or not recommendation 1 should be applied as well.

* P1-08416   closure **by** buckling
→ closure *realised by* buckling

4) Maintain normal word order as indicated by the general grammar of the language in which the terms are expressed.

* P1-18933   primary repair of torn ligament of knee, collateral
→ primary repair of torn collateral ligament of knee
* P1-10501   replacement of prosthesis of extremity, bioelectric or cineplastic
→ replacement of biolectric or cineplastic prosthesis of extremity.

5) Limit term length to what (at least) a skilled human reader can easily understand

* P1-11823   open treatment of craniofacial separation, Lefort III type with wiring and/or local fixation, complicated, fixation by head cap, halo device, multiple surgical approaches, internal fixation, and/or wiring of teeth

6) Use co-ordination with extreme care:

6a) *Use it only for "nearest neighbours" belonging to the same syntactic or semantic class and that share all modifications expressed in the term*

P1-00052   "01"   incision **and** reexploration for second look
P1-03052   "01"   radical excision with en bloc resection of regional organs **and** tissues

* P1-00052   "02"   incision **and** reexploration of recent operation
→ incision *with* reexploration of recent operation

It is obvious that in the example above the prepositional phrase post-modifies only the reexploration and not the co-ordination of incision and reexploration.

* P1-38374   thrombectomy with catheter of iliac vein by **abdominal and leg** incision
→ thrombectomy with catheter of iliac vein by incising *abdomen* **and** leg

In the example above, many parsers would have big difficulties in combining the adjective "abdominal" with the noun "leg".

6b) *Make special features of coordinations explicit*

* P1-01007   incision **and** packing of wound
→ incision *followed by* packing of wound [21]

6c) *Avoid too complex cases of ellipsis in co-ordinated constructions*

---

[21] Although the term as such does not give any clues for a proper interpretation, from the place in the hierarchy we know that the wound does not exist prior to the incision but that it is the result of it.

```
* P1-31620    "01"   cardiac catheterization, right heart and retrograde left
  P1-31620    "02"   combined right heart catheterization and retrograde left
                     heart catheterization [22]
```

5. Discussion: what is (or not) to be expected from the use of controlled languages in healthcare

In the previous sections, we have used ENV 12381 to show how close a (semi-formal) regimented language can be related to a controlled language. In addition, we highlighted some problems related to the free style in which terms in nomenclatures (in casu Snomed International) are expressed, or otherwise stated, we showed how these terms are too far from a controlled language. We argued that such systems would benefit from the principles behind such languages.

Throughout the literature, controlled languages are recognised to have several advantages and disadvantages [xl].

First, especially for people, the reduction in lexical and structural ambiguity and the prescription of stylistic rules directly improve the readability and understandability of the text. As a consequence, a text that is easier to read and understand is obviously also easier to maintain and update. However, controlled languages may drastically reduce the power of expression, depending on the severity of the restrictions imposed. At least in the beginning, writing in a controlled language may require so much thinking about what words and what syntactic constructions to use that it reduces the speed with which writers can produce texts. Effective writing also requires a considerable amount of training.

It is argued that a certain tension exists between the need of domain experts to communicate among themselves in an efficient subsystem of their shared natural language, and the need to communicate technical information from the experts to "outsiders" (e.g. medical students, transcribers) in some understandable way. Within a human perspective, controlled languages merely serve the purposes of the second group. A closer mirroring of sublanguage grammatical features during controlled language design is expected to improve acceptance among both the producer and user communities [xli]. Whether or not at the same time, and in the medical domain, the requirements imposed for automatic language processing can be met, needs further to be explored.

In the light of recent developments in formal terminological systems [xlii], and their use for text analysis [xliii], text generation [xliv] and automatic knowledge acquisition [xxx], controlled languages may have a considerable impact. First, to remember the medical informatics community that our excitement regarding the "discovery" of the concept-based approach should not make us forget about the terms. The fact that in

---

[22] Surprisingly, the better formulation is available in Snomed though not marked as preferred term

recent proposals for desiderata for controlled medical vocabularies requirements for term-formation (i.e. naming concepts) do not anymore appear [xlv], might be seen as ominous, if not deplorable. After all, although manipulation of medical terminologies will in the future mainly be mediated by software, the communicative dimension of the language will not become less important.

The biggest advantage for controlled language use is undoubtedly the easier computational processing. The reduction in lexical and structural ambiguity and the prescription of stylistic rules makes it easier to process the texts computationally. Depending on the actual restrictions imposed by the controlled language, it may even be possible to guarantee that certain computational processes succeed. As a consequence, many applications will become possible.

Automated translation is an obvious example. Nowadays, billions of dollars are spent in translating medical nomenclatures and classifications, or in annotating medical concepts in various languages.

Accurate mapping of nomenclatures and classifications on the basis of their terms, by using deterministic natural language processing techniques is a second possibility. Though the development of a unique and universal model of medicine as is currently being conducted within the Galen Organisation Ltd [xxvi], could finally make such mappings obsolete, one never can guarantee that users will prefer to use other systems with which compatibility needs to be maintained. It is also to be expected that when users move from one classification system to another, they will require backward compatibility with data registered earlier.

Controlled language use in terminological systems - we don't dare to propose (yet) also in clinical narrative - would make automatic knowledge acquisition more feasible [xlvi]. Experience has shown that even when building systems such as Galen, domain experts figuring as modellers prefer to use near-natural language tools than bare formal languages such as Grail [xlvii]. And that to improve consistency amongst modellers a methodology is proposed wherein "paraphrases" are used to make the knowledge in the rubrics of classification systems more explicit and less ambiguous [xlviii], is another indication that the notion of controlled language is gradually being introduced in the healthcare telematics community. Because building formal terminological systems involves huge validation efforts, the use of controlled language checkers - whether used after texts have manually been written or during the editing work itself - will prove to be highly profitable.

6. Conclusion

The development of formal terminological systems, thesauri and classifications in the domain of healthcare will make language and "traditional" terminology not less important. The complexity of natural language makes the development of reliable natural language understanding or generation systems extremely difficult. Moving

towards the use of controlled languages in areas where much benefit and few opposition is to be expected, might prove to be extremely promising. The terms used in medical nomenclatures and classifications might be a good starting point to apply generally recognised principles for controlled language design. All it takes, is to make the decision, to write down the principles, and to let the machines do the work...

References

[i]   ISO 31-1 : 1992, Quantities and units. Pt1, Space and Time.

[ii]   ISO 8601 : 1988, Data elements and interchange formats - Information interchange - Representation of dates and times.

[iii]   Buekens F, Ceusters W, De Moor G. *The Explanatory Role of Events in Causal and Temporal Reasoning in Medicine*. Meth Inform Med 1993; 32: 274-8.

[iv]   Ceusters W, Buekens F. Towards a High Level Framework Model for the Description of Temporal Models in Healthcare Information Systems. In: Hoopen ten AJ, Hofdijk WJ, Beckers WPA (eds), Proceedings of MIC '92, Publicon Publishing Rotterdam, 1992, 41-50.

[v]   Ceusters W, Buekens F, De Moor G, Bernauer J, De Keyser L, Surjan G. *TSMI: a CEN/TC251 Standard for Time Specific Problems in Healthcare Informatics and Telematics*. International Journal of Medical Informatics 1997 (in press).

[vi]   Kittredge R., Lehrberger J. (eds.) : Sublanguage : studies of language in restricted domains. de Gruyter, Berlin, 1982.

[vii]   Harris Z. Mathematical Structure of Language. John Wiley & Sons. New-York, 1968.

[viii]   Harris Z. A theory of language and information. Clarendon Press, Oxford, 1991.

[ix]   Ceusters W., Spyns P., De Moor G., Martin W. (eds.) : *Syntactic-semantic tagging of medical texts: the MultiTALE-project*. IOS Press, Amsterdam, 1997.

[x]   Janton, P. Esperanto: Language, Literature, and Community. State University of New York, Albany, 1993.

[xi]   Okrand M. Conversational Klingon. Simon & Schuster Inc, New York,1992.

[xii]   Okrand M. The Klingon Dictionary. Simon & Schuster Inc, New York,1992.

[xiii]   Schreurs D., Adriaens G. Controlled English (CE): from COGRAM to ALCOGRAM. In: O'Brian Holt P, Williams N (eds.) Computers and writing: state of the art. Intellect, Oxford, 206-221, 1992.

[xiv]   Mitamura T, Nyberg E. Controlled English for Knowledge-Based MT: experience with the KANT system. Carnegie Mellon University, Pittsbergh, 1995.

[xv]   Ogden C. Basic English: A General Introduction with Rules and Grammar. Paul Treber & Co. Ltd., London, 1930.

[xvi]   AECMA. A guide for the preparation of aircraft maintenance documentation in the international aerospace maintenance language. AECMA, Paris, 1989.

[xvii]   Pulman SG, Rayner M. Computer Processable Controlled Language. SRI International, Cambridge Computer Science Research Centre, 1994.

[xviii]   Fuchs NE, Schwitter R. Attempto Controlled Natural Language for Requirements Specifications, In: ILPS'95, Seventh Workshop on Logic Programming Environments, Portland, December 1995, pp.25-32.

[xix]   Dodd T. Prolog: a logical approach. Oxford University Press, 1990.

[xx]   Alan J. Natural language Understanding. The Benjamin/Cummings publishing company Inc., Menlo Park, 1988.

[xxi]   Côté R.A., Rothwell D. (eds.), *Systematized Nomenclature of Medicine - SNOMED International*, College of American Pathologists, Chicago, 1993

[xxii] Rector AL, Nowlan WA, Glowinski A. Goals for Concept Representation in the GALEN project. In Safran C. (ed). *SCAMC 93 Proceedings*. New York: McGraw-Hill 1993, 414-418.

[xxiii] Rector AL, Glowinski A, Nowlan WA, Rossi-Mori A. Medical concept models and medical records: an approach based on GALEN and PEN&PAD. *Journal of the American Medical Informatics Association* 1995, 2: 19-35.

[xxiv] Rector AL, Nowlan WA, Kay S. Conceptual Knowledge: the core of medical information systems. In Lun KC, Degoulet P, Piemme TE, Rienhoff O (eds.). *MEDINFO 92 Proceedings*. Amsterdam: North - Holland 1992, 1420-1426.

[xxv]   Rector AL. Compositional models of medical concepts: towards re-usable application independent medical terminologies. In Barahona P & Christensen JP (eds.) *Knowledge and decisions in health telematics*. Amsterdam: IOS Press 1994, 133-142.

[xxvi] Rogers, J. and Rector, A. (1996). The GALEN ontology. Medical Informatics Europe (MIE 96), Copenhagen, IOS Press. 174-178.

[xxvii] Ceusters W, Deville G. A mixed syntactic-semantic grammar for the analysis of neurosurgical procedure reports: the Multi-TALE experience. In Sevens C, De Moor G (eds.) *MIC'96 Proceedings*, 1996, 59-68.

[xxviii] Ceusters W, Lovis C, Rector A, Baud R. Natural language processing tools for the computerised patient record: present and future. In P. Waegemann (ed.) *Toward an Electronic Health Record Europe '96 Proceedings*, 1996:294-300.

[xxix] Ceusters W, Spyns P. *From Natural Language to Formal Language: when MultiTALE meets GALEN.* In: Pappas C, Maglaveras N, Scherrer JR (eds.) Medical Informatics Europe '97, 396-400, IOS Press, Amsterdam, 1997.

[xxx]   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. In: *Proceedings of IMIA WG6 Conference on Natural Language and Medical Concept Representation*. Jacksonville 19-22/01/97, 71-80.

[xxxi] Ceusters W., Waagmeester A., De Moor G. Syntactic-semantic conventions for a medical treebank: the Cassandra approach. In: Proceedings of MIC'97, VVAA, Utrecht, 1997.

[xxxii] Sager JC. A Practical Course in Terminology. John Benjamins Publishing Company, Amsterdam, 1990.

[xxxiii] Frawley W. Linguistic Semantics. Lawrence Erlbaum Associates, London, 1992.

[xxxiv] Foley W, van Valin R. Functional syntax and universal grammar. Cambridge University Press, Cambridge, 1984.

[xxxv] ISO/DIS 860:1993, International harmonisation of concepts and terms.

[xxxvi] ISO/TR 12618:1994. Computational aids in terminology - creation and use of terminological databases and text corpora.

[xxxvii] ISO/CD 704: Terminology work - Principles and methods.

[xxxviii]   ISO/TC 37/SC3/WG1 12620.2. Computational aids in terminology - Data element categories.

[xxxix] Suonuuti H. Guide to terminology. Tekniikan Sanastokeskus, Helsinki, 1997.

Ceusters W, Steurs F, Zanstra P, Van der Haring E, Rogers J. From a time standard for medical informatics to a controlled language for health. International Journal of Medical Informatics 1998;48: 85-101.

[xl] Adriaens G, Havenith R, Wojcik R, Tersago B. (eds) Proceedings of the First International Workshop on Controlled Language Applications. Centre for Computational Linguistics, Leuven, 1996.

[xli] Kittredge R. Towards a friendlier relationship between sublanguage and controlled language. Unpublished guest lecture, University of Leuven, 18-06-96.

[xlii] Rector A. Thesauri and formal classifications: terminologies for people and machines. In: *Proceedings of IMIA WG6 Conference on Natural Language and Medical Concept Representation*. Jacksonville 19-22/01/97, 183-195.

[xliii] Rassinoux AM, Miller RA, Baud RH, Scherrer JR. Modeling just the important and relevant concepts in medicine for medical language understanding: a survey of the issues. In: *Proceedings of IMIA WG6 Conference on Natural Language and Medical Concept Representation*. Jacksonville 19-22/01/97, 53-68.

[xliv] Wagner JC, Solomon WD, Michel P-A, Juge C, Baud RH, Rector AL, Scherrer JR. Multlingual natural language generation as part of a medical terminology server. In: Greenes RA, Peterson H, Protti D (eds.) MEDINFO 95 Proceedings. North Holland, Amsterdam, 1995.

[xlv] Cimino JJ. Desiderata for controlled medical vocabularies in the twenty-first century. In: *Proceedings of IMIA WG6 Conference on Natural Language and Medical Concept Representation*. Jacksonville 19-22/01/97, 257-267.

[xlvi] Pulman G. Controlled Language for Knowledge Representation. In: Proceedings of the 1st Int. Workshop on Controlled Language Applications, 1996, pp.233-242.

[xlvii] Rogers JE, Solomon WD, Rector AL, Pole P, Zanstra P, van der Haring E. Rubrics to dissections to Grail to classifications. In: Pappas C, Maglaveras N, Scherrer JR (eds.) Medical Informatics Europe '97, 241-245, IOS Press, Amsterdam, 1997.

[xlviii] Galeazzi E, Rossi Mori A, Consorti F, Errera A, Merialdo P. A cooperative methodology to build conceptual models in medicine. In: Pappas C, Maglaveras N, Scherrer JR (eds.) Medical Informatics Europe '97, 280-284, IOS Press, Amsterdam, 1997.