Ceusters W. Language Engineering as an enabling technology for clinical terminology harmonisation. In: CEC-DGXIII (ed.), Important Issues in Today's Telematics Research, TAP 1998 Conference, Barcelona, 1998;:168-173.

Language Engineering as an Enabling Technology for Clinical Terminology Harmonisation

Werner CEUSTERS

Summary: In the Galen-In-Use project, syntactic-semantic tagging of terminology collections is used as a language engineering technique to populate language and application independent models of medical knowledge. The technique makes differences in conceptual and linguistic categorisation explicit and allows the automatic generation of lexicons and grammars.

1. Introduction

GALEN (*Generalised Architecture for Languages, Encyclopaedias and Nomenclatures in Medicine)* started as a research and development project in the Third Framework Programme to develop a semantically sound model of clinical terminology [1]. In the Fourth Framework's Galen-In-Use project, the model is further expanded in the domain of surgical procedures and put in practice at various sites for different clinical purposes. The model is built using medical classifications and nomenclatures in various languages, each of them designed for a different purpose. Yet a specific characteristic of the model is its independence from any particular language or application. It is intended to deliver a large number of services to clinical end-user applications, including language understanding and generation.

Within the Galen-In-Use project, the model started to be populated mainly manually by bringing together contributions from modelling centres in various Member States. In addition, corpus-based natural language analysis techniques are used to speed up the process. The first experiments of these techniques revealed a paradox: by striving for language independence, mapping from concepts to language could still be realised (though sometimes with pedantic or odd sounding results), but mapping from language to concepts, was far more cumbersome. Two reasons could be identified. Firstly, more often than expected a clinical categorisation of medical concepts does accord with a linguistic categorisation. Secondly, during the modelling technique itself, valuable linguistic information was thrown away and as a consequence was not available for linguistic processing afterwards [2].

2. The challenges

2.1 Harmonising linguistic representations with conceptual representations

While working on the language engineering aspects of Galen-In-Use, numerous examples were found where linguistic principles were in conflict with conceptual principles. Physicians want to see medical concepts organised in a framework that reflects their clinical way of thinking. As an example, the Galen model categorises the concepts of "filling" and "injecting" as specialisations of a "LiquidInstallingProcess" that itself is a child of "InstallingProcess". This categorisation is useful from a clinical perspective where from the place in the hierarchy it can be derived that the concepts of injecting and filling have to do with the installation of liquid (though not necessarily exclusively as the Galen model supports multiple parents). This categorisation is however in conflict with the linguistic structures that (at least in European languages) are used to express installing, filling and injecting events. From a language understanding perspective, it would be better to categorise these motion events according to the way the thematic roles of *goal* and *theme* may surface in sentences expressing these events. As can be seen from Figure 1, no straightforward relationship can be drawn between the two categorisations, a situation that forced us to bring the differences between a linguistic representation and a conceptual representation closer to the attention of the "formal medical terminology" community [3].

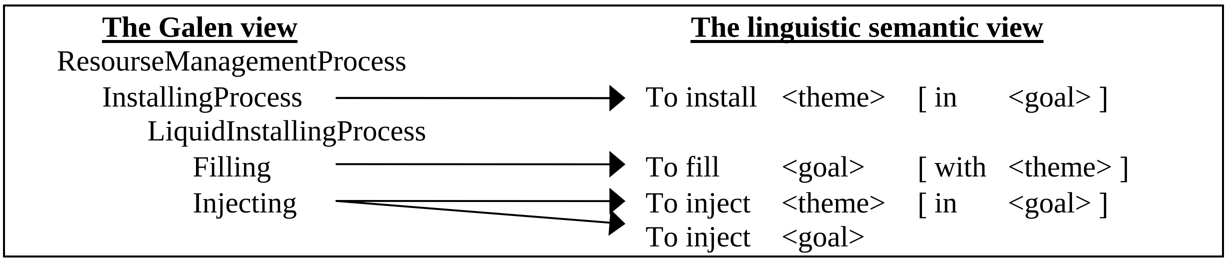
The Galen view
ResourseManagementProcess
InstallingProcess
LiquidInstallingProcess
Filling
Injecting
The linguistic semantic view
To install <theme> [in <goal>]
To fill <goal> [with <theme>]
To inject <theme> [in <goal>]
To inject <goal>

Figure 1: Differences in linguistic and conceptual categorisation

In addition, part-whole relationships demonstrate the differences in categorisation and actual expressions. Clinicians want to have the fingernail classified as part of the upper extremity, following a long chain of transitivity over "distal phalanx", "finger", "hand", "lower arm" and "arm", while they would never actually say that "a fingernail is a part of the upper extremity".

2.2 Preserving linguistic information when developing conceptual models

The Galen representation language (GRAIL) being too complex for direct manipulation by clinical modellers, an intermediate representation (more a notation than a true formalism) has been developed [4]. Within this notation, the actual expressions, possibly in different languages, that "inspired" modellers towards a conceptual representation of the expressions, are preserved as a whole, however without mappings between the constituent elements (Figure 2).

```
RUBRIC            "valgiserende osteotomie van humerus"
ENGLISH_RUBRIC    "valgising osteotomy of humerus"
PARAPHRASE        "osteotomy of humerus with purpose to create a valgising position"
MAIN              cutting
                      TO_ACHIEVE Deed:valgising
                          ACTS_ON Pathology:pathological posture
                      ACTS_ON Anatomy: humerus
```

Figure 2: GALEN dissection as intermediate representation for modelling procedures

3. The solution: Cassandra tagging on medical corpora

The goals of the Cassandra tagging are multiple. First, the tagging makes the relationships between the constituents of the phrases in the "rubrics" and "paraphrases" on the one hand, and the "main"-statement on the other hand, explicit. Second, it connects "linguistic" concepts and relationships to the "conceptual" representation of Galen. Third, it projects the conceptual representation on the surface structure of the expressions. Lastly, it allows automatic generation of lexicons, grammars and even a conceptual-linguistic cross-categorisation scheme on the basis of the tagged corpus at a later stage. As such, it combines the advantages of the pure conceptual approach (clean categorisation of medical concepts) with more corpus-linguistic oriented approaches [5, 6].

At the heart of the Cassandra tagging technique is a bracketing and encoding convention that relates the surface structure of a sentence to a linguistic representation and a conceptual representation. As an example, the sentences "*excision of cicatrix of skin*" and "*debridement of skin*" are respectively tagged as:

(1) *(excision)35 {[of]111 ((cicatrix)2120 {[of]216 (skin)474}0)0}0*
(2) *(debridement)82 {[of]142 ({palmar}1785 (skin)474)0}0*

where the different types of brackets categorise a sentence constituent as referring to a concept, a link (i.e. conceptually, or a thematic role linguistically), or a criterion (i.e. a link applied to a concept). This notation provides a fairly adequate bridge between the "topic-attribute-value" paradigm adhered to in Galen, and the predicate paradigm on which our linguistic engineering work is based. The numbers in each tagged example refer to a semantic lexicon that, restricted to the phrases presented above, can be represented as in Table 1.

RefId	Prototype	Conceptual repr.	Linguistic repr.
35	excision	excising	excising
82	debridement	debriding	debriding
111	of	ACTS_ON	THEME
142	of	ACTS_ON	SOURCE
216	of	HAS_LOCATION	SOURCE
474	skin	skin	skin
1785	palmar	[IS_PART_OF](palm)	[LOCATIVE](palm)
2120	cicatrix	cicatrix	cicatrix

Table 1: Semantic lexicon used in the Cassandra tagging technique

Some additional conversion rules are needed to generate the desired representations from the tagged sentences as not always (not to say seldom) a direct structural correspondence between the two representations is attainable. Clinicians for instance want to express that they "operate on pathologies that are located somewhere in the body", while they don't care about motion events and thematic roles at all even if they express it in that way (Figure 3).

Conceptual representation	**Linguistic representation**
excising	excising
ACTS_ON cicatrix	THEME cicatrix
HAS_LOCATION skin	SOURCE skin

Figure 3: Conceptual and linguistic representations of "excision of cicatrix of skin".

The Cassandra technique has also some particular features to cope with special phenomena such as "semantic gapping" as occurs in noun concatenation [7], e.g. the use of the asterisk in:

(3) *(division)49 {[of]84 ({(joint)129 [*]217}0 (cartilage)511*
{[of]217 ((foot)983 @and#622 (toe)984)0}0)0}0

4. Conclusion and recommendations

Galen-In-Use is now delivering medical concept models that can be used in a variety of clinical applications. Language engineering techniques are used to populate the model not only faster, but also to produce content based on empirical evidence. Building a model that encompasses the entire domain of medicine must indeed rely on priorities. On the other hand, performing linguistic analysis on the terms and expressions found in traditional terminological systems, revealed inconsistencies in wording and phrasing of these systems. As such, the language engineering techniques applied in the project do help both terminologists and knowledge engineers.

From this (ongoing) work two main recommendations can be proposed. Firstly, medical terminologists, more specifically those who develop coding and classification systems, nomenclatures and thesauri, should pay more attention to the language they use in these systems. Secondly, there still is a far too wide barrier between the medical language engineering community and the general computational linguistics community. Both have a lot to offer to each other. This potential should be further exploited in future European R&D projects.

5. References

[1] Rector AL, Solomon WD, Nowlan WA, Rush TW. (1994) "A Terminology Server for Medical Language and Medical Information Systems" In Scherrer JR (ed.) *Proceedings of IMIA Working Group 6*, Geneva.

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

[3] Ceusters W, Buekens F, De Moor G, Waagmeester A. (1979) "The distinction between linguistic and conceptual semantics in medical terminology and its implications for NLP-based knowledge acquisition." In Chute C (eds.) *Proceedings of IMIA WG6*, Jacksonville.

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

[5] Marcus M, Santorini B, Marcinkievicz MA. (1993) "Building a large annotated corpus of English: the Penn Treebank." *Computational Linguistics*, 19: 27-45.

[6] Bateman JA, Henschel R, Rinaldi F. (1995) *Generalized upper model 2.0*. Technical report, GMD/Institute für Integrierte Publikations- und Informationssysteme, Darmstadt, Germany.

[7] Ceusters W, Waagmeester A, De Moor G. (1997) "Syntactic-semantic tagging conventions for a medical treebank: the CASSANDRA approach." *Proceedings of MIC'97*, Velthoven, The Netherlands.