Ceusters W, Rogers J, Consorti F, Rossi-Mori A. Syntactic-semantic tagging as a mediator between linguistic representations and formal models: an exercise in linking SNOMED to GALEN. Artificial Intelligence in Medicine 1999; 15: 5-23.

# Syntactic-semantic tagging as a mediator between linguistic representations and formal modals: an exercise in linking SNOMED to GALEN.

Werner CEUSTERS
Language & Computing NV
Het Moorhof
Hazenakkerstraat 20
B-9520 Zonnegem
Belgium
Tel: +32 53 62 95 45
Fax: +32 53 62 95 55
Email: werner.ceusters@rug.ac.be

Jeremy ROGERS
University of Manchester
Department of Computer Science
Medical Informatics Groups
Oxford Road
Manchester M13 9PL
United Kingdom
Email: jeremy@cs.man.ac.uk

Fabrizio CONSORTI
Ist. 4 Clinica Chirurgica
Univ. "La Sapienza" Roma
V.le del Policlinico
I-00161 Roma
Italy
tel. +39+6+49970634
fax. +39+6+49970622
Email: consorti@axrma.uniroma1.it

Angelo ROSSI-MORI
Reparto Informatica Medica
Istituto Tecnologie Biomediche
Consiglio Nazionale delle Ricerche
viale Marx 15
I-00137 Roma
Italy
tel. + 39 6 - 827 71 01
fax + 39 6 - 827 36 65
Email: angelo@color.irmkant.rm.cnr.it

**Abstract**

Natural language understanding applications are good candidates to solve the knowledge acquisition bottleneck when designing large scale concept systems. A necessary condition is however that systems are built that transform sentences into a meaning representation that is independent of the subtleties of linguistic structure that nevertheless underly the way language works. The Cassandra II syntactic-semantic tagging system fulfils this goal partially. Within the GALEN-IN-USE project, it is used to transform linguistic representations of surgical procedure expressions into conceptual representations. In this paper, the proctology chapter of the SNOMED V3.1 procedure axis was used as a testbed to evaluate the usefulness of this approach. A quantitative and qualitative analysis of the data obtained is presented, showing that the Cassandra system can indeed complement the manual modelling efforts being conducted in the GALEN-IN-USE project. The different requirements related to linguistic modelling versus conceptual modelling can partly be accounted for by using an interface ontology, of which the fine tuning will however remain an important effort.

**Keywords:**

concept representation

linguistic semantics

syntactic-semantic tagging

natural language understanding

automated knowledge acquisition

# 1. Introduction

## 1.1 The problem

Medicine is one of these complex domains where new knowledge is accumulated at a daily basis, and at an exponential rate. Most of this knowledge resides in textbooks and papers, or more loosely structured in patient records. Despite the growing tendency to make this knowledge available in electronic format, the era of large scale knowledge based systems does not seem to have dawned yet. On the one hand, we are perhaps close, issues such as complexity and expressive power of knowledge representations being better understood. On the other hand, there is still a very long way to go as representing large quantities of knowledge is a major bottleneck if we ever want to develop systems that don't "fall of the knowledge cliff" [20].

One of the research domains that might come up with solutions for the knowledge acquisition bottleneck is natural language understanding. A necessary condition is however that systems could be built that transform sentences into a meaning representation that is independent of the subtleties of linguistic structure that nevertheless underlie the way language works [1]. Yet, here also is a bottleneck, be it in the form of a chicken and egg problem. Systems showing this much wanted behaviour must base their inferences on knowledge already available to them. The knowledge required comes in two different flavours. First, there is that kind of knowledge that is described as "linguistic semantics", the rules and principles explaining how literal meaning is grammaticalised or encoded in language [12]. It is this kind of knowledge that enables us for instance to understand the "sense" of an expression or sentence, i.e. the set (or network) of sense-relations that hold between it and other expressions within the same language [18], and that allows us to identify the same meaning independent of whether a given sentence is in the passive or active form. This knowledge is different from "conceptual semantics" or "conceptual knowledge" that describes what entities there are in the world that can be denoted by language. In the light of this distinction, an intriguing question is whether or not gaps in conceptual knowledge can be discovered by available linguistic knowledge, and if yes, how this can be achieved.

In this paper, we answer part of this question by generating conceptual representations on the basis of linguistic representations, and by analysing the problems and shortcomings detected.

## 1.2 The context

GALEN stands for Generalised Architecture for Languages, Encyclopaedias and Nomenclatures in Medicine. It started as a research and development project in the CEC's Third Framework Programme to develop a semantically sound model of clinical terminology [23, 24]. 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 on the basis of concepts that are found in 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. As such, it is intended to deliver a large number of services to clinical end-user applications and to assist the development of new or specialised classifications suiting local needs.

Natural language understanding and generation are part of the research conducted around potential uses for the Galen model. Text generation has even become part of the quality assurance of the knowledge acquisition process, a functionality that convinced the French government to take on the GALEN methodology for developing its new classification of medical procedures [5].

In addition, research is carried out on how natural language analysis techniques can be used to speed up the process of populating the model [8, 9, 10]. Originally, the model started to be populated manually by bringing together contributions from modelling centres in various European Union countries, a slow and labour intensive endeavour. Now it is hoped that bringing the adequate linguistic machinery in place, existing terminology collections can be processed for automatic knowledge acquisition purposes.

In this paper, we describe how semi-automatic syntactic-semantic tagging can be used as a vehicle to project linguistic representations on formal representations. The proctology chapter (codes P1-58xxx) of the procedure axis of SNOMED V3.1 was used as a testbed. Linguistic representations of the rubrics were transformed into Galen dissections which subsequently were processed further to extract new information, or to identify errors. As such we investigated whether or not this approach can be used as an alternative or an additional support to manual modelling.

## 2. Modelling surgical procedures in GALEN-IN-USE

Adding surgical procedure concepts collected from classification systems to the central GALEN model (CORE), is done in a two step approach.

The first step is a manual process during which a human modeller has to rewrite a surgical procedure rubric in the form of a “dissection” (Fig. 1). A dissection is a kind of intermediate representation used by the domain modellers that allow them not to be confronted with the complexity of the GALEN representation language itself (GRAIL) [14, 15]. The intermediate representation is not as complicated to use and learn as GRAIL, but it is also not an alternative notation for it as it is far less expressive than GRAIL. Dissections allow to represent concepts by means of “descriptors”, and relationships between concepts by “links”. Writing dissections is done on the basis of rubrics of existing classifications or nomenclatures, or of phrases from text books, patient records or other corpora. Some modelling centres derive dissections directly from the original rubrics, while others produce first paraphrases by removing ambiguity from the original phrases and possibly also by adding pragmatic knowledge that is not explicitly present in the phrases [13]. Tools have been developed to improve both quality and consistency at the level of dissection building.

The second step is a semi-automatic process performed by the TIGGER (Template Interpreter and Grail GEneratoR). TIGGER transforms the dissections into pure GRAIL, the native GALEN representation language [25], such that the knowledge represented in the dissections, becomes part of the CORE.

The process of pushing a dissection through TIGGER into the CORE model involves many stages at which the process may fail. The first step, a simple syntax check, is followed by a second step in which specific style normalisation constraints are checked. Following the directives of CEN ENV 1828 [6], it is for instance verified whether or not all surgical deeds referred to in the dissection, “act” on something, or

as described in CEN ENV 1828, have a “direct object”. Next, unmapped descriptors (step 3) and unmapped links (step 4) are identified. They usually indicate the introduction of knowledge that is not yet available in the model. Step 5 consists of checking whether or not the combinatorial constraints of known descriptors and links are respected. These constraints are not the same as the final CORE model constraints, but are an indication that the particular combination is one to check manually. The constraints are also used in the SPET, a modelling tool assisting modellers in writing valid dissections. Hence they are further referred to in this paper as “SPET-constraints”, and the totality of the knowledge as the “SPET-model” to mark the difference with the Galen CORE-model. Violation of SPET-constraints does not necessarily mean that the dissection contains an error. An unrecognised combination of descriptors and links, might indeed be a conceptual valid one, though not yet formally recognised as such by the system. In step 6, it is checked whether or not the declared GRAIL mapped entities of the dissection actually exist in the CORE model. At this stage, it might turn out that none of the mappings - usually for the links - actually applies in a specific case because link mappings are often context dependent, and when a link is used in a context which is not explicitly described, then the GRAIL expansion process fails. A last check may cause the candidate GRAIL representation to be rejected by the GRAIL classifier as nonsense. This is rare, because normally dissections are manually authored by skilled modellers and as a consequence they are probably correct.

The work described in this paper is an alternative to the manual modelling effort of representing rubrics as dissections. On the basis of terminological phrases found in the proctology chapter of the SNOMED procedure axis, a linguistic representation of each phrase is generated. This linguistic representation is then transformed afterwards into a dissection. The results are then processed by the TIGGER up to step 5 as described above. Because the dissections are automatically generated, one cannot be sure that they are semantically correct and as a consequence, automatic integration of “new” knowledge still requires manual validation. Part of the purpose of this work is to estimate how reliable the generated dissections are.

## 3. Linguistic representations of terminological phrases: the Cassandra II approach

A disadvantage of the manual modelling technique is the loss of linguistic information. During the modelling process, relationships between natural language constituents on the one hand, and GALEN-template elements on the other hand are not formally represented. Nevertheless these links do exist and have been used “mentally” by the modellers to represent the phrases in the format of a dissection. But when they are not explicitly represented, it is not possible to make use of this information afterwards when developing natural language understanding systems intended to build dissections automatically.

### 3.1 Syntactic-semantic tagging of medical treebanks

In [11], we described a syntactic-semantic tagging technique called Cassandra. The purpose of the Cassandra tagging technique is to re-introduce in an explicit and formal way the links between the semantic model and the surface language [7]. 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.

Fig. 2 gives the result of applying the Cassandra technique to the dissection of Fig. 1. Cassandra tagging of dissections consists of placing a number of explicitly labelled markers ("tags") in the original dissection according to a predefined syntax and following precise semantic conventions. Though the MAIN-statement of such a dissection is already structured according to predefined syntax, and hence is fully parsable, no formal relationships are maintained between the constituents of the MAIN-statement and the other components.

The general format of a tag is "premarker - constituent - postmarker - label", a specific example being "(removal)1", where "(" is the premarker, "removal" is the constituent, ")" is the postmarker, and "1" is the label. Labels can be compared to the indices used as the notational device for coreferencing noun phrases in linguistic analyses of sentences with this difference that the labels are also used at the level of the conceptual representation (i.e. the MAIN statement of a dissection).

There are various possibilities for what can be a constituent, depending on the place in the dissection where the tags appear. At the level of a MAIN-statement, a constituent corresponds to one of the basic semantic building blocks of the GALEN intermediate representation. At the level of a RUBRIC- or PARAPHRASE- statement, a constituent is a word or a group of words used in the statement.

The pre- and postmarkers indicate what kind of semantic building block the constituent corresponds with. The labels are a mechanism to mark explicitly the relationships between corresponding constituents across the various statements in a dissection. Between statements that are expressed by means of natural language such as RUBRIC and PARAPHRASE, these relationships are of type "synonymy" or "translation" depending on whether the related constituents are expressed in the same or a different language. Between the MAIN-statement and the other statements, the relationship is of type "has meaning", or its inverse "is grammaticalised through".

The complete tag set of the current version is outlined in Table 1. As shown, specific pre- and postmarkers are formally connected to each other, such that a premarker "opens" an item, and the corresponding postmarker "closes" it. As a consequence, tags can be made up of other tags to form "compound tags" without sacrificing syntactic context-independence when tags are embedded recursively. Stated otherwise, tagging of sentences is done in such a way that the normal word order of the sentence is not disturbed. In addition, embedding of tags is only allowed according to predefined combinatorial conventions based on semantic grounds (Table 2). Hence Cassandra tagging can also be seen as a bracketing technique combining phrase structure tagging with semantic tagging. In some cases such as in the linguistic phenomenon of ellipsis, or when world knowledge is not grammaticalised in the sentence, "empty tags", represented by an asterisk, are to be inserted (Fig. 3).

Though this technique is useful in contrastive studies to see how medical concepts are expressed in various natural languages, it is without modifications not directly useful for natural language understanding purposes. The Galen intermediate representation is not only "independent" from any particular language, it is also just "too far away" from natural language. Where the former can be seen as an advantage, the latter is certainly a disadvantage that only can be resolved by means of an additional knowledge interface. The gap between the Galen representation and the language is most prominent at

the level of the links both in terms of meaning and bracketing conventions. In Fig. 2, the tagging at the level of the RUBRIC is not the one that normally would be obtained in a general linguistic setting when a linguist (or natural language parser) was requested to do the job. Traditional semantic linguistic theories would not attach “from rectum” as a prepositional phrase to “foreign body”, the rectum being case-qualified as locative, but would rather attach “from rectum” to the main predicate of removing, hence being case-qualified as “source” and at the same level as “foreign body” which would be case-qualified as “theme”. The Galen reading of the phrase would be less “deviant” if the preposition “of” was used instead of “from”: *removal of foreign body of rectum*. Also the second prepositional phrase is differently understood: “under anaesthesia” would in a linguistic context be attached at the same level and case-marked as “time” (probably subspecified further according to a specific tense system), and not, as is done by the domain modeller, by interpreting it as a conjunction of two independent predicates.

Another disadvantage of this approach is that phrases and phrase constituents only are linked to their own semantic equivalents in Galen, but that there are no formal relationships between constituents beyond the borders of the dissection. By doing contrastive analyses, one can see that the “ACTS_ON” in Fig. 2 is grammaticalised differently in Fig. 3, but the Cassandra technique as such does not give any clues why that is the case.

These and other observations forced us to expand the first “naïve” approach into a second, more linguistically oriented one: Cassandra II.

### 3.2 Linguistic representations as precursors to conceptual representations.

In Cassandra II tagging, only natural language expressions are tagged, while dissections are intended to be generated automatically. Tagging in this case is a semi-automatic procedure where a parser tries to identify the correct syntactic-semantic structure of a sentence based on a semantic lexicon and a linguistic model generated during earlier parses. As for the development of the Penn Tree Bank [19], corrections are done manually where needed. We refer to the result of this process as a “linguistic representation”. The Cassandra II linguistic representation of the sentence in Fig. 2 is “*(removal)37 {[of]111 (foreign body)39}0 {[from]142 (rectum)1016}0 {[under]1439 (anesthesia)8}0*", while the one in Fig. 3 is represented as "*{({fine}119 (needle)117)0 [*]98}0 (biopsy)21 {[of]142 (rectum)1016}0*”. As in the first Cassandra approach, 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). As such, this notation provides still 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. Contrary to what happened in Cassandra I, the figures - when not zero - refer to a semantic lexicon that provides both a linguistic and conceptual interpretation of the constituents.

Table 3 shows the relevant parts of the semantic lexicon for the sentences in Fig. 2 and Fig. 3. The prototype field is just given here for better reading. Actually, words are not stored in the semantic lexicon, but retain their position in the corpus and are linked to the lexicon through the RefId that is unique for each row in the table. The conceptual representation field contains a representation of the table entries according to the Galen dissections. The linguistic representation anchors the table entries to a

linguistic model inspired on traditional predication theory [12, pp 198]. As such, the linguistic entity of "removing" is seen as a predicate that can take independent individuals as arguments. In the example given, these arguments are "foreign body" and "rectum", each of them instantiating the predicate in a different way. The linguistic model of Cassandra II is in fact a typology of the different ways that arguments instantiates predicates. This typology is intrinsically different from the typology adhered to in Galen. Whereas in Galen the typology is based on "medical facts", our typology is based on how such "medical facts" are grammaticalised in language. In Galen, "surgical removing" is seen as a "resource management process" that "acts on" a body structure. In our linguistic model, "removing" is seen as a predicate of the type "negative directed movement" that can take arguments such as the thing being moved, called the "theme", and the place from which the theme is negatively moved, called the "source". Both the notions of "theme" and "source" (according to various linguistic theories called "cases" or "thematic roles") have predictive power as they go hand in hand with specific syntactic phenomena in language. The preposition "from" for instance is a good case marker for the source in negative directed movement predicates. In the SNOMED V3.1 procedure axis, the preposition "from" appears 355 times, and in each single case, it marks the source of a negative directed movement ! One indeed can come up with counter examples . In the disease axis of SNOMED, "from" predominantly case-marks a causal role as in "D8-20430: Uterine scar from previous surgery affecting pregnancy". But here the predicate is not a negative directed movement, but a pathology that can have arguments fulfilling the thematic roles of "cause", "location", etc. Again, both of these roles do have some analogous relationship in the Galen model with respect to pathologies, but then only by virtue of medical reality and not because there are specific syntactic phenomena in natural language expressions related to these roles.

The task of transforming a linguistic representation into a conceptual one according to the Galen intermediate representation provisions, comes down to identifying the relationships between the two models and to setting up mechanisms that allow transformations when no direct relationships are found. In the most simple cases, this transformation consist of substituting the thematic role with the Galen link, as is often the case with THEME on the one hand and ACTS_ON on the other hand. Also when one thematic role can denote various Galen links, simple substitution is possible. In both cases, the required information can be derived from the semantic lexicon. In other cases, also the analysis tree needs to be rearranged (see section 6.1). In that case, additional conversion rules need to be used, a feature that was not implemented before the experiment described.

## 4. Methodology

The proctology chapter of SNOMED V3.1 (codes starting with P1-58) was used as a testbed to evaluate whether or not linguistic representations of the rubrics could be used to generate dissections. 5 out of 361 rubrics were not included in the experiment as from these rubrics, no linguistic representation could be generated. Linguistic representations were generated following the Cassandra II conventions as outlined above. They were then transformed into dissections automatically by the Cassandra II converter, taking the Galen guidelines for dissection development into account [15]. The dissections were then in batch processed by the TIGGER up to step 5 (see section 2) to evaluate potential incorporation in the Galen

model. At the beginning of the experiment, it was known that the Galen intermediate representation would not yet be expressively adequate enough to deal with all the conceptual information that would arise from the linguistic representation. As an example, conjunctions and disjunctions are represented somewhat differently in the generated dissections to account for phenomena that could not be expressed in one single dissection following the normal dissection conventions. For this reason, the TIGGER was adapted to display the following kind of information: 1) whether syntactic or semantic transformations were done on the generated dissections, 2) whether in the generated dissections descriptors or links were used that not (yet) had an equivalent concept or link in the SPET model, and 3) whether links were used in combination with concepts that not (yet) were sanctioned (i.e. allowed) by the SPET constraints. As a result, each generated dissection could then be classified in 9 different groups, depending on the combination of occurrences of the three types of findings (Fig. 4). The result was quantitatively and qualitatively analysed to find out: 1) what modifications would be needed at the side of the Cassandra II converter, and 2) whether the intermediate representation itself, or the knowledge contained in the SPET model, had to be updated.

## 5. Results

### 5.1 Quantitative analysis

Out of the 356 dissections generated, 134 (37.6%) were accepted by the TIGGER without modifications, while 222 (62.4%) were marked as requiring some manual intervention. 459 revision marks were given in total, what clearly indicates that most often more than one mark was given for a single dissection. In 195 cases, a link or descriptor was used that was not defined in the SPET model. Often, one rubric contained more than one such “unmapped” entity. In 177 cases a violation of SPET sanctions was detected. This could occur in cases where a link was simply not known (hence giving some overlap with the “unmapped” problem) or when concepts where used in combination with links without this combination being sanctioned by the SPET constraints. In 87 cases, transformations were done on the generated rubrics in order to bring them in the required format for the TIGGER. Fig. 4 gives an overview of the distribution of these cases over the dissections.

59 unmapped descriptors were responsible for 132 revision marks, where 10 unmapped links were responsible for the remaining 63 revision marks. Out of these 63, the link “HAS_SOURCE” was responsible for 41 of the revision marks.

The most frequent sanction violations were found to be the following: 1) in 20 cases the linguistic link “ATTRIBUTED_TO” was used without any counterpart in the SPET model, 2) in 44 cases the linguistic links “HAS_DESTINATION” and “HAS_SOURCE” were used in combination with procedures involving movement, whereas in Galen these links can only be used to specify a relationship between a device and a body part or pathology, while finally 3) in 32 cases, a mistake at the level of four concept annotations in the SPET model was responsible (see 5.2). The remaining 81 cases of sanction violation could not further be grouped in meaningful classes and are to be studied individually.

### 5.2 Qualitative analysis

Though the long term objective of the Cassandra system is to "understand" and represent conceptually what is uttered by physicians, its objective in this exercise was to discover knowledge that is not yet represented in the Galen model. As such, the 134 dissections that were accepted automatically by the TIGGER, are the least interesting ones: they did not bring on board new information. The remaining 222 dissections yield the most valid information: they can be used to see where the Cassandra system on the one hand, but perhaps even more the Galen model on the other hand can be improved.

When the TIGGER in our experiment marks a descriptor as being "unmapped", this might mean two things: either a concept does not yet exist in the SPET model that is referred to by the descriptor, or the concept does exist, but the descriptor (as a "word") is not yet known to be a reference for the concept. In the former case, the SPET model must be updated, while in the latter, the Cassandra system should use the "canonical" descriptor. Also some of the transformations performed by the TIGGER are indeed transformations at this level, be it with respect to links instead of descriptors such as replacing the link "HAS_REASON" by "MOTIVATED_BY".

In rare cases, and certainly theoretically, it might also be that Cassandra proposes on linguistic grounds a descriptor to be the denotation of a concept, while on medical grounds the domain modellers might argue that such a concept does not have a place in the Galen model. An example of this is the descriptor "disjuncted". Disjunctions sometimes pose problems when converting a linguistic representation into a dissection. As an example, whereas the sentence "*removal of anal seton or marker*" could be written in one single dissection using the "/" operator to denote the disjunction between "seton" and "marker" that both are modified by the adjective "anal", that is not anymore the case for the sentence "*removal of anal seton or unlisted marker*" (P1-58560) where the local scope of the adjective "unlisted" interferes with the scope of the "/" operator. Human domain modellers do indeed write down two separate dissections to feed this knowledge into the Galen model. From a linguistic viewpoint, this is however not acceptable as otherwise the actual expression would not be represented faithfully.

For sanction violations, two similar considerations apply: either Cassandra proposes "valid" combinations of links and descriptors that not yet are sanctioned by the SPET model, or, the output of Cassandra is wrong. A third possibility is that a combination is not sanctioned because one of the descriptors is "misunderstood" by Galen. This would mean that not the unavailability of information causes the combination not to be sanctioned, but an error in the SPET model as such. Given the great care with which the model is designed, this would be highly exceptional, but nevertheless, it did occur in this experiment: "incontinence" turned out to be a known descriptor referring to a "deed" instead of a "pathology", "cauterising" is known to the model as a "feature" instead of a "deed" and "dilating" refers to a "pathology" instead of a "deed". Finally the descriptors "complex" and "simple" are known as "position". These 4 errors in the model (or the "linguistic annotations" towards the model at the level of the TIGGER) are responsible for 40 (22.6%) of the 177 sanction violations ! A fifth inconsistency was found for the descriptor "fissure", where in the TIGGER it is used for an anatomical fissure (e.g. on the skull) while in Cassandra it is used for a pathological fissure, the "normal" fissure being referred to as "anatomical fissure". This accounted for an additional 3 sanction violations.

Many of the remaining sanction violations (except those related to the "HAS_SOURCE", "HAS_DESTINATION" and "ATTRIBUTED_TO" links) can be brought back to "allowable sanctions", and as such discovering them contributes to the knowledge acquisition process. Examples of not accepted combinations are "DigestiveSystemAnatomy - IS_LOCATION_OF - Lesion" as in "P1-58E32: *manual reduction of prolapsed rectum*", "Lesion - IS_LOCATION_OF - Lesion" as in "P1-58185: *incision of thrombosed hemorrhoid*", "BodySubstance - IS_PART_OF - BodySubstance" as in "P1-58374: *excision of lesion of perirectal tissue*", and "Deed - MOTIVATED_BY - Lesion" as in "P1-58554: *perirectal injection of sclerosing solution for prolapse*". Most often, a too narrow base or domain of the link is the cause for the sanction violation. It is however a matter of debate whether or not these sanctions should be relaxed, whether new links should be introduced, or whether other mechanisms must be put in place. This will further be dealt with in the discussion (see section 6).

In very few cases in which a sanction violation was marked, the generated dissection was simply wrong due to an erroneous linguistic representation. In "*destruction of lesion of rectum by chemicals*" for instance, "by chemicals" was attached to "rectum" instead of to "destruction". In some other cases, dissections are "formally wrong" according to the Galen style of modeling, but serious linguistic evidence prevented the linguists to meet the demands of the modellers at the level of the linguistic representation. For injection procedures for instance, as in "P1-58550: *injection of sclerosing solution into hemorrhoids*", the dissection conventions require the "hemorrhoids" as destination to be linked to the injected substance (in this case "sclerosing solution") while linguistically, "hemorrhoids" figure as "goal" in the positive directed movement predicate "injecting". This requires further processing beyond the principles described in section 3.2, a feature that is not yet implemented in the Cassandra II converter, nor in the TIGGER. It is however obvious that one of the goals of the Cassandra II system is to provide Galen with the representation it expects.

Most of the transformations can be brought back to link substitution or to resolving the disjunction-conjunction problem raised above. Additional transformations occur to bring a number of indentations in the linguistic representation back to a status of no indentation at the level of the dissection. This is for instance the case for the link "HAS_PATIENT" that according to the Galen dissection principles, should be put at the same level as the "MAIN" statement. This cannot be achieved linguistically as this would disturb the predicate representation dramatically. But again, an additional step will be needed in the analysis process to improve the performance of the Cassandra system.

## 6. Discussion

### 6.1 Linguistic modelling versus conceptual modelling

Language understanding is a process that traditionally is recognised to be the result of various kinds of knowledge: phonological, morphological, syntactic, semantic, pragmatic and world knowledge [1]. This separation of knowledge is normally maintained in natural language understanding systems, though usually in one common framework. The design of the pragmatic or world knowledge base is guided by principles relating to the language understanding task. This results in different designs than for instance are used in expert systems.

Galen's primary aim is not to serve natural language processing applications, but to build models that preserve a clean separation of medical taxonomies relating to different viewpoints. At the heart of Galen is a novel description logic explicitly designed for medical applications through which an ontology has been produced which is claimed to be highly effective at meeting its primary goals of mediating amongst medical terminologies [22] and supporting user interfaces [16]. A so called "multilingual module" has been built in Galen, be it based on simple mechanisms which rely on linguistic annotations indexed by the primary ontology. In language generation experiments, or perhaps more precisely "paraphrase" generation, this approach has proved to be reasonably successful [28, 29]. However, much of the information required for more sophisticated linguistic processing relates to the lexical items themselves and their internal organisation in a sentence rather than to the domain concepts which they represent. Also the criteria for modelling adopted for constructing a domain model on the one hand, and for constructing a model suitable for driving linguistic processing on the other hand, do not always coincide [4]. Though in a medical informatics context in general, and certainly in the context of our work, a model is required that drives linguistic processing towards some given target representation, we deliberately did not want the target representation and its well-formedness criteria as part of the linguistic model.

Given the nature of surgical procedures where most often anatomical structures are being displaced or worked upon, we opted for a linguistically inspired high level ontology where events are distinguished from entities, and further subdivided into states, acts, inchoatives and resultatives [12, pp183-184]. For each of those, motional events are distinguished from non-motional events. As a consequence, procedures most often are motional acts involving thematic roles such as THEME, DESTINATION, SOURCE, LOCATION and PATH.

Having two ontologies sitting next to each other, and because the goal of our work is to derive conceptual knowledge from linguistic information, mechanisms were to be provided to pass from one to the other. One option would have been to unify both ontologies. Other authors have made some progress in this direction by unifying the top layers of the domain-oriented Cyc ontology [17] with the language-oriented ontology of PANGLOSS [26]. However, there is a strong belief based on these experiences that unification will not always be possible or desirable and that transformation between domain-oriented and language-oriented ontologies will be a key strategy. Our approach is to maintain two different ontologies, and to develop transformation mechanisms based on an "interface ontology" between them [7]. The information to do this is partly stored in the semantic lexicon of Cassandra (Table 3), and partly in the link-conversion rulebase of the Cassandra II converter. In the future, a Cassandra-version of the TIGGER is one of the options further to be investigated.

### 6.2 Generic linguistic ontologies versus medical linguistic ontologies

Adopting an interface ontology is an established method for insulating an application from natural language constructs in a practically clean and theoretically well-understood fashion. One of the currently most established linguistically motivated interface ontologies is the Generalised Upper Model [2], a multilingual extension of the Penman Upper Model [3]. As a linguistically oriented ontology, the GUM is fundamentally different in design from domain- or world-knowledge oriented ontologies in that it

captures those distinctions which have influences for grammatical expressions in distinct languages without committing to just what the grammatical distinctions of any particular language are. This therefore provides a powerful point of language localisation that maintains theoretical independence from particular linguistic theories and language engineering techniques.

A relatively similar, though more simple approach is used in EuroWordNet [27]. In this project, semantic databases like WordNet1.5 [21] for several languages are combined via a so-called inter-lingual-index (ILI). This allows language-independent data to be shared over the languages, while language-specific properties are maintained as well in each individual database. The only organisation provided to the ILI is via two separate ontologies. The first one is the top-concept ontology which is a hierarchy of language-independent concepts, reflecting explicit opposition relations. The second is a hierarchy of domain labels. Both the top-concepts and the domain labels can be transferred via the equivalence relations of the ILI to the language-specific meanings and, next, via the language-internal relations to any other meaning in the individual database of a specific language (Fig. 5).

Within the Cassandra ontology, also generic linguistic principles dominate in the design. However, because the system is (at least currently) only intended to work within the medical domain, more domain-dependent configurations can be found. To maintain close compatibility (though not dependency) with the Galen model, higher level concepts available in the Galen model are reused in Cassandra. The hierarchical relationships amongst them are however different, and purely based on linguistic evidence. 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 does however not line up 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 is better to categorise these motion events according to the way the thematic roles of *goal* and *theme* may surface in sentences expressing these events, more precisely how syntactic roles or prepositions mark a preferential thematic role (Fig. 6).

### 6.3 Automatic versus manual modelling

A question to be investigated further is whether the linguistic technique is an alternative to manual modelling, or whether it is to be considered an additional tool to be used by modellers. When the Cassandra generated dissections were compared with those produced by human modellers on the same SNOMED rubrics, some more differences, other than those related to the use of linguistic roles instead of conceptual ones, became apparent. Most of them have to do with issues such as modelling style and normalisation of dissection building. Whereas in linguistic generated dissections the deed following the MAIN label is directly derived from the semantic head of the rubric, one modelling centre in the Galen In Use project requires this deed being restricted to the topographical or morphological interpretation of what the procedure actually carried out or created. Examples of such deeds are draining, shunting, repairing, reshaping, etc. Elementary deeds such as cutting, inserting or removing are also proposed to be

used only as arguments of the BY_TECHNIQUE link. Functional interpretations of what should be achieved by a surgical procedure as a whole, are in the same way proposed to appear only as arguments of the TO_ACHIEVE_OVERALL link. These modelling conventions would guarantee that linguistic expressions such as “relocation of mammary artery” (in which figures an elementary deed), “coronary artery shunt” (in which figures a morphological interpretation of the deed), and “myocardial revascularisation ” (in which the same deed is expressed by referring to its intended functional result) are all classified in the same way.

It is clear that in the current phase the Cassandra system is not able to fulfil these requirements. This would require a large world knowledge base to perform the necessary transformations on the linguistic representation. Building this knowledge base (the Galen CORE model) is exactly the purpose of the human modellers. However, when the knowledge base grows, human modellers will have problems to remember what is already there, and what isn’t. Even a large collection of surgical procedure expressions such as the SNOMED procedure axis cannot guarantee that all surgical procedures are covered. Hence other knowledge sources have to be consulted. It is however obvious that it makes no sense to model all rubrics of this new knowledge source as one can expect most of the rubrics being already represented in the conceptual model. Here the Cassandra tool will prove to be valuable for separating out those rubrics that cannot be processed automatically and hence are to be reviewed manually. The more linguistic and conceptual knowledge that is available, the less manual reviewing will be required.

## 7. Conclusion

The results of this experiment show clearly that generating dissections out of linguistic representations of terminological phrases by using the Cassandra system, can assist the manual modeling process. The technique can be used to extract unknown concepts from natural language texts and to verify whether or not sanctions in the conceptual model should be relaxed or further narrowed. The technique also proves to be valid for detecting inconsistencies in the conceptual model itself.

On the other hand, the linguistic approach is not able to deal easily with certain modeling demands such as splitting up dissections in which disjunctions or conjunctions with a mixed scoping phenomenon occur. Also decisions on what is to be identified as an atomic concept can be motivated differently when linguistic versus conceptual arguments are taken into account. All this requires further fine tuning of the machinery put in place: perhaps at the level of the Cassandra II converter, the TIGGER or the intermediate representation, but certainly at the level of the interface between them. As such, this work shows again that conceptual modeling must be complemented by linguistic modeling. This does not mean that linguistic models and conceptual models are to be seen as different views on one more generic model. Rather we are convinced that despite their different nature the development of an interface ontology connecting linguistic models to medical conceptual models is an area of research that further must be expanded.

## 8. References


1 J. Allen, Natural Language Understanding (The Benjammin/Cummings Publishing Company Inc, Menlo Park California, 1987).

2 J. Bateman, R. Henschel and F. Rinaldi, Generalised Upper Model 2.0: documentation, GMD/Institute for integrated publication and information systems Technical Report, Darmstadt, Germany, 1995.

3 J. Bateman, R. Kasper, J. Moore and R. Whitney, A general organisation of Knowledge for natural language processing: the PENMAN Upper Model, USC/Information Sciences Institute, Marina del Rey, California., 1990.

4 R. Baud, C. Lovis, L. Alpay, A-M. Rassinoux, J-R. Scherrer, A. Nowlan, and A. Rector, Modelling for Natural Language Understanding, in: C. Safran, ed., Proceedings of SCAMC 93 (McGraw-Hill Inc., New York, 1993) 289 - 293.

5 R. Baud, J-M.Rodrigues, J.C. Wagner, A-M. Rassinoux, C. Lovis, P. Rush, B. Trombert-Paviot, and J-R. Scherrer, Validation of Concepts Representation Using Natural Language Generation. Submitted to SCAMC Fall symposium 1997.

6 CEN, ENV 1828:1995, Medical Informatics - Structure for classification and coding of surgical procedures (CEN, 1995).

7 W. Ceusters, F. Buekens, G. De Moor, and A. Waagmeester, The Distinction between Linguistic and Conceptual Semantics in Medical Terminology and its Implications for NLP-Based Knowledge Acquisition, in: C. Chute, ed., Proceedings of IMIA WG6 Conference on Natural Language and Medical Concept Representation (IMIA WG6, Jacksonville, 1997) 71-80.

8 W. Ceusters, G. Deville, and G. De Moor, Automated extraction of neurosurgical procedure expressions from full text reports: the Multi-TALE experience, in: J. Brender, J.P. Christensen, J.-R. Scherrer, P. McNair, eds., Proceedings of MIE 96 (Ios Press, Amsterdam, 1996) 154 - 158.

9 W. Ceusters, G. Deville, O. Streiter, E. Herbigniaux, and J. Devlies, A Computational Linguistic Approach to Semantic Modelling in Medicine, in: W.P.A. Beckers, A.J. ten Hoopen, eds., Proceedings of MIC'94 (Velthoven, The Netherlands, 1994) 311-319.

10 W. Ceusters and P. Spyns, From Natural Language to Formal Language: when MultiTALE meets GALEN, in: C. Pappas, N. Maglaveras, J.-R. Scherrer, eds., Medical Informatics Europe '97 (IOS Press, Amsterdam, 1997) 396 - 400.

11 W. Ceusters, A. Waagmeester, and G. De Moor, Syntactic-semantic tagging conventions for a medical treebank: the CASSANDRA approach, in: J. van der Lei, W.P.A. Beckers, W. Ceusters, and J.J. van Overbeeke, eds., Proceedings MIC'97 (Veldhoven, The Netherlands, 1997) 183-193.

12 W Frawley, Linguistic Semantics (Lawrence Erlbaum Associates, Hillsdale, 1992).

13 E Galeazzi, A Rosi-Mori, F Consorti, A Errera, and P. Merialdo, A cooperative methodology to build conceptual models in medicine, in: C. Pappas, N. Maglaveras, J.-R. Scherrer, eds., Medical Informatics Europe '97 (IOS Press, Amsterdam, 1997) 280-284.

14 GALEN Consortium. Guidelines and Recipes for Completing templates. Internal document VUM02/96.

15 GALEN Consortium. Links and Templates Summary. Internal document VUM/03/96.

16 J. Kirby and A. Rector, The PEN&PAD Data Entry System: From prototype to practical system. in: AMIA Fall Symposium Proceedings (Hanley and Belfus Inc.,Washington DC, 1996) 709-713.

17 D. Lenat and R.V. Guha, Building large knowledge-based systems: representation and inference in the CYC project (Addison-Wesley Publishers, New York.,1989)

18 J. Lyon, Linguistic Semantics, an introduction (Cambridge University Press, Cambridge, New-York, Melbourne, 1995).

19 M. Marcus, R. Santorini, MA Marcinkievicz, Building a large annotated corpus of English: the Penn Treebank. Computational Linguistics, 19/2 (1993) 313-330 .

20 R.S. Michalski, Understanding the nature of learning: issues and research directions, in: R.S. Michalski, J.G. Carbonell and T.M. Mitchell, eds., Machine Learning, an artificial intelligence approach, vol II (Morgan Kaufmann Publishers Inc., Los Altos, 1986) 3-25.

21 G.A. Miller, R. Beckwidth, C. Fellbaum, D. Gross, and K.J. Miller , Introduction to WordNet: An On-line Lexical Database, International Journal of Lexicography 3/4 (1990) 235-244.

22 A. Rector, Coordinating taxonomies: Key to re-usable concept representations, in: Fifth conference on Artificial Intelligence in Medicine Europe (Springer, Heidelberg, 1995) 17-28.

23 A. Rector, W. Nowlan. and A. Glowinski, Goals for Concept Representation in the GALEN project, in: C. Safran, ed., 17th Annual Symposium on Computer Applications in Medical Care (McGraw Hill, New York, 1993) 414-418.

24 J. Rogers and A. Rector, The GALEN ontology, in: J. Brender, J.P. Christensen, J.-R. Scherrer, P. McNair, eds., Proceedings of MIE 96 (Ios Press, Amsterdam, 1996) 174-178.

25 J. Rogers, W. Solomon, A. Rector, P. Pole, P. Zanstra, and E. van der Haring, Rubrics to Dissections to GRAIL to Classifications, in: C. Pappas, N. Maglaveras, J.-R. Scherrer, eds., Medical Informatics Europe '97 (IOS Press, Amsterdam, 1997) 241 - 245.

26 W. Swartout, R. Patil, K. Knight and T. Russ, Towards distributed use of large-scale ontologies, in: Knowledge Acquisition Workshop, Banff, Alberta, Canada, 1996.

27 P. Vossen, P. Diez-Orzas, and W. Peters, The Multilingual Design of the EuroWordNet Database. in: Proceedings of the IJCAI-97 workshop on Multilingual Ontologies for NLP Applications, Nagoya, August 23, 1997.

28 J. Wagner, A.-M. Rassinoux, R. Baud, and J.-R. Scherrer, Generating noun phrases from a medical knowledge representation, in: P. Barahona, M. veloso, and J. Bryant, eds., Proceedings of Twelfth International Congress of the European Federation for Medical Informatics, MIE-94 (Lisbon, Portugal, 1994) 218-223.

29 J. Wagner, W. Solomon, P. Michel, C. Juge, R. Baud, A. Rector, J.-R. Scherrer, Multilingual Natural Language Generation as Part of a Medical Terminology Server, in: R. Greenes, H. Peterson, D. Protti, eds., Proceedings of MEDINFO 95 (North-Holland, Amsterdam, 1995) 100-104.

RUBRIC "removal of foreign body from rectum under anesthesia"
PARAPHRASE "removal of intraluminal foreign body from rectum under anesthesia"
CODE "P1-58378"
MAIN removing
    ACTS_ON foreign body
        HAS_LOCATION rectum
        HAS_POSITION intraluminal
WITH anesthesia procedure
    ACTS_ON patient

Figure 1

RUBRIC "(((removal )$_1$ {[of ]$_{10}$ ((foreign body)$_2$ {[from ]$_7$ (rectum )$_3$}$_8$)$_{11}$}$_{12}$)$_{13}$ &under$$_9$ (anesthesia)$_4$)$_{17}$"
PARAPHRASE "(((removal)$_1${[of ]$_{10}$ ({intraluminal}$_5$ (foreign body)$_2$ {[from ]$_7$ (rectum)$_3$}$_8$)$_{11}$}$_{12}$)$_{13}$
&under$$_9$ (anesthesia)$_4$)$_{17}$"
CODE "P1-58378"
MAIN (((removing)$_1$
    {[ACTS_ON]$_{10}$ ((foreign body)$_2$
        {[HAS_LOCATION]$_7$ (rectum)$_3$}$_8$
        {[HAS_POSITION]$_{18}$ (intraluminal)$_6$}$_5$)$_{11}$}$_{12}$)$_{13}$
&WITH$$_9$ ((anesthesia procedure)$_4$
    {[ACTS_ON]$_{14}$ (patient)$_{15}$ }$_{16}$)$_4$)$_{17}$

Figure 2

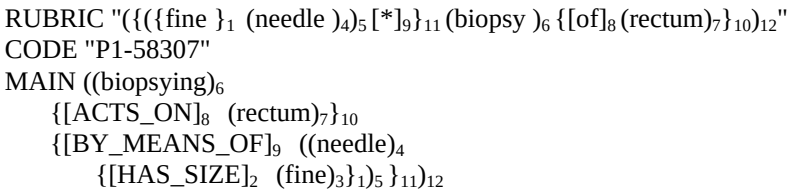
RUBRIC "({({fine }1 (needle )4)5 [*]9}11 (biopsy )6 {[of]8 (rectum)7}10)12"
CODE "P1-58307"
MAIN ((biopsying)6
{[ACTS_ON]8 (rectum)7}10
{[BY_MEANS_OF]9 ((needle)4
{[HAS_SIZE]2 (fine)3}1)5 }11)12

Figure 3

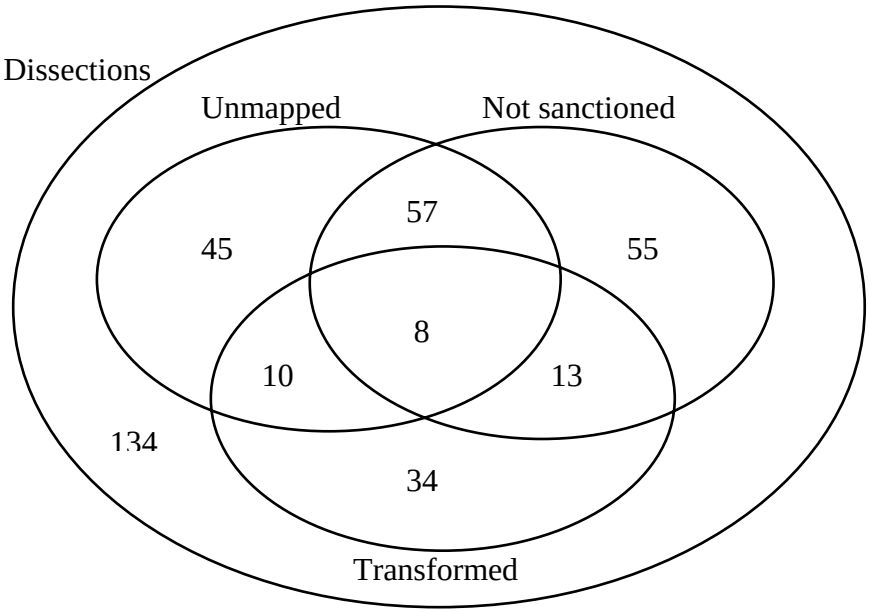
Dissections
Unmapped
Not sanctioned
57
45
55
8
10
13
134
34
Transformed

Figure 4

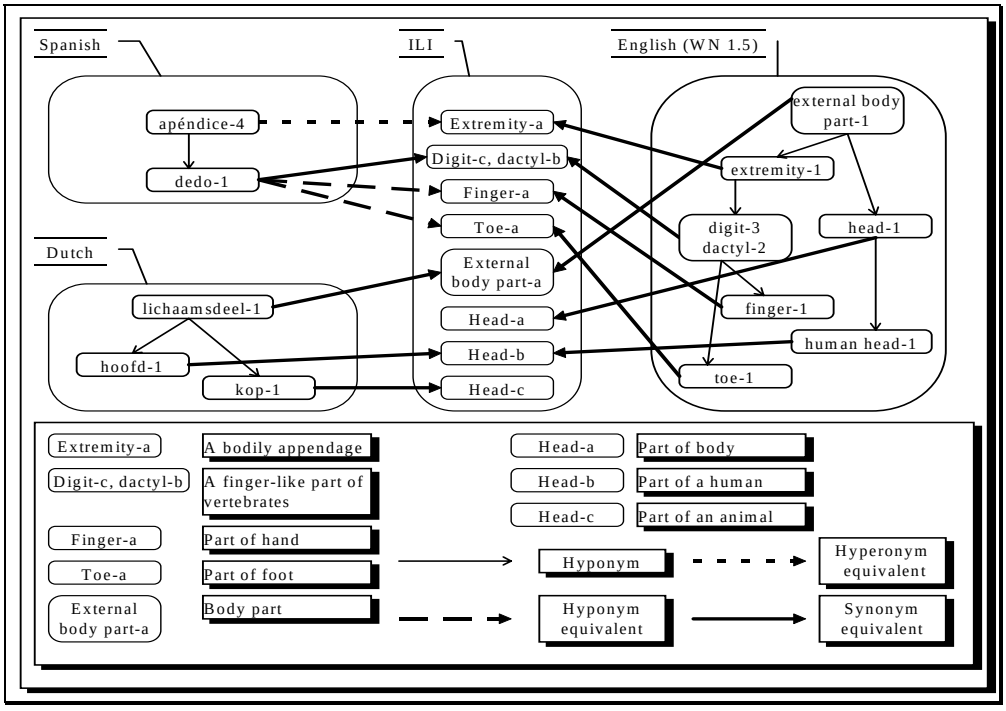
Spanish
ILI
English (WN 1.5)
apéndice-4
dedo-1
Extremity-a
Digit-c, dactyl-b
Finger-a
Toe-a
External body part-a
Head-a
Head-b
Head-c
external body part-1
extremity-1
digit-3 dactyl-2
head-1
finger-1
human head-1
toe-1
Dutch
lichaamsdeel-1
hoofd-1
kop-1
Extremity-a
A bodily appendage
Digit-c, dactyl-b
A finger-like part of vertebrates
Finger-a
Part of hand
Toe-a
Part of foot
External body part-a
Body part
Head-a
Part of body
Head-b
Part of a human
Head-c
Part of an animal
Hyponym
Hyponym equivalent
Hyperonym equivalent
Synonym equivalent

Figure 5

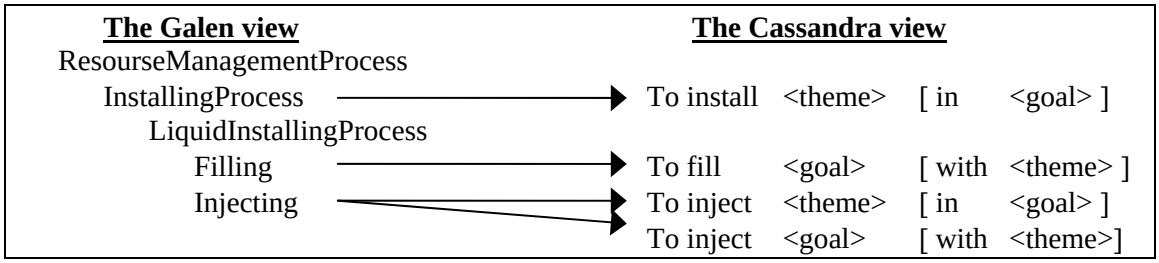
The Galen view
The Cassandra view
ResourseManagementProcess
InstallingProcess
To install <theme> [ in <goal> ]
LiquidInstallingProcess
Filling
To fill <goal> [ with <theme> ]
Injecting
To inject <theme> [ in <goal> ]
To inject <goal> [ with <theme>]

Figure 6

**Figure Captions**

Fig. 1: Manually generated dissection of a rubric from the SNOMED procedure axis. The RUBRIC level specifies the original sentence, the PARAPHRASE is a sentence contructed by the domain modellers to make the meaning more explicit, the CODE label specifies a reference to the corpus used, and the MAIN label introduces the semantic representation proper.

Fig. 2: Dissection tagged according to the Cassandra technique (see text for explanation of notation).

Fig. 3: Dissection in which an empty tag (indicated by *) is used to maintain normal word order when mapping phrase constituents to intermediate representation constituents.

Fig. 4 : Distribution of 459 revision marks generated by the TIGGER over the dissections generated from a linguistic representations. One dissection can have no or at most three revision marks.

Fig. 5 : In EuroWordNet, the ILI is used as an interface ontology to account for semantic differences in various languages. (Reproduced with kind permission from [27])

Fig. 6 : Galen categorisation versus linguistic categorisation.

| Pre- and post-marker | Relationship with the GALEN ontology (exhaustive) | Relationship with natural language phenomena (examples) |
|---|---|---|
| […] | semantic link | explicit in prepositions, or implicit in adjectives |
| (…) | descriptor (concept) | nouns, idioms |
| {…} | criterion (link + concept) | adjectives, adverbial constructions |
| @…# | local conjunctions | “and”, “or” |
| & ... $ | “MAIN”-conjunctions | “and”, “with”,”including” |
| \…/ | not represented in GALEN | function words such as articles, possessive pronouns, etc. |
| <…> | criterion modifier | adverbs |

Table 1

| Tag embedding | Use |
|---|---|
| { [a]1 (b)2 }3 | a “link” with a concept makes up a “criterion” (e.g. tag 8 in Fig. 2) |
| ({a}1 (b)2)3 | one or more “criteria” applied to a concept make up a new concept (e.g. tag 11 in Fig. 2) |
| ((a)1 &b#$ (c)3)4 | a coordination of tags of the same type make up a new tag of the same type (e.g. tag 17 in Fig. 2) |
| (\a/1 (b)2)2 | combining a “GALEN”-tag with a non-GALEN tag gives an embedded tag with the same meaning as the GALEN-tag |
| { <a>1 {b}2 }3 | modification of a criterion gives a new criterion |

Table 2

| RefId | Prototype | Conceptual repr. | Linguistic repr. |
|---|---|---|---|
| 8 | anaesthesia | anesthesia procedure | anaesthesia_procedure |
| 21 | biopsy | biopsying | biopsying |
| 37 | removal | removing | removing |
| 39 | foreign body | foreign body | foreign body |
| 98 | with | BY_MEANS_OF | INSTRUMENT |
| 111 | of | ACTS_ON | THEME |
| 117 | needle | needle | needle |
| 119 | fine | [HAS_SIZE](fine) | [SIZE](fine) |
| 142 | from | HAS_LOCATION | SOURCE |
| 1016 | rectum | rectum | rectum |
| 1439 | under | WITH | OCCURS_DURING |

Table 3

**Table captions**

Table 1: Cassandra tag-set for linking constituents in natural language to constituents of the Galen intermediate representation.

Table 2: Tag embedding rules as a syntactic-semantic bracketing technique to link phrase structures to conceptual model constituents. The ordering of tags within a block is irrelevant at the level of the rules as this is dictated by the principle of not interfering with word order.

Table 3: Semantic lexicon of the Cassandra II system. “RefId” is the unique identifier for an entry, “Prototype” is a typical word as found in natural language expressions. “Conceptual repr.” links the entry to the Galen intermediate representation while “Linguistic repr.” fullfils a similar role towards a linguistic typology of links and concepts.