Syntactic-Semantic Tagging Conventions for a Medical Treebank: the CASSANDRA approach

W. Ceusters (1,2), A. Waagmeester (1), G. De Moor (2)

1 *Language & Computing NV, Hazenakkerstraat 20, B-9520 - Zonnegem, Belgium*
2 *RAMIT VZW, University Hospital, De Pintelaan 185, B-9000 Gent, Belgium*

A treebank is a corpus of tagged and bracketed sentences capturing the linguistic properties of a (sub)language in an empirical way. The CASSANDRA treebank is developed as a sideline to the GALEN-IN-USE project in which it serves to make the relationships between natural language phenomena and semantic representations of medical expressions explicit, and to assist in the quality assurance of the modelling centres. The end result is a multilingual linguistic knowledge repository from which lexicons and grammars of various types can be derived in an automatic way.

1. Introduction

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 [1]. At the heart of the project is the development of a reference model for medical concepts (CORE) supported by a formal language for medical concept representation (GRAIL) [2]. A particular characteristic of the approach is the clear separation of the pure conceptual knowledge from other types of knowledge, including linguistic knowledge [3], in order to arrive in the future to application-independent medical terminologies [4].

In the GALEN-IN-USE project, various centres are collaborating to build an exhaustive model for surgical procedures [5]. An initial hypothesis was that this modelling work could be speeded up by semi-automatic processes relying on natural language processing techniques. The MultiTALE-I 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 ENV1828:1995 standard “Structure for the classification of surgical procedures.” [6,7].

Given the promising results of the MultiTALE-I system as compared to four similar systems described in the literature [8], it was first investigated whether the tagger could be used for the automatic generation of GALEN-dissections from natural language expressions out of the SNOMED procedure axis. These dissections are a kind of intermediate representation used by the domain modellers in order not to be confronted with the complexity of the GRAIL language itself [9, 10]. This turned out to be feasible

indeed, although a lot of efforts and resources had to be invested in providing sufficient medical knowledge to the new MultiTale-II parser for the delivery of acceptable results [11]. In fact, it became clear that contrary to what originally was expected, far more extra-linguistic knowledge was required to transform surgical procedure natural language expressions automatically into GALEN-templates with the expected level of detail. In addition, from the surface language of surgical procedure expressions alone, not enough conceptual knowledge could be derived to produce GALEN templates with a sufficient level of detail. As a result, the researchers working on the improvement of the MultiTALE-I tagger were in fact duplicating the work being done by the modellers. Meanwhile, they could not take advantage of the modelling work as during this process relationships between natural language constituents at the one hand, and GALEN-template elements at the other hand were not represented.
To overcome these limitations, the CASSANDRA technique was developed. 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 [12]. 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.

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

Adding surgical procedure concepts collected from classification systems to the GALEN CORE model, is done in a two step approach [13]. 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". Tools such as the SPET have been developed to improve both quality and consistency at this level. The second step is an automatic process performed by the TIGGER (Template Interpreter and Grail GEneratoR). TIGGER transforms the dissections into pure GRAIL.
In Figure 1 it is shown how the rubric "valgiserende osteotomie van humerus" (i.e. incising the humerus to create a valgising position), is represented by means of a dissection [14].

RUBRIC "valgiserende osteotomie van humerus"
ENGLISH_RUBRIC "valgising osteotomy of humerus"
PARAPHRASE "osteotomy of humerus with purpose to create a valgising position"
SOURCE "WCC"
CODE "5-781.21"
MAIN cutting
    TO_ACHIEVE Deed:valgising
        ACTS_ON Pathology:pathological posture
    ACTS_ON Anatomy: humerus

Figure 1: GALEN dissection as intermediate representation for modelling procedures

A dissection has a header which is built up of a number of labelled constituents. The RUBRIC-label introduces the literal expression, in the original language, for the surgical procedure as found in the classification being studied. The PARAPHRASE - label is used to formulate more precisely the intended meaning of the rubric (sometimes also to introduce knowledge that is not grammaticalised in the rubric but implicitly present by virtue of the exact position of the rubric in the hierarchy of the classification system), while the (optional) ENGLISH_RUBRIC - label is a close translation of the original rubric in English. The SOURCE and CODE - labels identify the originating coding scheme, while the COMMENT and CODING_METACOMMENT - labels (both not being used in the

example of figure 1) are respectively used to note general comments or to represent specific instructions to coding clerks implicitly or explicitly present in the classification system. The actual semantic representation of the rubric is introduced by the MAIN-label.

Processing of this dissection by TIGGER results in the GRAIL statement of figure 2.

```
(SurgicalDeed which
    isMainlyCharacterisedBy (performance whichG
        isEnactmentOf ((Incising which playsClinicalRole SurgicalRole) whichG <
            hasSpecificGoal ( (Valgising which playsClinicalRole SurgicalRole)
                whichG LocativeAttribute PathologicalStandingPosture) actsSpecificallyOn Humerus>)))
  hasProjection  (('WCC' schemeVersion 'default') code '5-781.21' 'code');
  extrinsically hasDissectionDetails (DissectionDetails which <
      hasDissectionRubric 'valgiserende osteotomie van humerus'
      hasDissectionEnglishRubric 'valgising osteotomy of humerus'
      hasDissectionParaphrase  'osteotomy of humerus with purpose to create a valgising position'
      hasDissectionCode '5-781.21'
      hasDissectionSource 'WCC' >)
```

Figure 2: GRAIL representation of “valgiserende osteotomie van humerus”

3. The CASSANDRA-tagging technique for dissections

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. Applying CASSANDRA-tagging to the dissection of figure 1, would give the following result:

START
DUTCH_RUBRIC ({valgiserende}$_5$(osteotomie)$_1${[van]$_3$(humerus)$_2$}$_4$)$_{22}$
ENGLISH_RUBRIC ({valgising}$_5$(osteotomy)$_1${[of]$_3$(humerus)$_2$}$_4$)$_{22}$
ENGLISH_PARAPHRASE ((osteotomy)$_1${[of]$_3$(humerus)$_2$}$_4${[with purpose]$_6$((to create)$_8$
{[*]$_{13}$(\a/$_9$({valgising}$_{11}$(position)$_{10}$)$_{12}$)$_{12}$}$_{14}$)$_7$}$_5$)$_{22}$
SOURCE "WCC"
CODE "5-781.21"
MAIN ((cutting)$_{21}$
{[TO_ACHIEVE]$_6$((Deed:valgising)$_7$
{[ACTS_ON]$_{17}$(Pathology:pathologicalposture)$_{18}$}$_{19}$)$_{20}$}$_5$
{[ACTS_ON]$_3$(Anatomy:humerus)$_2$}$_4$)$_{22}$
STOP

Figure 3: GALEN dissection tagged according to the CASSANDRA technique

The general format of a tag is:	“premarker” “item” “postmarker” “label”	
a specific example being :	(osteotomie)1	
where :	“(“	is the “premarker”,
	“osteotomie”	is the “item”,
	“)”	is the “postmarker”, and
	“1”	is the “label”.

There are various possibilities for what can be an “item”, depending on the place in the dissection where the tags appear. At the level of a MAIN-statement, an item corresponds to one of the basic semantic building blocks of the GALEN intermediate representation. At

the level of a RUBRIC- or PARAPHRASE- statement, an item is a word or a group of words used in the statement.
The pre- and postmarkers indicate what kind of semantic building block the item corresponds with.
The labels are a mechanism to mark explicitly the relationships between corresponding items across the various statements in a dissection. Between statements that are expressed by means of natural language (such as RUBRIC, PARAPHRASE, ENGLISH_RUBRIC, etc) these relationships are of type "synonymy" or "translation" depending on whether the related items 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. A main characteristic of the tagging convention is that a closed (and in the future fully "formal") relationship is maintained between the semantics according to the GALEN ontology, and the linguistic phenomena that can be encountered.

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

Table 1 - Characteristics of the CASSANDRA-tagging.

As shown in table 1, 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. In addition, embedding of tags is only allowed according to predefined combinatorial conventions based on semantic grounds.
Some examples of combinatorial conventions are described in table 2.

Tag embedding	Use
{ [a]1 (b)2 }3	a "link" with a concept makes up a "criterion" (e.g. tag 4 in figure 3)
({a}1 (b)2)3	one or more "criteria" applied to a concept makes up a new concept (e.g. tag 22 in figure 3)
((a)1 @b#2 (c)3)4	a coordination of tags of the same type make up a new tag of the same type (in the example a concept)
(\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 (e.g. tag 12 in figure 3)
{ <a>1 {b}2 }3	modification of a criterion gives a new criterion

Table 2 - Conventions for some tag combinations.
To allow for some peculiar linguistic phenomena, additional mechanisms are foreseen.

Tags can appear as discontinuous structures such as in expressions as “by abdominal approach”, where the criterion formed through the link “by approach” applied to the concept of “abdomen”, is grammaticalised by means of a prepositional phrase in which the semantic head is not the noun “approach”, but rather the adjective “abdominal”. Discontinuous tags are represented by closing the first element with the postmarker “&” and by starting the second element with the premarker “$”.

e.g.: {[HAS_APPROACH]$_2$ (abdomen)$_1$ }$_3$ → {[by&$_2$(abdominal)$_1$$approach]$_2$}$_3$

Another linguistic phenomenon that needs to be accounted for is “gapping”, i.e. omissions of phrase constituents that otherwise would have to be repeated without providing useful information to humans. It is obvious that the expression “amputation of left and right hand” does not refer to just one amputation on a hand having the characteristics of being left and right at the same time, but that the expression is “shorthand” for “amputation of left hand and right hand”, or “amputation of left hand and of right hand”, or even, “amputation of left hand and amputation of right hand”. This is clearly different from a syntactically similar phrase such as “cleaning of open and infected wound”. If actually two different wounds would have been cleaned, then one can assume that this event would have been registered differently in order to avoid the ambiguity.
The possible taggings for the phrase “amputation of left and right hand” are described in Figure 4. What tagging scheme is to be used, depends on the situation. For multi-lingual contrastive analysis of expressions with the same meaning, some other scheme might be more appropriate than when expressions are explicitly mapped upon their meaning according to the GALEN model.

1) ((amputation)$_1$ {[of]$_2$(({left}$_3$ (*)$_5$)$_8$ @and #$_6$({right }$_4$ (hand)$_5$)$_7$)$_9$ }$_{12}$)$_{15}$
2) ((amputation)$_1$ {{[of]$_2$({left}$_3$ (*)$_5$)$_8$ }$_{11}$ @and #$_6$ {[*]$_2$({right }$_4$ (hand)$_5$)$_7$ }$_{10}$ }$_{12}$)$_{15}$
3) (((amputation)$_1$ {[of]$_2$({left}$_3$ (*)$_5$)$_8$}$_{11}$)$_{13}$ @and #$_6$((*)$_1$ {[*]$_2$({right }$_4$ (hand)$_5$)$_7$}$_{10}$)$_{14}$)$_{15}$

Figure 4: Possible CASSANDRA tagging schemes for the expression “amputation of left and right hand”, with explicit labelling of gaps indicated by “ * ”.

4. From tagged sentences to a medical treebank

In a traditional sense, a treebank is a collection (also called “corpus”) of sentences upon which both “part of speech” tagging (POS) and “bracketing” is applied. Through part of speech tagging an explicit lexical category is assigned to each word in the corpus, whereas through bracketing the structure of sentences is made explicit. The existence of treebanks is motivated by the overall consensus that significant progress in language processing can be achieved by studying the phenomena that occur in naturally occurring unconstrained materials, and by trying to extract automatically information about language from very large corpora. An example of such a treebank is the Penn Treebank [15], in which the syntactic structure of sentences in general English is made explicit. The total Penn Treebank consists of hundreds of thousands of sentences containing all together several millions of words. Figure 5 shows how tagging and bracketing of the sentence “Casey should have thrown the ball” is realised in this treebank.

```
(S      (NP-SBJ      Casey /NNP)
        (VP    should /VBP
               (VP    have /VB
```

```
(VP     thrown /VBN
        (NP     the /DT ball /NN)))))
```

Figure 5: An example from the Penn Treebank.

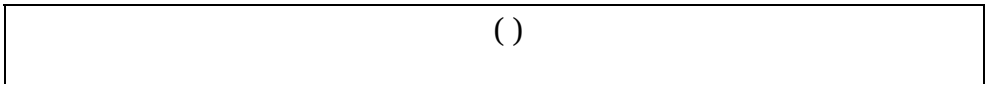
Comparing Figure 3 and 5, it is obvious that there are some important differences between the CASSANDRA treebank and the Penn Treebank.
First, POS-tagging and bracketing are tidily interconnected in the CASSANDRA treebank. POS-tags such as “determiner”, “noun, sg”, etc. are not given. Instead, the brackets at the deepest level of embedding, function as POS-tags. Second, CASSANDRA bracketing is based on semantic rather than on syntactic principles. In addition, the actual structural tags given don’t appear in the sentence itself, but in the MAIN-statement of the dissection. Whereas the Penn Treebank is a collection of sentences, the CASSANDRA treebank consists of blocks of statements, some of which are sentences and others semantic representations of the sentences. Numeric labels are used as a mechanism to mark explicitly the relationships between corresponding items across the various statements within the scope of one block. At the other hand, the scope of the semantic representations in the MAIN-statement covers the entire treebank, i.e. the MAIN-statement constituents are the lingua franca within the CASSANDRA treebank, and in addition serve as link to the GALEN-model. As a consequence, the CASSANDRA treebank “hides” at the same time a linguistic model and a conceptual model, while in addition a formal link between both of them is maintained. This makes the CASSANDRA treebank unique in its kind as according to our knowledge no large semantically tagged treebanks or corpora are currently available, and certainly are not mentioned in a recent world-wide survey [16].

5. The CASSANDRA treebank as a linguistic knowledge repository for healthcare

Though nobody doubts the value of machine readable dictionaries, current trends suggest that work in this area is at a turning point [17]. Processing of large corpora, preferably fully tagged, holds more promises than ever. From such corpora, dictionaries can be created by automatic means, while also grammars can be derived from the annotated text.
The CASSANDRA approach has the advantage that it is strictly independent from known grammatical formalisms, but at the same time conversions are easy to make. It is for instance possible to turn the conventions for tag combination (Table 2) into a set of rewrite rules such that CASSANDRA tagging can be looked at as a phrase structure grammar. For instance, the (“toy”) phrase structure grammar and associated lexicon that can analyse the expression of the English rubric in Figure 1, is presented in Table 3, while the corresponding parse tree is outlined in Figure 6.

of	→	[ ]	[ ] ()	→	{}
osteotomy	→	( )	$\{\}_{0\text{-}n}$ ( ) $\{\}_{0\text{-}m}$	→	( )
valgising	→	{}			
humerus	→	( )			

Table 3: Phrase structure grammar and associated lexicon to analyse the sentence “valgising osteotomy of humerus”.



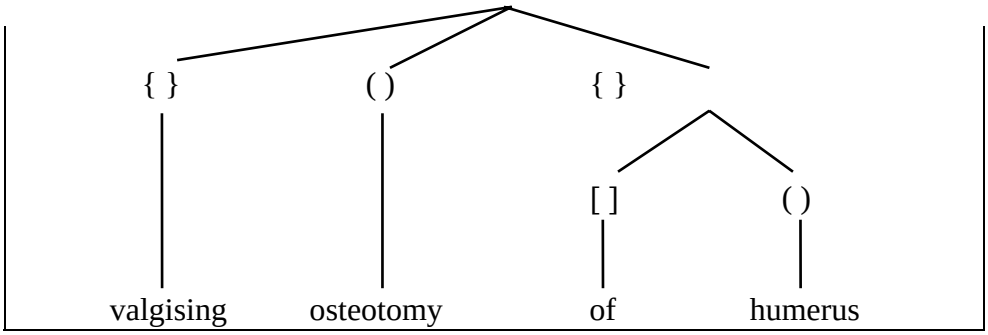


Figure 6: Parse tree for the sentence “valgising osteotomy of humerus” resulting from the grammar of table 3.

Also a dependency grammar [18] can be derived from the CASSANDRA Treebank. According to such a formalism, the tree structure for the same sentence would be:

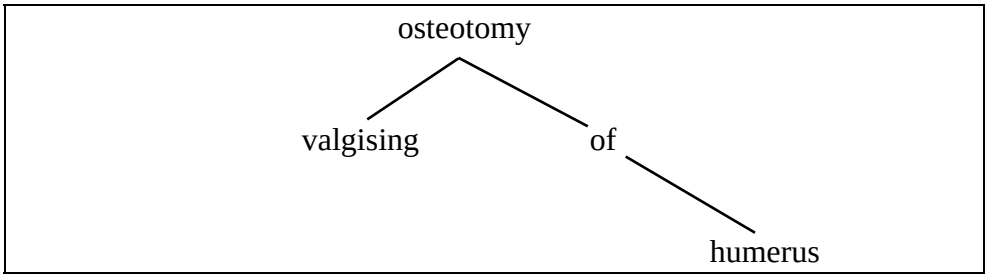


Figure 7: Dependency analysis of the sentence “valgising osteotomy of humerus”.

A dependency grammar can be derived directly from the tag combination conventions by specifying 1) what item within a combined tag links that combined tag to the parent node, and 2) to what kind of parent nodes it may link. This requires an additional notational convention. In Table 4, the dependency grammar able to produce the analysis tree of Figure 7 is presented. The arrows in this grammar are not to be interpreted as part of a rewrite rule, but indicate that the tag at the left may be linked to the tag at the right.

of	→	[ ]	[ ]	→	( )
osteotomy	→	( )	{}	→	( )
valgising	→	{}	( )	→	[ ]
humerus	→	( )			

Table 4: Dependency grammar and associated lexicon to analyse the sentence “valgising osteotomy of humerus”.

Grammatical theories also take into account the notion of “feature”. Semantic features are explicitly present in the CASSANDRA treebank through the GALEN descriptors and links in the MAIN-statement of a dissection. Syntactic features such as “number” and “gender”, and their associated feature values such as “plural” and “masculine”, are not. They only appear in the CASSANDRA treebank after replacing the words in the rubrics with indices pointing to a “traditional” full-form syntactic lexicon.

6. Quality assurance for GALEN modellers and language annotators

Although real benefits from the CASSANDRA approach are only to be expected when large parts of specific medical subdomains are covered, the work being done is immediately useful as a quality assurance mechanism for the modelling centres and language annotators in the GALEN-IN-USE project [5]. For instance, according to the modelling methodology developed, the PARAPHRASE (see Figure 1) should be a close verbal representation of the dissection. CASSANDRA tagging (and subsequent grammatical analysis) of the PARAPHRASE- and MAIN statements revealed that this is not always the case (Fig. 8).

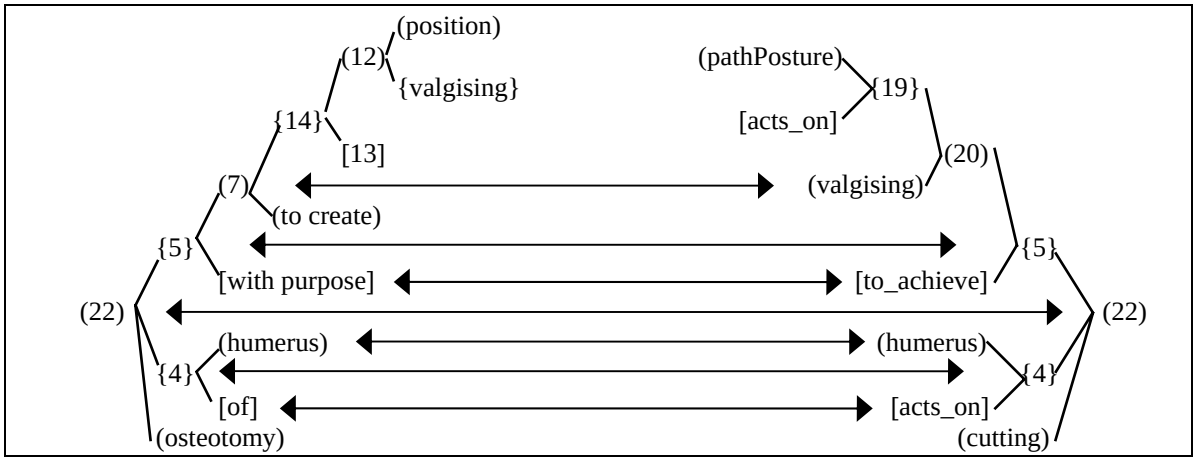


Figure 8: Analysis tree comparison based on the CASSANDRA tagging of the PARAPHRASE and MAIN-statement of figure 1.

In Figure 8, there is only correspondence where bi-directional arcs are drawn. From this it follows that at the level of the sentence, there is indeed correspondence between the sentence and its semantic representation, but not necessarily deeper down the analysis trees. Notice however that “tree correspondence” does not mean that the structure of both trees should be identical. This would mean that semantic structure and syntactic structure are in a 1 to 1 relationship, and that obviously is not the case. The GALEN-concept “valgising” for instance (a leaf in the semantic representation tree) is grammatically expressed as (or “means”) “to create a valgising position” (a branch in the paraphrase tree). However, there should be no “hanging branches (or leafs)” in none of the trees. A hanging branch in the paraphrase tree means that for that particular structure no corresponding GALEN-entity is found, e.g. “osteotomy” and “to create” (at word level) or “valgising position” (at node level). A hanging branch at the semantic representation tree means that conceptual structures are used that not are found in the paraphrase, e.g. “ACTS_ON PathPosture” and “cutting”. These observations are indications (though no proof) that the modelling done is not entirely adequate and should be reconsidered. In the example given, one could argue for instance that the hanging branch “ACTS_ON PathPosture” provides additional world knowledge to the concept of “valgising”.

7. Conclusion

The CASSANDRA tagging technique is a useful mechanism for making the relationships between language and meaning explicit. This is very often forgotten or ignored, even in highly referential works such as UMLS and SNOMED. Within the GALEN-IN-USE project, the CASSANDRA approach has the advantage of recovering this kind of knowledge that otherwise would have been thrown away, while at the same time it is used

for quality assurance. The resulting corpus will in the long run lead to a comprehensive multilingual medical treebank from which specialised grammars and lexicons can be derived in an automatic way.

8. References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

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

[6] 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.

[7] 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.

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

[9] GALEN Consortium. *Guidelines and Recipes for Completing templates*. Internal document VUM02/96 version 1.0.

[10] GALEN Consortium. Links and Templates Summary. Internal document VUM/03/96 version 1.0.

[11] 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.

[12] 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.

[13] JE Rogers WD Solomon AL Rector P Pole P Zanstra E van der Haring. 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.

[14] Nationale Raad voor de Volksgezondheid. WCC-Standaard Classificatie van Medische Specialistische Verrichtingen, versie 2.3, Zoetermeer, 1995.

[15] Marcus M, Santorini B, Marcinkievicz MA. Building a large annotated corpus of English: the Penn Treebank. Computational Linguistics, vol 19, 1993.

[16] Cole R, Mariani J, Uszkoreit H, Zaenen A, Zue V (eds). Survey of the State of the Art in Human Language Technology, CEC-DGXIII-E, Luxembourg, 1995.

[17] Wilks Y, Slator B, Guthrie LM. Electric Words: dictionairies, computers and meanings. The MIT Press, Cambridge, Massachusetts, London, England, 1996.

[18] Hudson R. Word Grammar. Blackwell, London, 1984.