From Natural Language to Formal Language: when MultiTALE meets GALEN.

Ceusters W [a, b] , Spyns P [b]

[a] *Office Line Engineering NV, Hazenakkerstraat 20, B-9520 - Zonnegem, Belgium*
[b] *RAMIT VZW, University Hospital, De Pintelaan 185, B-9000 Gent, Belgium*

In the GALEN project, the syntactic-semantic tagger MultiTALE is upgraded to extract knowledge from natural language surgical procedure expressions. In this paper, we describe the methodology applied and show that out of a randomly selected sample of such expressions, 81% could be analysed correctly. The problems encountered are summarised and areas of further investigation identified.

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 [i]. 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) [ii]. A particular characteristic of the approach is the clear separation of the pure conceptual knowledge from other types of knowledge, including linguistic knowledge [iii], in order to arrive in the future to application-independent medical terminologies [iv]. Although on a theoretical basis the feasibility of these objectives is debatable [v], actual work within the GALEN-IN-USE project shows that on a relatively concise domain such as surgical procedures, distributed collaborative modelling can be achieved over linguistic borders. As could be expected, the process is however extremely slow. Formal "naming" and subcategorisation of new concepts at the one hand, and (in)consistent modelling of natural language expressions using the building blocks of the CORE that already are available, turn out to be the most frequent reasons for discussion. Given the very promising results of the MultiTALE semantic tagger for neurosurgical procedure reports [vi, vii, viii], it was investigated whether or not this manual modelling work could be speeded up by using MultiTALE as an automatic modelling device.

2. Material and methods

100 English surgical procedure expressions were randomly selected from the SNOMED International V3.1 procedure, excluding generic (codes P1-0xxxx) and anaesthetic (codes P1-Cxxxx) procedures. These expressions were then processed by the original MultiTALE tagger. The results were analysed to identify possible shortcomings at the level of the lexicon, the syntactic-semantic grammar and the desired format of the output, i.e. GALEN templates [ix, x]. Based on this analysis, a stepwise lingware refinement methodology was adopted, until a satisfactory number of expressions could correctly be processed.
The purpose of this study was then to investigate 1) whether the high level ontology of GALEN and the representational power of the GALEN surgical procedure templates were sufficiently elaborated for use in natural language understanding, 2) to identify what

additional linguistic knowledge was needed to improve the results, and 3) to investigate whether the SNOMED expressions themselves could unambiguously be understood using the available conceptual and linguistic knowledge.

3. From MultiTALE to MultiTALE II

Prior to any modification, MultiTALE analysed the expression *P1-11E52: closed reduction of fracture of zygoma or zygomatic arch* as an action of type repair which has as direct object a pathology, namely a fracture of zygoma or zygomatic arch (fig.1). The semantic links discovered (action and do), as well as the semantic types (repair, path, anat) have their origin in CEN/ENV 1828:1995 [xi]. In addition, for the individual concepts discovered, the SNOMED International code is given.

```
(1) action repair noun    closed reduction > P1-10E30
(2)  -      -    prep    of
(3) do     path   sg      fracture of zygoma or zygomatic arch
(4)  -     path   sg          fracture of zygoma
(5)  -     path   sg             fracture > M-12000
(6)  -      -     prep           of
(7)  -     anat   sg             zygoma > T-11168
(8)  -      -     coor         or
(9)  -     anat   adjnoun       zygomatic arch > T-11167
```

Fig. 1: MultiTALE analysis of the sentence “closed reduction of fracture of zygoma or zygomatic arch”.

Notice that the correct final results given in lines 1 and 3 originate from an erroneous intermediate processing at lines 8 and 9 where the coordination is attributed at the wrong constituents. This is entirely due to the tagging nature of MultiTALE (as opposed to traditional parsers) according to which only the segmentation at the highest level matters.
With the objectives of GALEN in mind, this approach was no longer feasible as a more detailed analysis was required. The MultiTALE II output of the same sentence is given in fig. 2 and fig. 3.

```
np    {{Closed reduction} of {fracture of {zygoma or zygomatic arch}}}
np       { Closed reduction }
adj         Closed
noun        reduction
prep     of
np       { fracture of { zygoma or zygomatic arch } }
noun        fracture
prep        of
np          { zygoma or zygomatic arch }
noun          zygoma
conj          or
noun          zygomatic arch
```

Fig. 2: MultiTALE II syntactic output of the expression “Closed reduction of fracture of zygoma or zygomatic arch”

```
RUBRIC "Closed reduction of fracture of zygoma or zygomatic arch"
MAIN reduction
      ACTS_ON fracture
            HAS_LOCATION zygoma / zygomatic_arch
      HAS_APPROACH closed
```

Fig. 3: MultiTALE II semantic analysis of the expression “Closed reduction of fracture of zygoma or zygomatic arch”, presented in GALEN-template format.

In order to achieve these results, the following changes to the original system were needed.

1 Implementation of a refined model for surgical procedures

ENV 1828:1995 recognises only four semantic links: *deed*, *direct object*, *indirect object* and *means*. Especially the links *indirect object* and *direct object* turned out to be underspecified for being useful within a natural language understanding environment, and lead to "non-monotonic like" semantic analyses. See for instance:

(1) Injection (*deed*) of antibiotic (*direct object*)
(2) Injection (*deed*) of cyst (*direct object*)
(3) Injection (*deed*) of antibiotic (*direct object*) in cyst (*indirect object*)
(4) Irrigation (*deed*) of cyst (*direct object*) with antibiotic (*means*)

For this reason, more refined links are foreseen such as *has_location*, *has_source*, *has_target*, *has recipient*. As internally in MultiTALE II these links stand in a n-to-1 relationship to the original links, output can still be given according to the specifications of the ENV. However, in order not to duplicate the work of the modellers in the GALEN project, the conceptual model was not more enhanced than needed for an unambiguous interpretation of the expressions, leaving out the details required for generation purposes. In addition, only that part of the GALEN ontology that surfaces grammatically in the expressions, was incorporated [xii].

2 Implementation of a concept hierarchy

The MultiTALE tagger was directly based on the "flat" concept model of ENV 1828:1995, lexemes being encoded directly as *surgical_deed*, *anatomy*, *pathology* or *instrument*. To resolve certain linguistic ambiguities, a hierarchical model was needed. The relevant parts of the hierarchy needed to analyse the sentence of figure 3 , and the restrictional constraints on how some concepts may be linked, are outlined in figure 4.

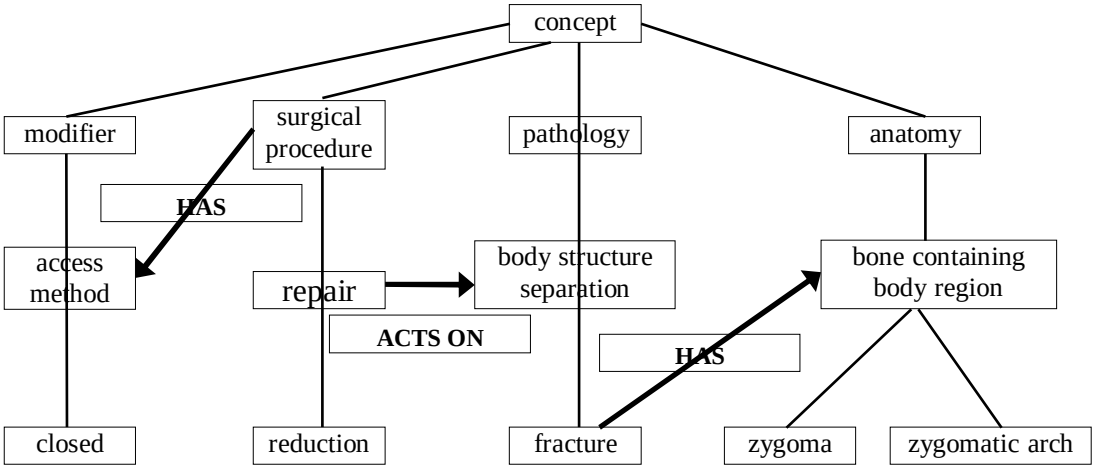

Fig 4: Relevant part of the concept hierarchy for the sentence *closed reduction of fracture of zygoma or zygomatic arch.*

3 Implementation of mechanisms for knowledge discovery

As MultiTALE II is designed to enrich the GALEN CORE and linguistic annotation modules (semi)automatically, mechanisms had to be foreseen for dealing with unknown words in the input. This was achieved using a bottom-up parsing strategy where both syntactic and semantic configurations limit each others possible interpretations. In fig 5, the sentence "injection of xyz" (were xyz obviously is an unknown word) is analysed by MultiTALE II with one possible syntactic solution (xyz being a noun), and four possible

semantic interpretations. First, xyz might be a body part, body region or pathology in which a not specified chemical is injected, as in *P1-10542: injection of ligament*. In these three cases, the HAS_DESTINATION semantic link applies. Next, xyz might be the chemical itself, with no destination specified, as in *P1-05027: injection of gas.*

np	{ Injection of xyz }	RUBRIC "Injection of xyz"
noun	Injection	MAIN injection
prep	of	ACTS_ON xyz : chemical
noun	*xyz	
	RUBRIC "Injection of xyz"	
	MAIN injection	
	ACTS_ON chemical	
	HAS_DESTINATION xyz : body_part / body_region / pathology	

Fig 5: Syntactic and semantic analysis of the sentence “injection of xyz”.

4. Results

Out of the 100 randomly selected expressions, 10 could not be processed by MultiTALE II. For 7 of them, the required concepts or links were not yet available in the GALEN template-formalism, clearly a reason for failure outside the responsibility of MultiTALE. Of the remaining three, two showed peculiar (a)grammatical configurations while the other one contained deictic references and ellipsis, linguistic phenomena for which no grammar rules are currently implemented (*P1-B9846: Bilateral repair of inguinal hernia, one direct and one indirect*). Of the 90 expressions that could be processed, 73 (81%) were analysed correctly, giving the only one possible interpretation, 58 of which by using exclusively the links foreseen in the GALEN template formalism (an intermediate representation developed in order not to confront the domain modelling experts with the complexity of the GRAIL language), while for the remaining 15, additional semantic links were introduced. It would have been possible to map these extra links to the “garbage”-link HAS_OTHER_FEATURE that is allowed in the templates, but we choose deliberately for not doing so in order to preserve the depth of the interpretation. 17 expressions led to multiple interpretations, 48 all together. Of those 48, 36 (75%) could be judged being correct.

5. Discussion and conclusion

The results presented in this paper reflect not the final desired outcome of MultiTALE II, but are rather to be seen as a first evaluation of the actual stage of the system, with further improvement in mind. Ambiguities in the input phrases was the most important reason for multiple interpretations. E.g. *P1-A1122: Decompression of orbit only by transcranial approach*, where “only” can refer to the orbit (nothing else being decompressed), to the decompression (nothing else than a decompression being done on the orbit), or to the approach (no other approach allowed for giving this code). Coordination also led often to multiple interpretations, though the semantic constraints prevented all possible syntactic combinations, as can be seen in fig 2, where syntactically a possible bracketing would have been: *{{Closed reduction} of {{fracture of zygoma} or zygomatic arch}}.* This possible syntactic solution is however not retained on semantic grounds. Failure to reach an adequate interpretation was due to one of three reasons. For few sentences, the representational power of the GALEN-templates was not sufficient. It is for instance not yet possible to represent coordination amongst different semantic links that apply at the same time to one

concept, e.g. *P1-2682B: repair of internal or complex fistula of trachea*, where "internal" and "complex" specify two different features of "fistula". Also, the GALEN templates allow numbers to be linked to concepts using the HAS_NUMBER link, but quite often, an exact number cannot be deduced from the expression, as just a plural is given. See *P1-7AC34: Lysis of adhesions of spermatic cord*, where one can only infer that there must be more than one adhesion. For some other sentences, specific surface linguistic constructs turned out to be problematic. E.g. in *P1-19B05: Primary suture of ruptured ligament of ankle, collateral*, "collateral" obviously specifies "ligament", but no grammar rule could yet be implemented in such a way that this sentence would be analyzed correctly without introducing erroneous output for other sentences such as *P1-40141: Incision and drainage of hematoma, complicated*. Similar difficulties are caused by coordinated multiword units upon which ellipsis is applied, as in *P1-21A08: … rhinoplasty with lateral and alar cartilages …*. A third reason for incorrect results, is the lack of detailed anatomical knowledge such as the one required for correctly parsing *P1-17A26: Tenodesis for proximal interphalangeal finger joint stabilization*, where the system must know that "interphalangeal" refers to "joint" and not to "finger", in contrast with "abdominal wall mass" were "abdominal" refers to "wall".

The main conclusion of this work is that it is indeed feasible to develop a syntactic-semantic parser that quite satisfactorily translates natural language expressions into a predefined formalism for further processing. However, in order to be able to extract new knowledge from texts, a certain amount of background knowledge, both conceptual and linguistic, must be available. The precise boundaries of each of them are not yet clear, what requires further investigations.

6. References

i. 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.

ii. 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.

iii. 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.

iv. 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.

v. Ceusters W, Deville G, Buekens F. The chimera of purpose- and language-independent concept systems in healthcare. In Barahona P, Veloso M, Bryant J (eds.) *MIE 94 Proceedings* 1994, 208-212.

vi. 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.

vii. 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.

viii. 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.

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

x. GALEN Consortium. *Links and Templates Summary*. Internal document VUM/03/96 version 1.0.

xi. CEN ENV 1828:1995. Medical Informatics - Structure for classification and coding of surgical procedures.

xii. Ceusters W, Deville 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 (in press).