A Mixed Syntactic - Semantic Grammar for the Analysis of Neurosurgical Procedure Reports: the Multi-TALE Experience.

W. Ceusters [1,2] , G. Deville [3]

[1] Language & Computing NV, Hazenakkerstraat 20, B-9520 Zonnegem
[2] RAMIT VZW, UZ - Gent, Med. Informatica, De Pintelaan 185, B-9000 Gent
[3] FUNDP, Ecole de langues vivantes, Rue de Bruxelles, 61, B-5000 Namur

Abstract: The purpose of the Multi-TALE syntactic-semantic tagger is to identify surgical procedure expressions in full text reports and to assign semantic tags to its composing elements according to CEN ENV1828:1995. In this paper the general architecture of the English tagger is described, putting emphasis on the syntactic-semantic grammar that has been designed for the sublanguage used in neurosurgical procedure reports. The grammar consists of a set of rules that are to be used in a bottom-up approach to combine words and word groups in the sentences to be analysed, to more complex elements, up to the level of a predicate, or one of its arguments. The validation procedure of the tool is then shortly discussed. Finally, a comparison with other systems is given both in terms of architecture and results.

1. Introduction

For the last decade there is a growing need for automated natural language processing tools in medicine. But natural language processing, even on a medical subdomain such as neurosurgery, is so complex that it can only be tackled with a divide-and-conquer strategy, which implies that different tools or submodules are to be developed for the different stages of natural language processing.

Multi-TALE is a small-scale project that fitted into this view. The project resulted in two syntactic-semantic taggers for English and Dutch neurosurgery reports. The function of a tagger is to perform the first and essential pre-processing stage for any natural language processing application: the labelling of words with their grammatical (sub-)categories, only taking into account a limited context. In addition to this syntactic labelling, the Multi-TALE taggers also provide semantic tags to words and word groups found in the texts. More specifically, semantic decoration is provided on the basis of the model for surgical procedures as described in CEN ENV1828:1995. In this paper, only the English tagger is discussed.

2. CEN ENV1828:1995: Structure for Classification and Coding of Surgical Procedures

A project Team (PT002S) of CEN/TC251 (Technical Committee 251 on Medical Informatics of CEN (Comité Européen de Normalisation) was given the task to develop a structure for classification and coding of surgical procedures. The purpose of this standard is to *identify the concepts within the text of surgical procedure notes and to structure them to represent the concepts and their internal relations* [1]. According to this standard, a surgical procedure is conceptually composed of a *surgical deed* (i.e. deed that can be done by the operator to the patient's body during the surgical procedure) which is semantically linked to the concept fields *human anatomy*, *pathology* and *interventional equipment*. Potential semantic links are *direct object* (referring to that on which the surgical deed is carried out), the *indirect object* (referring to the site of the surgical deed) and the *means* (referring to the means with which the deed is carried out).

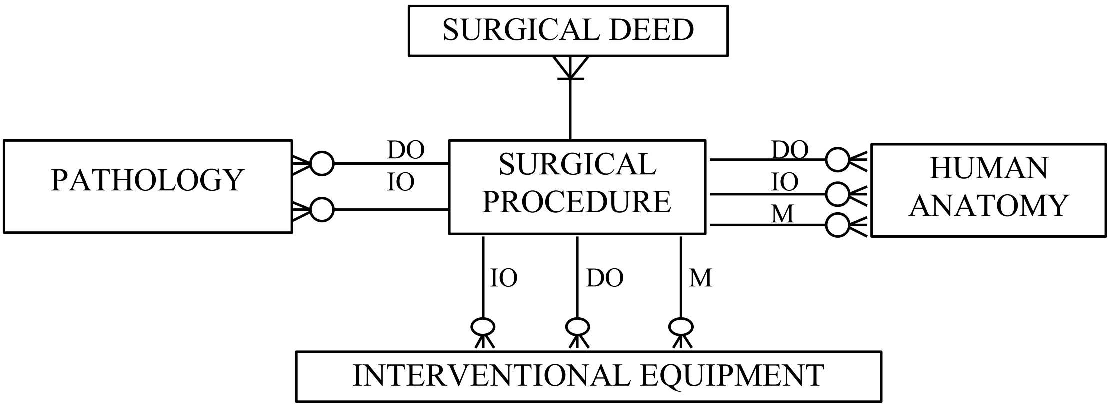

Fig. 1: Semantic types and central semantic links in ENV1828:1995

Although the standard was developed for the description of elementary surgical procedure expressions as they can be found in classification and coding systems, one of the hypotheses of the Multi-TALE project was that the same structure could be used to represent the particular tasks described in full text reports. To test this hypothesis, a functional approach to the medical language was adopted, recognising that the language structure is not to be separated from language use. The functional dimension of natural language has been convincingly introduced by Dik in theoretical linguistics [2], and successfully applied and adapted by Deville for the modelisation of administrative language [3]. By using the same approach, we could semantically model surgical procedure expressions as structures that consist of a predicate (surgical deed) with an adequate number of terms, specified by means of a semantic function that fulfills the role of argument of that predicate. We identified the semantic links *direct object*, *indirect object* and *means* as the expression of such semantic functions labelled *case roles*, that can be ascribed to the arguments of a predicate (i.e. concept of surgical deed) .We also redefined the notion of combinatorial rule as a specification at surgical deed level, constraining on semantic grounds the case roles that are prototypical of a given surgical deed. We finally defined 12 classes of surgical deeds on the basis of common syntactic and semantic properties by appealing to the notion of *predicate primitive*. To each class of surgical deed corresponds a cluster of arguments with specific syntactic-semantic constraints [4]

3. General Architecture of the English Multi-TALE tagger

The English Multi-TALE prototype is a modular system consisting of several programs which operate together under Microsoft's Word for Windows, version 6.0. A schematic overview of the analysis flow, together with the required linguistic resources, is given in figure 2.

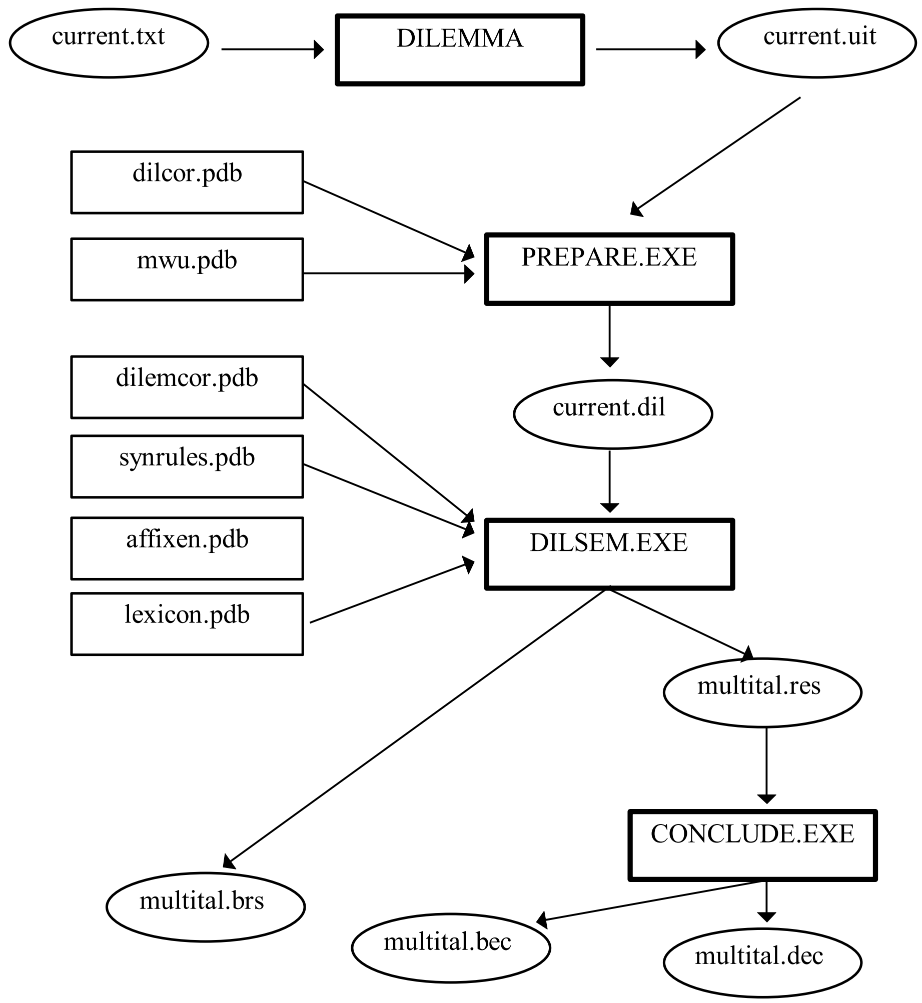

Fig. 2: General architecture of the English Multi-TALE tagger. Bold rectangular boxes represent the programme modules, light rectangles refer to the linguistic resources that are part of the system, and round boxes represent the intermediate and final output files generated.

A typical session with the Multi-TALE prototype starts by selecting in a word processor the portion of text to be processed. The sentences selected are temporarily stored in the file *current.txt* that becomes the input for the syntactic tagger DILEMMA [5]. This is a rule-based parts of speech tagger developed for general language. When applied to the particular sublanguage of neurosurgical procedure reports, its accuracy turned out to be 85% maximum [6]. We received DILEMMA as a closed system without the possibility to improve its performance. For this reason, an intermediate module (*prepare.exe*) was developed. The task of this module is twofold: first to correct certain systematic errors - independent of any context within the sentence being processed - produced by DILEMMA, and second, to allow the processing of multi-word units (e.g. *sella tursica*, *cerebellar artery*) as one entity. The output of this module is then passed to the syntactic-semantic tagger (*dilsem.exe*). This part of the system is developed on a language- and domain independent mode. The linguistic and semantic knowledge it exploits is stored in separate linguistic knowledge bases (LKB). A first LKB corrects context-dependent mistakes made by

DILEMMA. A second LKB contains the semantic lexicon while a third one holds the grammar.

Texts are processed by the semantic tagger sentence by sentence. The analysis of any given sentence proceeds in a deterministic way, except for lexicon lookup. When for instance in a sentence two words are found that each can have three meanings or alternative syntactic characteristics, 9 possible interpretations of the sentence are processed. The last module of the Multi-TALE prototype, *conclude.exe*, uses heuristic techniques to point out in such case the most likely interpretation.

4. The Multi-TALE reference lexicon

4.1 Macro-structural entities

A macro-structural entity of the MULTI-TALE augmented reference lexicon is defined here as a lexical entry as it is coded in the lingware of the application. This macro-structural entity can be of three basic types: bounded morphemes, words, and word groups. We will discuss these three types on the basis of linguistic and operational criteria. The examples illustrating this section are taken from the corpora of English and Dutch corpora of neurosurgical procedures [7].

4.1.1 Sublanguage bounded morphemes

The first type of macro-structural entities consists of sublanguage bounded morphemes. These morphemes are :

- each word stem under its different variation patterns, provided that these variations are unpredictable. In the following examples, the stems 'skull' and 'crani' are encoded as two different lexical entries (the alphanumeric string preceding the expressions are identification tags of corpus extracted expressions):

  SDY_E_88_011.2 *The resultant <u>crani</u>ectomy was then extended as a 2 cm strip a long fused metopic suture down to the <u>skull</u> base in the region of the crista galli.*

- the predictable orthographic variants of a given word. In the following example, the term 'haemostasis' can also be written as 'hemostasis'. In that case two different lexical entries are foreseen :

  SDY_88_E_010.7. *<u>Haemostasis</u> was achieved without difficulty and without use of Surgicel.*

Unpredictable orthographic variants (i.e. spelling errors) can only be taken into account by means of a matching algorithm that would check the user's input before processing it, and are not part of the lexicon.

4.1.2 Single words

The second type of macro-structure consists of the sublanguage's single words in their lemmatised version.

4.1.3 Word groups

The third type of macro-structure consists of word groups (in their lemmatised form) that are considered as one single conceptual entity. Note that the structure of these word groups cannot be computed. Examples of such compounds are respectively “spinous process(es)” and 'Infant delta Shunt' in the two following report abstracts :

SDY_88_E_055.1. *Under general orotracheal anaesthesia, in the prone position, through a midline incision, the muscles were* *separated from the spinous processes and laminae of L2 to the sacrum and levels identified by reference to the sacrum.*

SDY_88_E_037.5 *Peritoneum exposed through an epigastric incision and an Infant Delta Shunt, Performance Level I was tunnelled between the two incisions.*

4.2 Micro-structure of lexical entries

Each entry in the lexicon has the following structure, consisting of a set of 5 fields with syntactic semantic information.

```
<lemma>
  Meaning: <snomed-code>
  Superconcept: <supertype>
  Superord. Concept: <subtype>
  Main Syntactic Cat: <syntactic category>
 [ Roles:
              <role_desc 1>
              <role_desc 2>
              ...
              <role_desc n>
   Preps:
        <prep_desc 1>
        <prep_desc 2>
        ...
        <prep_desc n> ]
```

Fig. 3: micro-structure of entries in the Multi-TALE lexicon.

The <snomed-code>-field contains the SNOMED-code of the entry when available.

The <supertype> fields specifies the main semantic type to which the entry belongs. Potential values are the central semantic types as defined in ENV1828:1995.

The <subtype>-field refers mainly to the occurrences of concept types for surgical deeds such as *install*, *remove*, *inspect*, etc.

The <syntactic category> field specifies the syntactic categorial (and subcategorial) information regarding the lemma. Examples of such categories are Noun, Adjective, Verb, Adverb. In case of surgical deeds, potential entries for this field are “verb” and “noun”. The syntactic category does not of course influence the semantics of a lemma, but depending on whether a surgical deed syntactically is realised by means of a verb or a noun, specific prepositions may be used or not.

The <role>- field (only applicable for surgical deeds) indicates for the surgical deeds the prototypical case structure that is in relation with the deed in question, together with some semantic constraints -expressed in terms of other supertypes or specific concepts- on each of the cases. Theses cases are DO (direct object), IO (indirect object), and MEANS (manner/means).

The <prep>-field indicates the case markers (prepositions, or prepositional locutions) that univoquely mark the term they refer to with a given case. Note that one lemma can have more that one case marker, as the example of the surgical deed “removal” that combines with the preposition 'from' (marking an IO -indirect object-) and 'of' (marking a DO -direct object).

5. The syntactic-semantic grammar

SYNRULES.PDB contains the Multi-TALE syntactic-semantic grammar for English neurosurgery procedure reports. The purpose of this grammar is to combine in a bottom-up approach syntactic elements in the sentences[1] to be analysed, to more complex elements, up to the level of a “clause”.

A <u>sentence</u> is considered to have the following structure:

<sentence> ::= { <segment> }* [<segmentor> { <segment> }*]
<segment> ::= { <clause> }*
<clause> ::= { <intercon> }*
<intercon> ::= { <token> }*

E.g.: The <u>sentence</u> *A big fatty tumour was removed rapidly from the brain and the hole filled with pieces of artificial tissue*, is composed of the <u>segments</u> *A big fatty tumour was removed rapidly from the brain* and *the hole filled with pieces of artificial tissue*. The two segments are connected by the <u>segmentor</u> *and*. The first segment contains the <u>clauses</u> *a big fatty tumour*, *was removed*, *rapidly*, and, *from the brain*, while the second segment is composed of the clauses *the hole*, *filled*, *with pieces of artificial tissue*.

This aggregation process is shown in the file MULTITAL.BRS of the developers version of Multi-TALE, as outlined in figure 4.

```
SENTENCE 001, SOLUTION 001, SEGMENT 001: remove
do      path     detnoun 4        A big fatty tumour            <clause>
 -        -      art     -            A                         <token>
 -      path     adjnoun 2            big fatty tumour          <intercon>
 -        -      adj     -                big                   <token>
 -      path     adjnoun 2                fatty tumour          <intercon>
 -        -      adj     -                    fatty             <token>
 -      path     sg      -                    tumour            <token>
action  remove   papa    11a      was removed                   <clause>
 -        -      past    -            was                       <token>
action  remove   papa    -            removed                   <token>
 -      velocity adv     -        rapidly                       <token>
 -        -      prep    -        from                          <token>
io      anat     detnoun 4        the brain                     <intercon><clause>2
 -        -      art     -            the                       <token>
 -      anat     sg      -            brain                     <token>


SENTENCE 001, SOLUTION 001, SEGMENT 002: fill
do      mcr      detnoun 4        the hole               <clause>
 -        -      art     -            the                       <token>
 -      mcr      sg      -            hole               <token>
action  install  past    -        filled                        <clause>
 -        -      prep    -        with                          <token>
m       anat     adjnoun 16mcrOf  pieces of artificial tissue   <intercon><clause>
 -      mcr      pl      -            pieces                    <token>
 -        -      prep    -            of                        <token>
 -      anat     adjnoun 2            artificial tissue         <intercon>
 -        -      adj     -                artificial            <token>
 -      anat     sg      -                tissue                <token>
```

Fig. 4: bottom-up syntactic-semantic tagging in Multi-TALE

Each grammar rule has the following general format:

<u>IF</u> you find in the sentence to analyse a <token> or <intercon> (further called *left-element*) with a syntactic category occurring in the list <leftsyn>, and a semantic type featuring in the list <leftsem>,

[1] The “decision” on what parts of a full-text are to be considered sentences, is taken earlier in the process.

[2] In this version of MULTITAL.BRS, clauses starting with a preposition are displayed over two lines.

ANDIF (optionally) this <token> or <intercon> is followed by a <token> or <intercon> (further called *mid-element*) with a syntactic category occurring in the list <midsyn>, and a semantic type featuring in the list <midsem>,

ANDIF you find also after that a right <token> or <intercon> (further called *right-element*) for which applies respectively <rightsyn> and <rightsem>,

ANDIF before *left-element* there is (optionally) within a given <distance> a <token> or <intercon> with a syntactic category occurring in the list <left-context>,

ANDIF after *right-element*, described by <rightsyn> and <rightsem>, there is (optionally) within a given <distance> a <token> or <intercon> > with a syntactic category occurring in the list <right-context>,

<u>**THEN**</u> take the *left-element*, *mid-element* and *right-element* together as indicated by the <demon> of <midsyn>,

AND give the resulting <intercon> as syntactic category what is expressed in <syntactic_result>

AND finally assign the resulting <intercon> the semantic type as specified in <semantic_result>.

When an action can be performed on consecutive elements within a sentence, three <demon>s can be applied:

"join": take the elements together as they occur in the sentence
"reloc": join *left-element* and *right-element*, and put *mid-element* at the right of it
"break": insert a <segmentor> before *mid-element*

For instance, the rule:

```
sc("16start","@1",
    [scope("break","",1),scope("segm","",2)],
    ["pred","detnoun","noun","ind","sg","pl"],
    join(["prep"]),
    ["sg","pl","detnoun","noun","adjnoun","pers","prop","adjprop"],
    [], ["nil","anat","path","inte","mcr"], [], [], "@1").
```

specifies that at the beginning of a sentence, or when occurring 1 or 2 places after a <segmentor>, noun-phrases separated by a preposition, may be grouped together to form one constituent that receives the syntactic and semantic features of the left-element. This rule will cause in the sentence "*the catheter in the third ventricle was withdrawn"* the combination of "*the catheter*", *"in"* and *"the third ventricle"*, while in the sentence, *"I inserted the catheter in the third ventricle",* this combination will not be allowed as the rule would not fire.

All the rules are thoroughly documented following a rigorous schema. As a consequence, updates to the grammar can be implemented without mayor difficulties. The following documentation tags are provided:

- **RULE_ID :** the original identification label (digital or alphanumeric code) that has been used in the development of the grammar itself. As the rule set has evolved in the course of development, this label does not systematically indicate any sequential or logical order of the rules.
- **RULE_SHORTHAND :** as the RULE_ID gives few indications on the type and content of the rule, we identified each rule with an transparent shorthand (string of characters) on type and role content. This shorthand can appear in the multi-tale output in order to easily trace and check the rules that have fired.
- **RULE_FORMAT :** the rule as expressed under its original format
- **RULE_DESCRIPTION :** the rule as interpreted in a condition action structure
- **EXAMPLE_REF :** indicates the reference of the example using the tagging of deliverable D2-1 : English and Dutch Representative Base-Corpora of Neuro-surgical Procedure Reports [8].
- **EXAMPLE_TXT :** indicate the example taken from the English and Dutch Representative Base-Corpora of Neurosurgical Procedure Reports.
- **EXAMPLE_OUTPUT:** indicates the example in its Multi-Tale output format
- **COMMENTS :** motivates the rule both in a functional and linguistic perspective.

6. Validation

Two types of validation have been carried out: one to test the intermediate performance of the system with fine-tuning as primary objective, and a second one to assess the results of the final modifications [9]. Ten surgical reports (138 sentences) from the corpus, five having been used for the development of the syntactic-semantic rule-base (training sample), and five for which this is not the case (testing sample), have been manually validated. Two human experts (physicians) were used as gold standard. Recall (number of entities retrieved and relevant over number of relevant entities in the report) and precision (number of entities retrieved and relevant over number of entities retrieved (only for second validation)) were calculated separately for testing and training sample. Calculations were based on the correct recognition of syntactic entities such as sentences, segments (surface form of the predicates), clauses (surface form of predicate arguments), simple and complex noun phrases and verb phrases, as well as on semantic information (types and semantic links correctly identified). In total 2139 syntactic units (to be mapped into 6 categories) and 857 semantic-contextual entities (to be mapped into 8 categories) were to be retrieved.

Table 1 shows the results for both validations (* indicates test not performed). The first validation revealed an acceptable performance for syntactic tagging (except for complex noun phrases) and semantic type recognition, with only very modest results for case assignment. In addition, recall in the training sample appeared to be much higher than in the test sample. Fine-tuning of the system turned out to be very effective, and led to syntactic recall for the test sample of 95.7% (precision being 95.7% also) and 89.3% for semantic recall with a precision of 94.8%. Semantic labelling still appeared to be more successful than case-assignment (recall 93.4% versus 77.4%, 60.0% and 77.8%, $p < 0,01$).

	First Evaluation		Second Evaluation			
	Known	Unknown	Known		Unknown	
	Recall	Recall	Recall	Precision	Recall	Precision
Sentences	100	99	100	100	98.7	100
Segments	88	86	98.0	88.3	98.5	92.3
Clauses	91	80	91.9	92.6	95.2	93.3

Simple NP's	93	85	92.8	100	94.3	100
Complex NP's	79	56	88.6	93.9	86.7	100
VP's	93	85	92.6	98.9	95.2	100
Tot. Syntax	91	82	93.3	94.4	95.7	95.7
Deeds	*	*	96.5	94.3	98.1	97.1
Anatomy	*	*	93.9	95.8	98.4	98.4
Pathology	*	*	100	88.6	94.7	100
Instrument	*	*	87.5	97.2	71.1	96.4
Tot. Sem. Types	90	71	94.6	94.2	93.4	97.8
Action	97	82	96.5	90.1	97.1	93.5
Direct Object	84	69	83.3	88.7	77.4	92.9
Indirect Object	78	61	75.0	80.0	60.0	70.6
Means	68	50	70.6	100	77.8	93.3
Tot. Semantics	83	71	91.3	92.0	89.3	94.8

Table 1: Validation results of the Multi-TALE syntactic-semantic tagger.

7. Discussion

Some existing systems can be compared with Multi-TALE from a functional and implementational viewpoint. The LSP system was originally designed to extract factual information from medical reports [10]. The medical sublanguage is viewed by this system as consisting of 6 information formats onto which a total of 54 semantic classes (represented in the lexicon) can be mapped. The system uses a parser that can deal with incomplete analyses. When it was used to find 13 important details of asthma management in a total number of 31 discharge summaries (testing set), recall appeared to be 82.1% (92.5% counting only omissions instead of errors) and precision 82.5% (98.6%) [11].

Haug et al. report recall and precision rates of 87% and 95% for detecting clinical findings in 839 chest x-ray reports by using SPRUS [12], and rates of 95% and 94% respectively for the detection of diagnoses [13]. SPRUS is mainly semantically driven, and is not able to exploit syntactic information. Hence complex noun-phrases and whole sentences cannot be processed [12].

The CAPIS system was able to recall 92% of the relevant physical findings (156 in total in 20 reports on patients with gastro-intestinal bleeding), with a precision of 96% [14]. CAPIS uses a finite-state machine parser that is specifically designed for the more structured parts in medical narrative such as the clinical findings section.

8. Conclusion

The main conclusion from this work is that tagging, as a more simple procedure compared to parsing, may be an effective strategy to extract factual information from clinical narrative. The main advantage is that a complete parse of a sentence is not needed to map the semantic contents of a sentence onto a predefined classification or conceptual model. But at the same time, the approach has its limits. Full natural language understanding cannot be realised by a semantic tagger alone. It is our believe that when full understanding of a sentence is required, semantic taggers are however useful as they can enhance the performance of a parser used subsequently by rapidely eliminating alternatives that only after a long processing time would turn out not to lead to an acceptable solution.

9. Acknowledgements

The Multi-TALE project was partly funded by the Commission of the European Communities in MLAP-II contract 93-04.

10. References

1. CEN ENV 1828:1995. Medical Informatics - Structure for classification and coding of surgical procedures.
2. Dik S. A Theory of Functional Grammar, Foris, Dordrecht, 1989
3. Deville G. Modelisation of Task-Oriented Utterances in Man-Machine Dialogue System. Ph.D. Thesis, University of Antwerp, 1989
4. Deville G, Ceusters W, Mommaerts JL, Martin W. Semantic Analysis of the CEN/TC251/PT002S-model for Surgical Procedures; Applicability and Extendibility. Multi-TALE deliverable D2-2, 25-06-94.
5. Paulussen, H and Martin W. DILEMMA-2: a lemmatizer-tagger for medical abstracts, in Proceedings of the Third Conference on Applied Natural Language Processing (ACL), Trento, 141-146, (1992).
6. Mommaerts JL, Ceusters W, Deville G. Zijn taggers voor algemene taal bruikbaar voor medische subtalen ? Een case studie met Dilemma. In: Beckers WPA, ten Hoopen AJ (eds) Proceedings of MIC'94, Velthoven, The Netherlands, 25-26/11/94, 283-290, 1994.
7. Deville G, Ceusters W, Kramer A, Martin W. Multi-TALE Deliverable D2-3: Reference Lexicon for Neuro-surgery Reports in English as Syntactic-Semantic Knowledge Base for the Multi-TALE prototype, 1994-09-30.
8. Mommaerts JL, Ceusters W, Deville G, Martin W. Multi-TALE- Deliverable D2-1: English and Dutch Representative Base-Corpora of Neurosurgical Procedure Reports.
9. Ceusters W, Deville G, De Moor G. Automated extraction of neurosurgical procedure expressions from full text reports: the Multi-TALE experience. In: Brender J. et al. (eds.) Medical Informatics Europe '96, 154 - 158, IOS Press 1996.
10. Sager N, Friedman C, Lyman MS: Medcial Language Processing: Computer Management of Narrative Data. Reading, MA: Addison - Wesley, 1987.
11. Sager N, Lyman MS, Bucknall C, Nhan N, Tick LJ: Natural language processing and the representation of clinical data. J. Am Med Infomatics Assoc, 1, 142-160, 1994.
12. Ranum DL. Knowledge based understanding of radiology text. Proceedings of the 12th Annual Symposium on Computer Applications in Medical Care, Washington, DC, 141 - 145, 1988.
13. Haug PJ, Ranum DL, Frederick PR: Computerized extraction of coded findings from free-text radiologic reports. Radiology, 174, 543 - 548, 1990.
14. Lin R, Lenert L, Middleton B, Shiffman S. A free-text processing system to capture physical findings: Canonical Phrase Identification System (CAPIS). Proc-Annu-Symp-Comput-Appl-Med-Care. 1991: 843-7