

**THE INTEGRATION OF  
LINGUISTIC AND DOMAIN  
SPECIFIC KNOWLEDGE:  
CAT2 WITHIN ANTHEM**

Oliver Streiter and Antje Schmidt-Wigger  
IAI Martin-Luther-Straße 14  
66111 Saarbrücken Germany  
fax: +49-681-39 74 82  
catler@iai.uni-sb.de

Proceeding of  
**Conference on Health  
Telematics'95**  
Ischia, Italy.

**MT News International 11 1995**

Newsletter of the International  
Association for Machine Translation.

# THE INTEGRATION OF LINGUISTIC AND DOMAIN SPECIFIC KNOWLEDGE: CAT2 WITHIN ANTHEM

Oliver Streiter and Antje Schmidt-Wigger  
IAI Martin-Luther-Straße 14  
66111 Saarbrücken Germany  
fax: +49-681-39 74 82  
catler@iai.uni-sb.de

The aim of the LRE project ANTHEM is to develop a prototype of a natural language interface that allows users of Healthcare Information Systems to enter medical diagnostic expressions in a multilingual environment. Within ANTHEM the CAT2 MT system is used to analyse Dutch and French diagnostic expressions, translating these (a) into German, Dutch and French and (b) into a semantic representation which is then passed to the ANTHEM Expert System for automatic coding in ICD. For both purposes the interlingual approach has been adopted. The interlingua consists of a set of basic word-concepts identified by their SNOMED code and a limited set of semantic relations which link word-concepts to form larger concepts. In this paper we shall explain in detail how general linguistic, sublanguage-specific and domain-specific knowledge is integrated into one framework of analysis, relying on a rich system of lexical information which becomes operative through a few principles of syntactic and semantic composition.

# THE INTEGRATION OF LINGUISTIC AND DOMAIN-SPECIFIC KNOWLEDGE: CAT2 WITHIN ANTHEM

Oliver Streiter and Antje Schmidt-Wigger  
IAI Martin-Luther-Straße 14  
66111 Saarbrücken Germany  
fax: +49-681-39 74 82  
catler@iai.uni-sb.de

## 1 Introduction to ANTHEM

The aim of the LRE project ANTHEM<sup>1</sup> is to develop a prototype of a natural language interface that allows users of Healthcare Information Systems to enter medical diagnostic expressions in Dutch or French ([Ceusters et al.1994b]). Within ANTHEM the CAT2 MT system<sup>2</sup> is used to analyse these expressions, translating them into (a) German, Dutch and French and (b) a semantic representation which then is passed to the ANTHEM Expert System for automatic coding in ICD.

A first prototype running on a Unix Workstation was realized in 1994 (cf. [Ceusters et al.1994a]). Currently the lexical coverage and the functionalities of the system are being extended in order to allow for an application in a real life medical setting. The portation to DOS is under way.

## 2 The Construction of an Interlingua

It has been known for a long time in MT theory that the interlingual approach is the most promising in multilingual systems. Since in ANTHEM at least three natural languages and one language independent representation are involved, the interlingual approach seems to be the most natural to follow. In order to circumvent the main problem - the preliminary construction of an interlingual

---

<sup>1</sup> The ANTHEM consortium consists of RAMIT Ghent (coordinator), FUNDP Namur, IAI Saarbrücken, CRP-CU Luxembourg, the University of Liège, Datasoft Management nv(?) Oostende and the Military Hospital Brussels.

<sup>2</sup> CAT2 is a unification-based Machine Translation system developed at IAI Saarbrücken, the most up to date description of which can be found in [Sharp and Streiter1995].

classification of concepts for the whole domain (cf.[Arnold and Sadler1992]) - an existing classification in Medicine, the Systematized Nomenclature of Medicine (SNOMED) (cf. [Côté et al.1993]), was adopted for this purpose. For every language involved the coupling of words to concepts is done in the CAT2 lexicons, in which every lexical entry contains the associated SNOMED code (e.g. *snomed*= 'M-12000') apart from the lemma (e.g. *lex*=*fracture*) and its grammatical description.

In order to express generalizations about the concepts, these concepts are re-grouped into 27 classes (Semantic Types). Some classes are taken from the dimensions used in SNOMED (e.g. *topography*, *morphology*, *function*), others such as *living\_object* and *chemical* are identified by the paradigmatic relations which they maintain (for a complete description see [Ceusters et al.1994a]).

In order to describe the combination of Semantic Types, a set of Semantic Roles has been developed. Following linguistic models, each possible head of a structure was assigned an argument frame, i.e. a set of arguments to which the head can assign a Semantic Role, plus the restrictions on the Semantic Type, in order to control the access to the argument slot. The set of Semantic Types, Roles and Restrictions is called the 'ANTHEM Semantic Model' which represents the interlingual domain specific knowledge on which the analysis in CAT2 is based.

### 3 Implementation in CAT2

#### 3.1 Basic Schemes of Composition

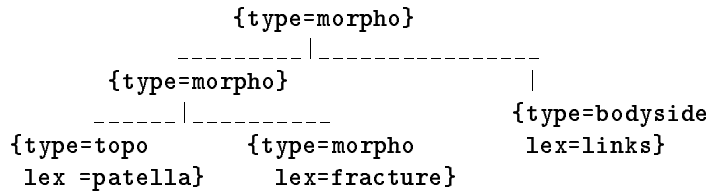
The syntactic and semantic analysis is carried out with a limited set of schemes of composition (building-rules) which represent the structures by which larger expressions are formed. The generic parser implemented in CAT2 uses these patterns of composition to build valid tree structures. The schemes are responsible for the construction of (1) head-argument structures, (2) functional projections (3) coordinated structures and (4) multi word units.

The main scheme of composition, the head-argument structure, takes into account only semantic (i.e. domain-specific) information and percolates the semantic properties of the semantic head (i.e. the part that functions as predicate) to the mother node. Every tree structure build by this scheme is submitted to a syntactic verification rule which may filter out illegitimate structures. This rule tests the position and inflection of adjectives and adjacency constraints (e.g. the position of of-phrases and genitive-phrases). Syntactic information is furthermore used in preference rules for disambiguation of the semantic analysis. Examples are the preference of *topologies* where the third localizes the second and not the first, or the preference of a nominal over an adjectival semantic head for some type distributions, in which, given the ANTHEM Semantic Model, both could function as semantic head.

Moreover the verification rule identifies the syntactic head (i.e. the daughter which shares its syntactic properties with the mother node). Through this 'split'



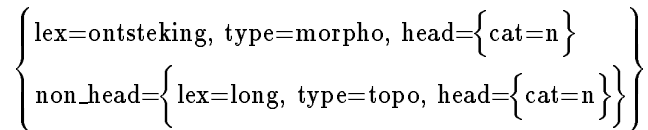
PATELLA.



The implementation of discontinuous structures is based on the slash principle as described in [Gazdar et al.1985]: If the *bodyside* slot of a *morphology* is not filled when its projections becomes the semantic non-head of a second structure, the empty slot is passed within the *slash* feature onto the semantic head.

### 3.3 Compounds

Compounding is a morphological process, which is equivalent to syntactic means to combine concepts. Dutch and German compounds are analysed into their parts: the head of the compound (the rightmost component) contains in its feature bundle a complete description of the non-head, as shown for the Dutch compound *longontsteking*.



A verification rule (f-rule) determines the semantic nature of the compound through the unification of the non-head with one of the argument slots of the head. For the purpose of translation head and non-head will be broken down into two separated tree structures, so that the translation is a compositional one, where each part finds its equivalent in the target language (SPIER => MUSCLE, SPASME <=> SPASME).

```

PARAVERTEBRALE SPIERSPASMEN      ('para-vertebral muscle spasms')
=>
SPASMES PARAVERTEBRAUX DANS LE MUSCLE  ('spasms para-vertebral in the muscle')

```

In medical sublanguage, however, compounds like MWUs often refer as a whole to one concept only. This is accounted for by a separate lexical entry for this concept, blocking the analytical translation process.

```

TENNISELLEBOGEN      ('tennis elbow')
=>
* COUDE DE TENNIS    ('elbow of tennis')
EPICOINDYLITE        ('epicondylitis')

```

### 3.4 Compounds and Extraposition

In the same way as phrasal structures, compounds allow for the extraposition of the *bodyside* and thereby the appearance of discontinuous structures, where the *bodyside* refers to the non-head of the compound. This structure is accounted for by the same *slash* mechanism: If a *morphology* is appearing as a non-head of a compound, the *bodyside* slot is transmitted within the *slash* feature to the head, from where the *bodyside* marker can be bound.

ENKELDISTORSIE LI            ('ankle-contortion left')  
POLSONTSTEKING RECHTS      ('wrist-inflammation right')

### 3.5 Lexical Functions

The concept of **Lexical Functions** has been developed by Mel'čuk in the framework of his Meaning $\Leftrightarrow$ Text Model (cf. [Mel'čuk1974]). The essence of this notion is that one word A selects a second word B in order to realize a special meaning related to A, which is called the lexical function LF. Assume A to be SMOKER. In order to realize a 'high degree' of it, which is not morphologically possible in English, A selects B = HEAVY as its modifier in order to realize through the expression HEAVY SMOKER the high degree of SMOKER. Lexical Functions merit a special treatment in MT since A but not B can be translated literally. In ANTHEM we find lexical functions within the combination of *severity* and *function*. In many cases the *function* selects one special *severity* operator for the Magnifier, and another, not necessarily related word for the Minifier; the examples are taken from the German corpus.

- SCHWERE/LEICHTE VERBRENNUNG (heavy/light burn)
- HOHES/LEICHTES FIEBER (high/light fever)
- STARKE/SCHWACHE SCHMERZEN (strong/weak pain)
- STARKE/LEICHTE RÖTUNG (strong/light reddening)

The different realizations of the high degree (i.e. the words SCHWER/HOCH/STARK) and the low degree (i.e. LEICHT/SCHWACH) receive the same interlingual representation. Which lexeme is to be chosen for its realization is determined in the lexicon for every *function*.

## 4 Conclusion

In the preceding discussion we have shown how the CAT2 system analyses the natural language input of ANTHEM based on an interlingual domain-specific semantics and syntactic and lexical restrictions. The main difficulties with the analysis and translation of this input arise from the mismatch of syntactic and semantic principles (e.g. principles of projection and principles of lexical selection). As a consequence, the semantic and syntactic constraints which represent the domain specific and the language specific knowledge respectively, apply in sequence in order to assure an efficient and correct analysis and translation.

## References

- [Arnold and Sadler1992] Doug Arnold and Louisa Sadler. 1992. Unification and machine translation. *Meta*, 37(4):657–680, December.
- [Ceusters et al.1994a] Werner Ceusters, Guy Deville, Emmanuel Herbigniaux, Pierre Mousel, Oliver Streiter, and Geert Thienpont. 1994a. The ANTHEM Prototype. Working paper 31, IAI, Martin-Luther-Straße 14, 66111 Saarbrücken, BRD.
- [Ceusters et al.1994b] Werner Ceusters, Guy Deville, Oliver Streiter, Emmanuel Herbigniaux, and Jos Devlies. 1994b. A computational linguistic approach to semantic modeling in medicine. In *Belgo-Dutch Congress on Medical Informatics '94*, pages 311–319, Veldhoven.
- [Côté et al.1993] Roger A. Côté, David J. Rothwell, Ronald S. Beckett, and James L. Palotay, editors. 1993. *Developing a standard data structure for the systematized Nomenclature of Human and Veterinary Medicine. SNOMED International. Introduction*. College of American pathologists & American Veterinary Medical Association.
- [Gazdar et al.1985] Gerald Gazdar, Ewan Klein, Geoffrey Pullum, and Ivan Sag. 1985. *Generalized Phrase Structure Grammar*. Basil Blackwell, Oxford, UK.
- [Mel'čuk1974] Igor Aleksandrovič Mel'čuk. 1974. *Opyt teorii lingvisticeskix modelej Smysl ⇔ Tekst. Semantika, sintaksis*. Izdatel'stvo "Nauka", Moskva.
- [Sharp and Streiter1995] Randall Sharp and Oliver Streiter. 1995. Applications in multilingual machine translation. In *Proceedings of The Third International Conference and Exhibition on Practical Applications of Prolog, Paris, 4th-7th April*.