Ceusters W. Medical natural language understanding as a supporting technology for data mining in healthcare. In: KJ Cios (ed.) Medical Data Mining and Knowledge Discovery, Physica-verlag Heidelberg, New York, 2001;:41-67.

Medical Natural Language Understanding as a Supporting Technology for Data Mining in Healthcare.

Werner Ceusters

In this chapter, we describe the role of language engineering techniques in text mining, a discipline focusing on information extraction from free texts. Indeed, text mining expands the idea of data mining in structured databases towards information discovering in natural language documents. After an introduction of the various tools and techniques that are available for text mining from the linguistic engineering point of view, we concentrate on a specific application in the domain of medicine.

1 Understanding the problem domain

1.1 Introduction

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, a big portion of them resulting from studies conducted on data and information accumulated in patient records. Despite the growing tendency to make this information available in electronic format, turning the information into knowledge is not an easy task. Indeed, faithful recording of patient data can only be achieved by using natural language. This was already stated in the early eighties by Wiederhold who claimed that *the description of biological variability requires the flexibility of natural language and it is generally desirable not to interfere with the traditional manner of medical recording* (Wiederhold 1980). At the other hand, it is evenly true that without proper mechanisms in place, free natural language registrations are impossible to be understood by machines, if not to say, quite often also by colleagues. Very often, medical statements are written down in a context that is obvious at the time of registration, but that is difficult to reconstruct later on by third parties, or even by the original source. Also, in order to allow a computer to process healthcare data further such as for data mining purposes, the data must be available in a coded and structured format. Making that happen in a transparent way for healthcare specialists, is the ultimate goal, if not even the “raison d’être” of natural language understanding applications in healthcare. 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 (Bateman et al. 1995).

Natural language processing systems are already looked at since the sixties, though mostly only in academic environments. Now it is recognised by major technology consultants in the healthcare domain as an emerging technology with great prospects for the near future. Real applications start to become available, and once the current problems related to continuous speech recognition will be solved, a massive penetration of natural language understanding applications will undoubtedly occur.

In this paper, we give an overview of the many faces of language engineering applications, focusing on their impact on data mining in healthcare. It is not the idea to give a detailed course in theoretical computational linguistics. In line with the new ideas on successful management in business, we prefer to pay attention to solutions, while not loosing time by focusing too much on the problems (Drucker et al. 1997).

1.2 The many faces of language engineering

1.2.1 Definitions

The development of a machine that understands a human being has been a great dream ever since the beginning of computers. Proof of that are the numerous science-fiction stories in which a computer is addressed in ordinary human speech, i.e. natural language, upon which the machine promptly answers with a metallic sounding voice. The obstacles between that dream and today’s reality are still enormous, but the light is beginning to shine in certain specialised domains.

Natural language processing applications come in many flavours. At the heart of the technology is a specific discipline of science called *computational linguistics*, aiming to develop computational models of language that explain how language works in human beings, and how this insight can be used to allow computers to work with language. If the focus is more on the development of practical applications rather than on theoretical studies, the term *linguistic engineering* is preferred.

As with many disciplines, sub-branches of linguistic engineering emerged very quickly. A first major division is to be recognised between *language processing* and *speech processing*. The basic aim of *speech processing* or *speech recognition* is to turn the sound wave generated by a speaking human being into a digitally represented text, f.i. by using the ASCI set of characters. Speech recognition is not be confused with *voice recognition* which aims to identify or authenticate individual people based on some physical characteristics of their voice.

The result of a speech recognition application can be used in word processors or printed on paper. The computer processing the speech signal has however no understanding of the meaning of what has been said, nor is the resulting text by any means a representation that is immediately understood by the machine. *Language processing* at the other hand starts with the verbal representation of - say - an ASCI text, and uses this format to do some further useful processing.

A second major subdivision that cuts orthogonally through the previous one, is whether or not understanding of speech or language is at stake. It is possible to do many tricks with language - and even to build very useful applications by doing so - without a need for true understanding of spoken or written texts. Many information retrieval packages operate in this way by doing string searches, some basic stemming procedures - such as transforming conjugated verbs or plurals to their base form - and counting words, with fairly adequate success. Also the *command & control* paradigm where a computer user can dictate commands to a computer instead of using a mouse, belongs to this class. For this kind of applications, the general terms *natural language processing*, versus *speech processing* apply, whereas if true understanding is achieved, the term *natural language understanding* is preferred.

A third division has to do with the direction of processing. While generally with natural language understanding, *natural language analysis* is understood (going from a text to its meaning), the opposite (going from a meaning representation to a text) is called *natural language generation*. For speech applications, the terms *speech-to-text* or *text-to-speech* are often used. Be aware that also here the understanding issue cuts orthogonally through the applications. It is perfectly possible to have text-to-speech applications that do not understand what is being said. Also specific paradigms of *machine translation* work quite well without deep understanding of the texts that must be translated.

Natural Language Understanding is being considered as one of the most complex problems in artificial intelligence. Up to now, a computer is not yet capable of really understanding the true meaning of ordinary human language. The necessary background knowledge is so extensive and complex that even today's description-possibilities are unable to describe everything. Food for thought is the idea that a human child needs at least six years to adopt a language and that even today's supercomputers don't posses even a fraction of the comprehension capabilities of the human brain.

However ! Under certain specific circumstances it is possible to have a computer understand natural language. Medical language, as a sub-language of ordinary human language, is a field that complies in an excellent way with the 'specific circumstances' required: a closed world with restricted domains and disciplines easily separated from each other, a relatively uniform terminology, and the availability of numerous descriptions (textbooks, classifications, …). Because the principles of understanding natural language in the world of medicine could have immediate and huge advantages, the conception of systems to make a machine understand medical language has been a field of research for nearly 20 years. The results of this research are now becoming available as *medical natural language understanding*.

1.2.2 Natural language understanding applications for healthcare telematics

There are numerous applications for which medical language technology may pay off. Quite a bit of those have an immediate added value in the present and future clinical-care organisations, most often as enabling tools in the field of traditional telematics. Medical Language Technology is the new engine that will provide the power to stimulate the next generation of medical software applications. Table 1 summarises the possibilities. Some are discussed more deeply in the following paragraphs.

1.2.2.1 Automatic encoding

To overcome the problems related to the use of natural language in communication and clinical registration, coding and classification systems have been introduced as interlingua. Systems such as ICD, Snomed International, ICPC, CPT and many others are now widely used to register medical findings, diagnoses or procedures. Similarly, terminological systems such as NIC, NANDA, ICNP and others are proposed to be used as interlingua in a nursing environment.

Table 1: Use of NLU in the healthcare domain

- semi- and full-automatic ICD-registration and coding based on full-text-reports.
- medical terminology-management on all levels (departmental, hospital, HMO, National)
- natural-language data-entry-facilities for EPR-systems
- tools for building, selection and evaluation of clinical guidelines on all levels (departmental, hospital, HMO, national)
- automatic translation of medical files into a multiple range of languages (for telematic or telemedicine-purposes)
- automatic conversion of medical files into different classifications and mapping between classifications (ICD9, ICD10, Snomed, CPT4, UMLS, Mesh, Read, ICPC, ICNP, CISP, …)
- tools for medical-data-cleaning and uniformisation for datawarehouses
- tools for full-text-retrieval and semantic searches
- access tools for internal and external knowledge-bases

Coding patient data means that a physician (or professional encoder) has to describe the patient data by means of codes that are a kind of placeholders for the concepts available in systems such as ICD. The requirements to be met in order to perform the coding task adequately are (Ceusters et al. 1996) : 1) a perfect understanding of the meaning of the patient data (the source concepts), 2) a perfect understanding of the meaning of the concepts available in the concept system (the target concepts), 3) at least a certain level of similarity and coherence between the source concepts and target concepts, 4) facilities to search the concept system for the target concept(s) that match(es) a given source concept as closely as possible.

It is common knowledge that coding performed by humans is of rather low quality, both in terms of recall/precision, inter-rater variability, and even reproducibility by the same team. Natural language understanding tools can improve coding quality dramatically.

1.2.2.2 Medical terminology management

Coded data are the most convenient way for computers to turn data into information. This is the main reason for the success of coding and classification systems. Hélas, the one omni-potent classification system that fulfils the needs of all doctors, nurses, hospital managers, governments, librarians and international organisations, has yet to be developed. We are even quite convinced it never will be built ! There always will be a need for local variations, for additional dimensions, for greater detail, etc. And as long as a variety of systems continues to be available, the need for integration, mappings and translations will also continue to exist.

That is why people working in the domain of medical natural language processing invest in the development of tools that allow them to work with various classifications, without however becoming too much dependent on them. Assisted by such language analysis tools, mappings can be created from local systems to any other, while guaranteeing that they will remain compatible with future and previous versions. By doing so, users can be sure that their precious data don't become worthless once a new version of an official classification system becomes available.

1.2.2.3 Natural language data entry

Continuous speech recognition software will soon become available at a level of quality that is acceptable to be used in routine medical practice. Discontinuous speech applications are on the market since many years but cannot be said to be a big success. Speaking discontinuously, i.e. pausing after each word or word cluster, is not really practical. Also "command and control" speech applications where - if we may say so - not just the keyboard is replaced by a microphone, but also the mouse movements are to be guided by the voice ("go to medication", "enter 3 tablets of Aspirin", "go back", …) are only useful in some uncommon situations where it is impossible to use the hands to operate a mouse, light pen, keyboard, or whatever other "conventional" input device.

The availability of continuous speech recognition software will have as consequence that the structured data entry of today will disappear gradually, probably even completely in a not so far future. This requires for powerful full text understanding systems that can capture the true semantics of what is said by the user. For specific domains (radiology, pneumology, …), such "text-to-meaning" applications are already available, and this in various languages. Interest in such systems is constantly growing thanks to XML, a format that is perfectly suited to capture the recursively embedded meaning-representations resulting from free text analysis.

1.2.2.4 Clinical trials and practice guidelines

Language understanding services are needed when free text entries (whether being full text or short phrases) entered in a certain context, are to be used for other purposes. A typical example is matching patient selection criteria for clinical trials. It is not easy for a physician seeing patients on a routinely basis to bear in mind constantly what clinical trials are running in his department, and what criteria must be met by a patient to enter a trial. It is not feasible to run over the inclusion criteria for each single patient during an encounter. It is more sensible to have a software "watchdog" that constantly monitors the data entered by a physician, and that produces an alert when specific criteria are met. If data are entered in free text, this means that such a watchdog must have enough language understanding power to identify "numbness in left lower leg since last week" as satisfying an inclusion criterion such as "sensory disorders of the limbs lasting for more than 24 hours".

The same goes for checking whether or not practice guidelines are followed when registering patient data, or to generate other alerts upon specific criteria.

1.2.2.5 Intelligent querying, information retrieval

Many electronic patient record systems keep collections of text documents (discharge summaries, referral letters, surgery reports, …) related to individual patients. Documents in these "result servers" are only accessible through general indicators such as the original source, the kind of document or the creation data. Searching documents on the basis of their content is seldom possible, or only by means of string search or some crude pattern matching mechanisms with jokers. Natural language understanding techniques can add a lot of functionalities to these primitive mechanisms.

Searches could be improved by using a thesaurus. A basic problem is however where to get one that is suitable for your needs. For medical bibliographic retrieval, one could use the well known MESH thesaurus from the National Library of Medicine. But this thesaurus is largely insufficient to be used in clinical practice. With the proper natural language processing tools, special purpose thesauri that respond to local demands can be built.

Having a thesaurus is not enough. The next step is to attach thesaurus entries to the documents. This is the problem of indexing. Traditionally, indexing is done manually. Professional indexers read a document, and assign the relevant thesaurus entries to it. Natural language understanding software is able to automate this process partly or even completely. The result is an electronic index that gives you fast access to documents on the basis of their “conceptual content” and not limited to the occurrence of specific words.

However, this is not the end of the story. Using a thesaurus to index properly a collection of documents, guarantees that you will find all (and not more) the documents that you need, provided that you know perfectly the terminology used in the thesaurus. To overcome this restriction, “query enhancement” techniques can be used to match a user’s query to one or more relevant thesaurus entries. This requires the use of a semantic network.

1.2.2.6 Text mining

Together with the recent interest in data mining, also *text mining* has been introduced as a new discipline. Text mining applications support knowledge workers who must extract meaning and relevance from large amounts of information available in textual format. Both text and data mining have much in common with archeology, because underlying each is the assumption that knowledge lies buried in a scattered mass of information.

Typical text mining applications cluster documents in sets that have a common semantic basis. The semantic basis can be queried beforehand by the user, or discovered automatically by the software using statistical techniques. Other text mining applications try to summarize documents, or pinpoint the user to parts of documents that contain information that the user probably did not see before.

2 Understanding the data

2.1 Types of knowledge

The different forms of knowledge that traditionally are claimed to be required for proper written text understanding are: *morphology*, *syntax*, *semantics*, *pragmatics* and *discourse* or *world knowledge* (Allen 1987). It is obvious that these forms of knowledge do not stand on their own, but that they are tightly related. At morphological level, inflection may be seen as a pure syntactic phenomenon, whereas compounding is merely guided by semantic principles. The actual form of a sentence depends amongst others on the situation under which a meaning is to be conveyed. As such pragmatics and discourse have an influence on syntax. Some authors even deny or reduce the distinction between some of these kinds of knowledge. Quine for instance showed that semantic knowledge and world knowledge cannot sharply be delineated (Quine, 1953).

When dealing with terminology rich domains and with automated knowledge acquisition from written text understanding as a primary goal, it is possible to simplify the picture and to adopt a rather reductionist view. First, we can abstract away from the discourse level. Authors of medical textbooks, developers of terminologies or physicians writing patient reports, merely want to convey facts, and not to invoke emotions or to initiate actions by the reader. As such, we can limit our analysis to what in the speech-act literature is known as *constative inscriptions*, sentences uttered in a descriptive context (Searle et al. 1980), however without being too narrow as is the case in the traditional formal linguistic semantics scene where sentence-meaning is viewed as being exhausted by propositional content and is truth-conditionally explicable (Bach 1989).

We also can abstract away from pragmatics - although not ignore its existence - as it is not our aim to provide theories on how context changes the surface forms of the expressions we are looking at. When looking to terminological phrases, we can certainly abstract away from indexical information. Terminological phrases by definition have to be self-explaining and do not refer to entities that are outside the domain covered.

In a monolingual environment, we could also ignore morphology, but as multilinguality is one of the main objectives in large scale text understanding, this would be too big a sacrifice. However, for the sake of simplicity and quietly assuming that the principles that govern word-formation are similar to the principles that govern syntax, we will not further deal with morphology in this chapter.

2.2 Linguistic and conceptual knowledge

In our reductionist view, we can see a medical text as the product of a process in which words or word groups that refer to concepts, are put together following linguistic rules to form larger word groups that refer to new concepts that have a certain relationship with the original concepts.

Since the early activities of CEN/TC251, the Technical Committee of the European Standardisation Centre that deal with healthcare informatics, references to *conceptual* models, *concept* systems and *conceptual* semantics are dominating the medical informatics literature (Rossi-Mori 1994). For the purpose of this book, we mean by *conceptual knowledge* that knowledge that exclusively deals with concepts and the organisation of these concepts in a structure that is independent of any language. This is not a fortiori the same as what in the linguistic literature is known as *conceptual semantics*, which is a particular theory on *meaning as conceptual structure* (Jackendoff, 1988; Lakoff, 1988). Central in this theory is that semantic structures (what we denote) and conceptual structures (what we mean) converge, or even are the same. However, this probably is the case in a terminology rich domain such as medicine. Hence the *semantics* (i.e. the linguistic meaning) of a medical expression can be said to be equal to the concept that is referred to.

In the light of our data mining objective starting from written text understanding, we mean by *linguistic knowledge* that knowledge that specifies the rules of how valid expressions in a particular language are formed. This kind of knowledge comes in different flavours, two of which in our reductionist view are of importance. First there is the pure grammatical or syntactic knowledge that f.i. dictates phrase constituent order. Typical examples are the adjective - noun order in English, and the noun - adjective order in French. Gender agreement between nouns and adjectives in French is another example.

A second kind of linguistic knowledge is the one that is influenced by meaning. It is this kind of knowledge that tells us that actions usually are denoted by verbs, and entities by nouns. It is also this kind of linguistic knowledge that dictates us that adjectives denoting colour must appear just in front of nouns, and after other adjectives. This knowledge is extremely interesting for our purposes, as it holds the key of the door that leads from denotation to meaning. The particular branch of semantics that deals with this issue is *linguistic semantics*: *the study of literal meanings that are grammaticalised in a language* (Frawley 1992).

2.3 Linguistic semantics

A first principle of linguistic semantics is that one looks only at the *literal*, i.e. *decontextualised* meaning of an expression. From the standpoint of literal meaning, the expression

(E-1) *removal of cardiac pacemaker from epicardium or myocardium.*

represents a state of affairs that involves an event of *removing* and certain entities namely a *cardiac pacemaker*, an *epicardium* and a *myocardium*. There is no discussion about that. If we know that this expression is the rubric-term for SNOMED-code *P1-315C4*, then we know also that the implicational, i.e. contextualised meaning of this expression is that if on a patient a cardiac pacemaker is removed from one of the two specified places, this may be registered in his medical record as *P1-315C4* if there is an agreement in the institution where the procedure is carried out that such interventions are coded in SNOMED. The notions *patient*, *institution*, *agreement*, etc, are required to understand the full semantics of the expression, but it is obvious that these notions are not encoded in the sentence itself. Hence they are not part of the linguistic semantics, or the *grammatical meaning* of the sentence.

At the other hand, the entities pacemaker, epicardium and myocardium appear in sentence (E-1) as structural categories, in casu nouns, that are essential to the formation of English sentences. From expression (E-1), we know also that it is the pacemaker that is the entity on which the event of removing acts, and not the epi- or myocardium. We are sure about that just because of the form of the expression in English, and not because we have to infer it from other information, e.g. because this expression is a rubric in SNOMED. It is the preposition *of* that marks the object that is removed, and the preposition *from* that encodes the source from which the removal is carried out.

2.4 Conceptual and linguistic ontologies

All knowledge based approaches rely on an *ontology*, a more or less formal representation - to be used in computer systems - of what concepts exist in the world, and how they relate to one another. Ontologies are often viewed as strictly language independent models of the world, especially in the medical informatics community (Rector et al. 1996), though the need for an ontology in natural language processing applications is generally well accepted (Bateman 1993). This is not to say that knowledge structuring based on a linguistic approach leads to the same result as when opting for a conceptual approach. A typical example is the ontological distinction between *nominal* and *natural kinds* (Kripke 1972), that in no language is grammaticalised just because the difference is pure definitional (Welsh 1988).

Situated ontologies - i.e. ontologies that are developed for solving particular problems in knowledge based applications (Mahesh and Nirenburg, 1995) - that have to operate in natural language processing applications, are better suited to assist language understanding when the concepts and relationships they are built upon, are linguistically motivated (Deville and Ceusters, 1994).

In the perspective of re-usability, two dimensions have however to be explored: (relative) independence from particular languages and (relative) independence from particular domains.

Linguistic semantics based analyses allow us to separate f.i. entities from events and property concepts, a rather crude distinction being the fact that in most languages these concepts are respectively grammaticalised by means of nouns, verbs and adjectives. Linguists are concerned on how these concepts give overt form to language, while from a computational point of view, these concepts also have to be "anchored" in an ontology.

A relative new notion related to ontologies is that of the *interface ontology*, standing between conceptual (or domain) and linguistic ontologies. Approaches based on interface ontologies differ in the "distance" between the interface ontology and the domain ontologies at the one hand, and the linguistic ontologies at the other hand. In the MikroKosmos initiative, an interface ontology is developed for machine translation purposes in the domain of commercial merges and acquisitions of companies (Mahesh 1996). Hence, it is more close to a given conceptual domain, although general concepts are included as well as unrestricted texts are envisaged to be processed. The KOMET project resulted in the "Generalised Upper Model 2.0", where a closer contact with linguistic realisations is maintained: *if there is no specifiable lexicogrammatical consequences for a 'concept', than it does not belong in the Generalised Upper Model* (Bateman et al. 1995 : p5). 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 (Vossen et al. 1997). In this project, semantic databases like WordNet1.5 (Miller et al. 1990) 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.

3 Preparation of the data for subsequent data mining

The application of natural language understanding that will interest most the readers of this textbook, is to prepare textual documents in such a way that the information they contain can be used for traditional data mining. This requires applications that can re-arrange the unstructured information that resides in texts, into a structured format.

In the remaining part of this chapter, we will first describe the prinicples and technologies that underly natural language understanding applications in the context of information discovery in healthcare. Then we'll describe a system that has achieved these objectives.

Various technologies are indispensable in the process of representing medical natural language in a format understandable by machines. Some of these technologies are used "off line", i.e. they assist in the development of resources that are needed to drive the "on-line" technologies such as syntactic-semantic taggers and parsers. We refer to these resources and technologies together as "lingware" because contrary to traditional informatics tools, they are specifically designed for linguistic processing, and software and knowledge bases are tightly interconnected.

3.1 Technologies required

3.1.1 Machine readable multilingual medical lexicons

Dictionaries are usually large books intended to be used by humans to look up the meaning of unknown words. Most electronic dictionaries currently available differ only from paper dictionaries in their being published on a digital medium. The major advantage is that they can be used from within the most popular word processors without the need for retyping. But their audience consist still of human readers…

For medical natural language understanding purposes, dictionaries have to be fundamentally different in nature: they are primarily intended to be used by machines ! Such dictionaries can be used by some of the knowledge extraction software to represent the meaning of full text documents. But they also can be integrated in third party systems for information retrieval, spell checking, automatic translation, etc.

3.1.2 Automatic term extractors

Linguistic engineering is a very labour intensive activity. Hence there is a need for tools that can be used to automatically extract new words and terms from text documents. Whether they are in English, Dutch, French or whatever other European language, the vast majority of typical expressions contained in documents pertaining to a specific domain should be extracted on the fly. In addition, semantic relationships between the content words of the documents (i.e. those words pertaining to the domain) are to be made explicit.

3.1.3 Taggers and parsers

These tools are linguistic analysers that are able to make the implicit knowledge available in texts more or less explicit.

Taggers are pieces of software that take as input a text, and that “decorate” these texts with syntactic and/or semantic descriptions pertaining to the sentence constituents identified.

Parsers provide a complete structural analysis of sentences. Here also, analysis can be limited to syntactic structure, or complemented by semantic decorations.

3.2 Preparing neurosurgical reports for data mining purposes: the MultiTale approach

3.2.1 Rationale

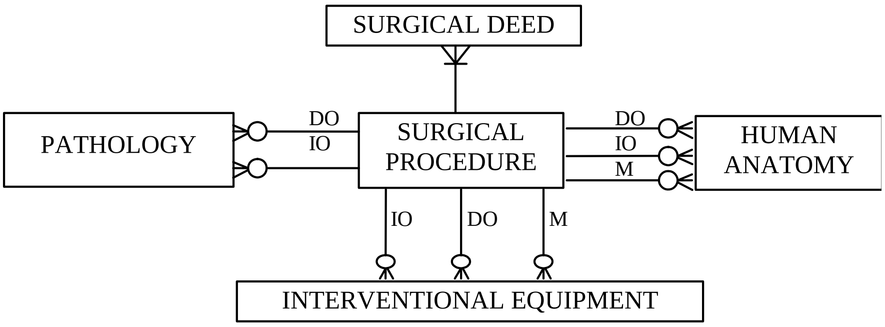
For data-collection in healthcare, two major issues are generally well recognised. Firstly, coherent models for the representation of data, information and knowledge are urgently needed. Indeed, only when such models are available, data coming from different sources may be compared and used for various purposes. Secondly, despite the increasing use of healthcare information systems in which data are registered in a structured and standardised way, the clinical narrative remains an important source of information when delivering optimal care to patients. However, the format in which this information is expressed, cannot readily be used for automatic data-processing, and is inefficient in terms of computer assisted medical management, data transfer between different information systems, quality assurance and surveillance.

The Multi-TALE syntactic-semantic tagger has been developed as a tool for factual information retrieval from neuro-surgical procedure reports, hence making this information suitable for further processing (Ceusters et al. 1998). In the remaining part of this chapter, we describe the underlying semantic model and the overall technical architecture of the prototype, sketch the results of two validations, and compare the results with those of similar systems.

3.2.2 The CEN/TC251 ENV on 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* (CEN ENV 1828:1995).

According to this standard, a surgical procedure is conceptually composed of a *surgical deed* (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).



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 (Dik 1989), and successfully applied and adapted by Deville for the modelisation of administrative language (Deville 1989). 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 or case, and functioning as arguments of that predicate. We identified the concept field *surgical deed* as predicate primitive and the semantic links *direct object, indirect object and means* as the expression of 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 by appealing to the notion of predicate primitive. To each class of surgical deed corresponds a cluster of arguments with specific semantic constraints.

4 Data mining

The function of an automatic *tagger* is mainly to perform the first and essential pre-processing for any natural language processing application: labelling words in sentences with their grammatical (sub-)categories, only taking into account a limited context. The output of an automatic tagger is a list of the words of a text with the appropriate labels. This output can be used for further processing, for example by a parser. With *parsing* is meant here performing a complete syntactic analysis of a sentence, distinguishing syntagms (noun phrase, verb phrase, etc.) with their syntactic functions (subject, direct object, main verb, etc.). In addition to this syntactic tagging, the Multi-TALE prototype also provides semantic tags to the texts. More specifically, semantic decoration is provided on the basis of the CEN\TC251model for Surgical Procedures as explained above.

The existing DILEMMA-1 tagger-lemmatizer for English general language was used as a starting-point for the analysis (Paulussen and Martin, 1992). Its output, a list of lemmatised words with their basic syntactic categories (adj, noun, verb, ...) was then processed by the semantic tagger. This tagger uses a rule-set to combine in a bottom-up approach 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. Each rule contains information regarding the syntactic and semantic conditions to be fulfilled by the word or word group in focus, and its left and right neighbours within in a specified distance.

Other resources used by the Multi-TALE prototype are a small syntactic lexicon to correct the systematic mistakes of DILEMMA-1 (Mommaerts et al. 1994), and a semantic lexicon of 2000 words (base-forms) with additional information on case-assignment and usage of prepositions.

Finally, a heuristic scoring procedure is used to select the best solution when more than one analysis for a given sentence is possible.

For example: the sentence “After insertion of the ventricular catheter, removal of cerebral fluid was performed.”, is analysed by Multi-TALE as:

Semantic link	**Semantic type**	**Syntactic tag**	**Rule**	**Syntagm**	**Meaning (Snomed-code)**
		prep		After	G-4004
action	install	sg		insertion	P1-05500
		prep		of	
do	inte	detnoun	4	the ventricular catheter	(T-A1600,A-26800)
		comm		,	
action	remove	sg	bac	removal was performed	P1-03000
		prep		of	
do	anat	adjnoun	2	cerebral fluid	(T-A0110,fluid)

Typical output of Multi-TALE

4.1.1 The Multi-TALE reference lexicon

4.1.1.1 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 that were used in the development of the system.

a) 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 craniectomy was then extended as a 2 cm strip a long fused metopic suture down to the skull 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. *Haemostasis 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.

b) Single words

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

c) 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.*

4.1.1.2 *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> ]
```

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).

4.1.2 The syntactic-semantic grammar

A separate linguistic knowledge base attached to the system 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 sentence is considered to have the following structure:

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

E.g.: The sentence "*A big fatty tumour was removed rapidly from the brain and the hole filled with pieces of artificial tissue*", is composed of the segments "*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 segmentor "and". The first segment contains the clauses "*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 the next table.

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

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


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

Bottom-up syntactic-semantic tagging in Multi-TALE

Each grammar rule has the following general format:

IF 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>,

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>,

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

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
- **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.

5 Evaluation

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. 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.

The next table 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

Validation results of the Multi-TALE syntactic-semantic tagger.

5.1 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 (Sager et al. 1987). 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%) (Sager et al. 1994). 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 (Ranum 1988), and rates of 95% and 94% respectively for the detection of diagnoses (Haug et al. 1990). SPRUS is mainly semantically driven, and is not able to exploit syntactic information. Hence complex noun-phrases and whole sentences cannot be processed. 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% (Lin et al. 1991). 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.

The main conclusion from this work was 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.

Another conclusion was that the CEN/TC251/PT002S standard for surgical procedure classifications can be used for syntactic-semantic tagging of neurosurgical procedures provided that the concept type *surgical deed* is thought of as predicate primitive and the semantic links *direct object, indirect object and means* as the expression of semantic functions labelled case roles that can be ascribed to the arguments of a predicate (i.e. concept of surgical deed). In this context, Multi-TALE is in line with the World Health Organisation's view in that *the challenge is now to clarify* *how we can pass from traditional encoding of medical data to automatic encoding of natural language, and how the universally accepted classifications with their advantages and disadvantages can be used in this context* (Jardel 1989).

6 Conclusion

Medical information systems are sufficiently large and varied such that no one vendor can expect to provide all of the systems needed in even a single hospital, let alone for the health service as a whole. Many of these varied systems would benefit from natural language interfaces and some, such as automatic linkage to abstracts of the literature, are even impractical without it. Generic multilingual solutions are required if the range of services to be built is to meet the demand. Furthermore, it is essential that the natural language processing components share the underlying concept structure used by the various applications.

Electronic patient record systems are no exception to this. A wealth of knowledge is needed to enter information in those systems consistently and to use the information afterwards for various purposes. Provided that a highly acceptable system can be designed in a specific environment, then developers surely will want to make it available to other users. Whilst much re-use of system components is feasible within a given market segment, there are significant costs associated with the 'localisation' of systems to the needs of other markets. Perhaps the most important of these costs is the localisation to the linguistic needs of each national market in Europe.

Medicine is a descriptive, language intensive activity, and the costs of developing, and perhaps more importantly maintaining, the linguistic resources needed to localise clinical systems are clearly high. Any practical approach to the management and exploitation of linguistic resources in large scale clinical information systems must be based on common methods and internal representations for linguistic information. This information must be reusable across a wide range of systems and local variants of those systems, and the cost of maintaining that information must be separable from those of maintaining the rest of the system.

Data mining is also one of these new disciplines that can benefit from natural language understanding applications. Numerical data are hard to get in domains such as medicine where most data are only available in textual format. Preprocessing textual information through natural language analysers will be one of the solutions to resolve the knowledge acquisition bottle neck.

7 References

Allen J. 1987. Natural Language Understanding. Menlo Park: The Benjamin/Cummings Publishing Company Inc.

Bach E. 1989. *Informal lectures on formal semantics.* Albany, NY: Suny Press.

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

Bateman JA. 1993. Ontology construction and natural language. *In Proc. International Workshop on Formal Ontology.* Padua, Italy, 83-93.

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

Ceusters W, Lovis C, Rector A, Baud R. 1996. Natural language processing tools for the computerised patient record: present and future. In P. Waegemann (ed.) *Toward an Electronic Health Record Europe '96 Proceedings*, 294-300.

Ceusters W, Spyns P, De Moor G, Martin W (eds.) 1998. Syntactic-Semantic Tagging of Medical Texts: the Multi-TALE Project. Studies in Health Technologies and Informatics, IOS Press Amsterdam.

Deville G. 1989. Modelisation of Task-Oriented Utterances in Man-Machine Dialogue System. Ph.D. Thesis, University of Antwerp.

Deville G, Ceusters W. 1994. A multi-dimensional view on natural language modelling in medicine: identifying key-features for successful applications. Supplementary paper in *Proceedings of the Third International Working Conference of IMIA WG6*, Geneva.

Dik S. 1989. A Theory of Functional Grammar, Foris, Dordrecht.

Drucker P, Dryson E, Handy Ch, Saffo P, Senge P. 1997. Looking ahead: Implications of the present. In: Harvard Business Review, sept-oct 1997.

Frawley W. 1992. Linguistic Semantics. Hilsdale, Hove and London: Lawrence Erlbaum Associates.

Haug PJ, Ranum DL, Frederick PR. 1990. Computerized extraction of coded findings from free-text radiologic reports. Radiology, 174, 543 - 548.

Jackendoff R. 1988. Conceptual semantics. In Eco U et al. (eds.) *Meaning and mental representation*. Bloomington: Indiana University Press, 81-97.

Jardel JP. 1989. Opening address for the International Working Conference on Natural Language Processing of the International Medical Informatics Association. In: Scherrer JR, Côté RA, Mandil SH (eds.) Computerized Natural Language Processing for Knowledge Engineering. Elsevier Science Publishers, p 1.

Kripke S. 1972. Naming and Necessity. In Davidson D & Harman G (eds.) *Semantics of* *natural language*. Dordrecht: Reidel, 253-355.

Lakoff G. 1988. Cogintive semantics. In Eco U et al. (eds.) *Meaning and mental representation*. Bloomington: Indiana University Press, 119-154.

Lin R, Lenert L, Middleton B, Shiffman S. 1991. A free-text processing system to capture physical findings: Canonical Phrase Identification System (CAPIS). Proc-Annu-Symp-Comput-Appl-Med-Care, 843-7.

Mahesh K. 1996. *Ontology Development for Machine Translation: ideology and methodology*. Technical Report MCCS-96-292, Computing Research Laboratory, New Mexico State University, Las Cruces, NM.

Mahesh K and Nirenburg S. 1995. A situated ontology for practical NLP. In *Proceedings of the Workshop on Basic Ontological Issues in Knowledge Sharing, IJCAI-95*. Montreal, Canada.

Miller GA, R. Beckwidth, C. Fellbaum, D. Gross, and K.J. Miller. 1990. Introduction to WordNet: An On-line Lexical Database, International Journal of Lexicography ¾, 235-244.

Mommaerts JL, Ceusters W, Deville G. 1994. Are taggers for general language useful for medical sublanguage ? A case-study with DILEMMA (Dutch). In: Beckers WPA, ten Hoopen AJ (eds) Proceedings of MIC'94, Velthoven, The Netherlands, 25-26/11/94, 283-290.

Paulussen, Hans & Willy Martin. 1992. *DILEMMA-2: a lemmatizer-tagger for medical abstracts*, in Proceedings of the Third Conference on Applied Natural Language Processing (ACL), Trento, 141-146.

Quine W. 1953. Two Dogma's of Empiricism. In Quine W (ed.) *From a logical point of view*, New York.

Ranum DL. 1988. Knowledge based understanding of radiology text. Proceedings of the 12th Annual Symposium on Computer Applications in Medical Care, Washington, DC, 141 - 145.

Rector AL, Rogers JE, Pole P. 1996. The GALEN High Level Ontology. In Brender J, Christensen JP, Scherrer J-R, McNair P (eds.) *MIE 96 Proceedings*. Amsterdam: IOS Press, 174-178.

Rossi-Mori A. 1994. Towards a new generation of terminologies and coding systems. In Barahona P & Christensen JP (eds.) *Knowledge and decisions in health telematics*. Amsterdam: IOS Press, 208-212.

Sager N, Friedman C, Lyman MS. 1987. Medcial Language Processing: Computer Management of Narrative Data. Reading, MA: Addison - Wesley.

Sager N, Lyman MS, Bucknall C, Nhan N, Tick LJ. 1994. Natural language processing and the representation of clinical data. J. Am Med Infomatics Assoc, 1, 142-160.

Searle JR, Kiefer F, Bierwisch M (eds.) 1980. *Speech Act Theory and Pragmatics*. Dordrecht: Reidel.

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

Welsh C. 1988. *On the non-existence of natural kind terms as a linguistically relevant category.* Paper presented at the Liguistic Society of America, New Orleans, LA.

Wiederhold G. 1980. Databases in healthcare. Stanford University, Computer Science Department, Report No. STAN-CS-80-790.