Methods of Information in Medicine
© F. K. Schattauer Verlagsgesellschaft mbH (1996)

**P. Spyns**

# Natural Language Processing in Medicine: An Overview

Division of Medical Informatics,
State University Gent,
Belgium

**Abstract:** An overview is given of natural language processing applications in medicine. An attempt has been made to enumerate the most important and known international projects and to summarize their goals, principles, methods and results. A section is devoted to projects involving the Dutch language. A more general discussion about the two fundamental approaches concerning medical language understanding is provided. An extensive bibliography may be useful for those wishing to explore this research domain.

***Keywords:*** Natural Language Processing, Computational Linguistics, Medical Language Understanding

## 1. Introduction

At scientific congresses and in various medical informatics journals, much attention has recently been paid to the electronic medical record (EMR) [1, 2]. The EMR is supposed to perform better than the actual paper-based record with respect to physical availability, ease of reading, completeness and, more importantly, accessibility of data [3]. As a general rule, the physician uses documents of the medical record as a memory support when the patient returns to hospital. The patient discharge summary, as a synthesis of the (previous) patient stay, is well suited for such a task. Much information (especially the patient discharge summary) is stored in free-text form. The use of natural language does not facilitate easy access to the wealth of information in the EMR. However, natural language still is the most frequently used and easiest way to transmit complex messages [4]. Hence, some authors consider the study and application of Natural Language Processing (NLP) in medicine as one of the most challenging issues in the field of medical information retrieval [5-7]. Natural Language Processing in Medicine is a promising research area that has already delivered some important solutions [6, 8, 9-13].

## 2. Material and Methods

We restrict the scope of this paper to the analysis of written texts. Issues such as report generation [14-16], respelling, paraphrasing, translation programs [17] and speech technology are not discussed [18]. Similarly, we do not present research projects that are mainly involved with domain modelling and knowledge representation[1] (e.g., UMLS [19], GALEN [11], the Canon Group's effort [20], and CEN TC251 WG2 [21]). Another criterion was the impact on current research in the field, which motivated the non-selection of older projects (e.g., [22]). Finally, statistical techniques (e.g., CAPIS [23] or SAPHIRE[2] [24, 25]) or vectorspace methods (e.g., Radtrac [26-28]) are not presented in the current overview.

The first section (section 3) describes briefly some theoretical issues about NLP in general, and NLP in the biomedical area. Subsequently (section 4), 20 international projects are presented. A distinction is made between "complete" (section 4.2) and "partial" (section 4.3) NLP systems. The former have a complete analysis and understanding chain (from separate words over sentences to complete texts), while the latter are limited to the word level (possibly jumping from the word level to the text level). Depending on the goals of the NLP application, it may not be necessary (or possible) to perform a "complete analysis and understanding process". Special attention is paid to projects concerning Dutch[3] medical language processing (section 4.4). Although it is not a widespread language, an important research effort is devoted to NLP in medicine. Before the conclusion (section 6), another section (section 5) will present a summarizing table containing the most characteristic features of the mentioned "complete" NLP projects. It serves as a basis to discuss the more fundamental issue of language dependence and domain independence (LD/DI) versus language independence and domain dependence

[1] In the *Journal of the American Medical Informatics Association* several papers about NLP in medicine have been published, in particular concerning domain modelling and knowledge representation.

[2] Note that SAPHIRE is available on the Internet (ftp medir.ohsu.edu/pub/saphire) and has a WWW page (http://www.edu.ohsu.edu/post-saphire.html).

[3] The Dutch language is spoken in the northern part of Belgium (Flanders) and The Netherlands.

(LI/DD) that plays an important role in the architecture of large-scale NLP in medicine projects. Finally, the references, although not exhaustive, give an interesting overview for people involved in this particular research area.

## 3. NLP and Medicine

Language is a form of communication between humans. One human emits a message represented by a specific combination of acoustic or graphic signs to another person *(receiver)* who shares some common sense knowledge with the *sender* which should enable the receiver to understand the *message*[4]. It was the philosopher Charles Morris who introduced the triplet "syntax-semantics-pragmatics" [29]. He stated that the study of *pragmatics* encompasses the complete environment of a person who speaks or hears. This includes the study of *semantics* that is concerned with the relationship of expressions to their meaning. The study of *syntax* examines the properties and structure of a language. The lexicon and morphology are sublevels of the syntactic level and concern the study of words and word formation (inflection, derivation, and compounding).

Each level can produce ambiguities that can be resolved by the subsequent levels. Some ambiguities can only be resolved by the knowledge of the real state-of-affairs that is described by the sentence. A good and complete NLP system should be able to handle (to a certain extent) all the mentioned levels. The question is how to integrate the three main levels. Some propose a cascaded architecture where the ambiguities of the lower levels are resolved sequentially by the subsequent levels. Another possibility is to work on one level and activate the "linguistic machinery" of the "higher" levels when the current level leaves too many ambiguities unresolved (integrated architecture).

Currently, in the field of linguistics it is considered as useful to restrict research to well-defined *sublanguages*. A sublanguage is a technical language that is used by the various actors in the technical field to pass specific messages. A technical language presents some characteristics that differentiate it from the general language. An evident point of difference is the *vocabulary*. In every sublanguage, normal words can be found in their "normal" meaning. But a considerable amount of general language words can take a more restricted and specific meaning in the context of a sublanguage. Other words have a general meaning, next to a more technical one, of which only the latter is used in the context of the sublanguage. Finally, for each technical domain there exists a large amount of very specific vocabularies that are mostly exclusively used in that particular technical domain [30]. Next to the vocabulary, the *sentence construction rules* also differ from the normal construction rules. Some expressions are used in a more concise way than is the case in normal language (omission of words that are not strictly necessary; telegram style is commonly used in medical documents [31]). Not all the possibilities of the general language are exploited. For instance, a patient discharge summary generally contains verbs only in the third person singular. Questions are seldom found in medical documents. Therefore, it is easier to build linguistic tools for sublanguages than for a general language. Every sublanguage also has its idiosyncratic expressions, which are difficult to understand when used outside the technical domain. These special expressions are created to describe adequately and concisely situations or objects that are typical of the concerned technical domain.

Although research in the medical informatics field only recently gained popularity, researchers applied *NLP techniques* on medical vocabulary for some time. Because *medical jargon* consists of many terms that are composed of Latin and/or Greek parts, it was attempted to describe the word formation rules for such words [32]. These rules would also determine the meaning of the entire word in function of its composing parts [33]. Such a complex word is very transparent across the various languages due to its Latin/Greek origin. Therefore, another characteristic of the medical sublanguage is the quasi-universality of the *medical notions*. For instance, appendectomy means the same worldwide [34]. This also holds for many non-transparent words. Moreover, *medical practice* in general is more or less universal. Physicians generally have a common way of examining and treating their patients. Within the same language, they express themselves in a similar way. Even the ways of storing and disseminating information, show similarities. For example, communication between a treating physician and a general practitioner about a patient generally takes place by means of a discharge summary. As both the expediter and receiver of the summary are physicians (treating the same patient), they share an amount of medical knowledge that will not be addressed explicitly in the discharge summary. The application domain is thus much more limited than in general language, which is advantageous for the development of language-processing tools.

These characteristics favored the application of Artificial Intelligence (AI) techniques for NLP in the biomedical area. Concepts like frames, scripts, and domain modelling became common terminology. These AI constructs try to make computers reason with premises and allow them to deduce valid and consistent conclusions from the original data (deductive reasoning). As the understanding of natural language is one of the most prominent features of intelligent human behavior, it is evident that AI, which wants to simulate intelligent human behavior by means of machines, devotes much attention to NLP. The emphasis lies on modelling, knowledge representation and inferencing (problem solving) for a particular application domain. The generality of syntactic knowledge was partly traded for domain-specific knowledge in order to simulate language understanding in a particular domain [35, p. 308].

From the end of the 1980s onward, computational linguistics and artificial intelligence were jointly (re-)applied to the medical field. Medicine as a science with an almost universal conceptual framework provides a suitable application domain for a linguistic approach privileging the underlying semantic ur

[4] As written language is considered to be a representation of spoken language, most authors in theoretical linguistics build their theories considering spoken language as the starting point.

versals with the application of AI techniques for knowledge inferencing. In addition, the medical language being a sublanguage reduces the set of linguistic problems to solve [36, 37] since not all the linguistic phenomena of a general language are present in a sublanguage. Influenced by progress in AI, specific theories (e.g,. Discourse Representation Theory) were developed to cope with linguistic phenomena that transcend the sentence boundary (coreferenciality, reference resolution, temporal reasoning). However, only few of the large-scale NLP in medical systems include a pragmatic level that treats these discourse phenomena. Currently, most attention is oriented towards domain modelling, of which the representation formalism is an important aspect.

## 4. NLP in Medicine

### *4.1 Preliminary Remarks*

Many of the partial NLP systems (section 4.3) operate jointly on the word level (morphological analysis) and the conceptual level (semantic analysis) and are generally (semi-)automatic encoding tools. In general, complete NLP systems (section 4.2) aim at the representation of the knowledge expressed by the sentences of the document. An important objective is information retrieval by means of user-defined queries. Indexing and encoding are derived applications.

The presented complete projects apply various methods to analyze the document and eventually represent its content. Some are primarily syntactically driven (LSP-MLP: section 4.2.1), others are essentially semantically based (MEDLEE: section 4.2.4, SPRUS: section 4.2.5, MEDI-CAT: 4.2.11). Many of these complete NLP systems have a language-independent concept system that validates the results of sentence analysis. Only medically sensible interpretations are allowed. The majority of the complete systems uses a different way of combining the syntactic and semantic levels. Some follow the sequential way (cascaded architecture), which implies a clear separation between the levels (SPECIALIST: section 4.2.2, MENELAS: section 4.2.6, RIME: section 4.2.8, MEDITAS: section 4.2.9). Others favor an integrated architecture; both levels are interwoven. In the case of SYMTEXT (section 4.2.5), ARISTOTE (section 4.2.7) and METEXA (section 4.2.10), the syntactic level activates the semantic, while in the RECIT system (section 4.2.3) the semantic level clearly outweighs the syntactic level. The choice between an integrated and cascaded architecture will be the main topic of the discussion (section 5).

### *4.2 Complete NLP Systems*

#### 4.2.1 The Linguistic String Project-Medical Language Processor

The Linguistic String Project-Medical Language Processor (LSP-MLP) of New York University is the first (and up till now the longest lasting) large-scale project about NLP in medicine [7, 12, 38, 39–43]. The LSP-MLP aims at enabling physicians to extract and summarize sign/symptom information, drug dosage and response data, to identify possible side effects of medications and to highlight or flag data items [44]. In short, tasks commonly denominated by the term "intelligent information retrieval".

Some years ago, the LSP-MLP has also been ported to French and is being ported to German, which illustrates the general applicability of its methodology and approach [45–50]. The reason for its generality lies in the use of a well-defined underlying linguistic theory (String Grammar), based on the distributionalism of Harris, and a scientifically based sublanguage approach [36, 51–53].

The following six steps are typical for the LSP-MLP processing chain:

1. The syntactic parsing stage. The parsing module structures the sentences of a medical document and represents the dependencies by means of parse trees. The parser can handle conjunctions.
2. The semantic selection stage. The semantic selection module uses the co-occurrence patterns to improve the parsing tree by resolving cases of structural ambiguity. Also, semantic characterization of parts of the parse tree is done on this level.
3. The transformation stage. The transformation module fills in gaps due to conjunction ellipses, reduces all sentence types to the affirmative type, completes relative sentences and regroups verbal splits.
4. The regularization stage. The regularization module transforms the semantically augmented parse tree into a canonical tree consisting of elementary sentences that correspond to the basic sublanguage sentence types. The inflected forms are replaced by their canonical form and the semantic host and modifiers are identified.
5. The information formatting stage. The formatting module maps the words of the elementary sentences into the appropriate fields of a format tree and constructs a binary connective-format tree for each sentence with the connectives as parent, and the phrases on which it operates as left and right children.
6. The normalization stage. The normalization module recovers implicit knowledge when possible and maps the format trees into the relational database structure.

All the information is stored in a relational database. Its columns represent semantic information templates ("information formats") (e.g., pt [patient], med [medication], body part, diag [diagnosis], etc.) while the rows (the sentences of a document) contain the normalized words ("strings") and sentence parts of the document [54].

Retrieval of information from a document in the database is done by means of SQL queries. In contrast to systems with a semantic and/or pragmatic level of analysis, the central notions of the LSP-MLP are the "string" (literal) and some conceptual primitives. Neither does an underlying semantic lattice with interrelated concepts exist [55]. A model-theoretical representation of the content of a document is not possible, which implies that deduction of implicit knowledge is only feasible by means of "external" special-purpose programs (e.g., the time program [12: chapter 9]).

This concerns especially the sentence and text analysis outside the LSP-MLP project. With respect to medical morphology, research has never ended (cf. sections 4.3.1 and 4.3.2).

Next to the "core NLP programs and knowledge bases" (lexicon, grammar, co-occurrence patterns, semantic labels, information formats, etc.) several important utility programs have been developed for large-scale corpus analysis [51] and (semi-)automated extension of the dictionary [30]. These programs largely facilitate the extension of the various knowledge bases (lexicon, grammar, co-occurrence patterns) and the porting of the application to another domain or language.

The LSP-MLP also acts as kind of "NLP family founding father". The Pundit system [56] is a Prolog implementation of the LSP parser and grammar, which was used by the US Navy for the analysis of repair reports [57]. In turn, the Pundit parser and grammar are reused in the Specialist system (cf. section 4.2.2), which brings them back to the area of NLP applications in medicine.

More recently, research concerning the LSP-MLP focused on the automatic encoding into SNOMED codes [7] and on the retrieval of information from discharge summaries for hospital management and clinical research [58, 59]. With respect to the information retrieval task, the system obtains a precision of 98.6% and a recall of 92.5% for the test sets [7]. The latest work includes the use of Standardized General Mark-up Language and World Wide Web Graphical User Interface technology to access and better visualize the requested information in the text (e.g., by highlighting words) [50].

### 4.2.2 Specialist

The Specialist system is being developed at the National Library of Medicine (NLM) [9, 60–64] and is intended to function as an information-extraction tool for biomedical knowledge bases, and the Medline abstracts in particular.

The lexicon was created using the MeSH (Medical Subject Headings), Dorland's Illustrated Medical Dictionary, as well as some general English dictionaries [65: p. 129]. An interesting utility (Lextool) facilitated the tedious work of dictionary extension [66]. This lexicon has been enlarged with the compound terms provided by the Meta-1 thesaurus (which is a meta-thesaurus of concepts that do not always appear in the medical vocabulary). Meta-1 has been derived from the Unified Medical Language System (UMLS), also developed at the NLM. At the moment, there are some 65,000 entries (canonical forms). The lexicon is stored as a relational database consisting of 10 tables or relations. However, the database tables are not completely normalized.

All the modules necessary for large-scale processing of real texts are present:

1. The Specialist system comprises a pre-processing module which takes into account abbreviations, acronyms and chemical terminology.
2. Besides the lexicon containing all the required linguistic information, a category guesser for unknown words is available.
3. The sentence analyzer, based on the Pundit-system[6] that reuses the Restriction Grammar-formalism (RG) (cf. infra [57, 56]), produces underspecified syntactic structures [67].
4. Concerning the semantic processing, predicate argument structures are produced whereby the arguments are labelled with semantic case roles.

To perform quick lookups in the dictionary and to determine the semantic links, MeshTool and MeshLink were developed. Together with the Meta-1 browser, they guarantee an easy extension of the domain knowledge. A large part of the semantic types used for the selectional restrictions of the Specialist system come from the UMLS [19]. Both the semantic type and the syntactic features characterize a lexical entry [68: p. 220]. A semantic network of 131 semantic types has already been built. These types can be attributed to each term of the metathesaurus (in total some 222,927 concepts). Some 50 links between semantic types have been defined.

The information-retrieving module operates mainly by combining conceptual structures, an indexing lexicon, i.e., the ULMS Semantic Network types and concepts, and Boolean operators. The terms of the indexing vocabulary function as entry points to the stored texts (abstracts and citation records of the Medline). The Medline database can already be consulted on a large scale by means of Specialist. During the testing, it appeared that the Specialist system generally had to deduce the underlying concepts (due to its domain knowledge) which means that relatively few search terms could be mapped directly onto the concepts of the stored texts. It is possible to enter nominal phrases as search or index terms. The retrieval process functions interactively whereby the lexicon browsers are used to fine-tune the query and take full advantage of the internal logic of the indexing strategy. In 1992, some 45% of the queries and 55% of the titles of the bibliographic database were claimed to be analyzed correctly [68]. The Specialist system is implemented mainly in C and Quintus Prolog.

### 4.2.3 RECIT

The Centre d'Informatique Hospitalière of the Hôpital Cantonal de Genève is investigating an electronic archiving environment with NLP facilities [4, 69]. In a first phase, patient records were archived in view of their subsequent computational treatment [70]. At a later stage, the LSP-MLP (cf. supra) has been adapted for French [47–49, 71]. And finally, a proper NLP system, namely RECIT (REprésentation du Contenu Informationnel des Textes médicaux) [5, 72–74] has been developed using an alternative analysis method called Proximity Processing [75]. The aim was to implement a robust and multilingual system able to analyze medical sentences and jargon, and to preserve the meaning of free text into a language-independent knowledge representation [6, 76]. Also, text generation is intensively studied [77].

Proximity Processing uses as early as possible during the sentence analysis, semantic domain knowledge; while the syntactic knowledge is only locally used for disambiguation purposes. As it concerns medical sublanguage (with many sentences written in a telegram style) (cf. [36] for a detailed study of the medical sublanguage), this method can be successfully applied. In addition, the robustness of the parser is enhanced. As

[6] The Pundit system as well as the Specialist lexicon and lexical tools are available for research purposes.

medical texts can contain many agrammaticalities, this issue cannot be neglected.

This alternative approach relies on the following principles: at the proximity of two words, the probability of one word determining (modifying) the other is very high. Thus, words next to each other are regrouped according to local syntactic and semantic compatibility rules ("proximity rules") [78].

The aggregation by proximity consists of six stages:

1. Decomposition of the text into paragraphs, and then into sentences.
2. Decomposition of each sentence into simple words, idiomatic expressions and abbreviations that are dictionary entries by means of morphological analysis and dictionary lookup of syntactic and semantic information.
3. Treatment of frequent associations, allowing the recognition of special expressions frequently used in the medical domain (temporal expressions, lab results, negations, pronominal verbs, etc.).
4. Resolution of syntactic ambiguities (syntactically driven contextual disambiguation rules).
5. Treatment of grammatical associations in so far as semantic compatibility rules exist.
6. Word grouping by functional group processing (organizing sentences into functional units).

The aggregation of words into word roups is a recursive process that ends hen the complete sentence consists of ome of these word groups. Using this ethod, sensible global results on the rounds of partial sentence analysis re-ults are apparently obtained.

Once the sentence is analyzed by aplying proximity rules, meaningful conepts are represented by means of the onceptual Graph (CG) formalism 9–81], that is becoming a de facto andard for medical knowledge repre-ntation [cf. also 82]. Such a semantilly-oriented approach is less language pendent than the more syntactically-ented approach. The RECIT system operational in French, English and, to lesser extent, German. The RECIT stem comprises, besides the analyzer d the proximity rules, a typology of ncepts, a dictionary with syntactic d semantic information, a set of conceptual relationships, and a set of canonical conceptual schemata. This semantic information presently relies on the medical model developed during the Galen (currently called the Galen-in-Use) project [11, 83, 84]. The RECIT system is also able (with a strong focus on nominal constructions) to generate (and thus to translate) medical expressions in natural language (NL) thanks to the use of the language-independent conceptual model [16, 85].

The RECIT system is implemented in Quintus Prolog running X-Windows (X11R4) and OSF Motif (for Sun, DEC and Hewlett Packard computers). A PC-based version exists as well. Actually, the system works mainly for nominal phrases and is said to attain a 95% success ratio. Reference resolution, pronominal and complex verbal constructions (e.g., relative sentences) still require more research. The lexicon currently contains more than 3,000 entries for French and English. Tools to extend the dictionary semi-automatically have also been developed [86]. When retrieving information, the CG of the query is matched with the CGs of the texts stored in the textual database which, in case of success, are returned to the user. An overview of the functionalities aimed at is given in [69] and [6].

#### 4.2.4 MEDLEE

The Columbia University of New York (together with the Columbia Presbyterian Medical Center) has developed an NLP system (MEDical Language Extraction and Encoding System) that identifies clinical information in narrative reports and transforms this textual information into a structured and conceptual representation [87]. The main goal is to represent the knowledge of chest X-ray radiology reports, store it in a database and allow physicians to query the knowledge base by means of a controlled vocabulary. Another realization is the integration of the NLP module with an automated decision-support system [88].

Although the MEDLEE system is primarily semantically driven, the necessity of integrating syntactic knowledge is recognized: the development of a syntactic grammar is foreseen [89]. The semantic grammar consists of 350 DCG rules, specifying well-defined semantic patterns, their interpretations, and the underlying target structures [87] (corresponding to a formal domain model, based on the information formats of the LSP-MLP; cf. section 4.2.1) into which they should be mapped. The grammar rules are directly interpretable by Prolog. Half a person year was devoted to the development of the semantic grammar.

Three separate phases of processing can be distinguished [88]:

1. The parsing phase. It determines the structure of the text and generates the preliminary structured output form for the clinical information. The parser uses a (semantic) grammar and a lexicon (containing 1,720 single-word entries and 1,400 multi-word phrases).
2. The phrase-regularization phase, which combines the structured outputs of noncontiguous expressions and standardizes them so that they correspond to the appropriate regular form, using a mapping knowledge base (consisting of the structural output forms of multi-word phrases that can be decomposed).
3. The encoding phase, which maps the standard forms into unique concepts associated with the controlled vocabulary using a synonym knowledge base that consists of standard forms and their corresponding concepts in the controlled vocabulary (the Medical Entities Dictionary [90]).

Note that the representation formalism used is the Conceptual Graph formalism [79]. Something more particular is the admission of phrases in the lexicon, because for some combinations it is said to be more efficient and more precise to handle these expressions as single units (e.g., "could not be evaluated"). The semantic lexicon also contains entries that are marked as irrelevant for semantic processing.

An evaluation of the MEDLEE system coupled to a medical decision-support expert system established that it did not behave significantly differently from a group of physicians who had to flag the presence or absence of six clinical conditions for a set of 200 admission chest radiography reports [91]. The system attains a recall level of 81% and a precision level of 98%.

Another realization of this group is AQUA (A QUery Analyzer), a natural-language front-end to an information retrieval system. It is implemented in Prolog, uses DCG, is based on the UMLS Semantic Net, and stores the conceptual representation as a CG. The prototype still needs to be modified and has not yet undergone thorough evaluation [92].

#### 4.2.5 SPRUS-SymText

A team of the University of Utah (Salt Lake City) has recently developed SymText (Symbolic Text Processor). Currently, concepts and words of 10 texts concerning chest X-ray reports are analyzed to fill tables consisting of conceptual slots. Therefore, SymText relies on an ATN grammar and parser coupled to a transformational grammar that is responsible for the final format of the parsing tree. For the clinical reasoning part, a model based on a Bayesian network is applied [93]. Several different versions of this mixed syntactic/semantic parser have been tested. Ultimately, SymText will function as a NLP server to an information retrieval system.

A previous development of the same group in the same clinical area is SPRUS (Special Purpose Radiology Understanding System). Although SPRUS is a purely semantically-oriented system, the developers recognized the need for (modest) syntactic processing [94] which has been integrated in Symtext.

SPRUS is a typically AI frame-based system consisting of three levels of processing:

1. Convert the words of the text into relevant pointers to text frames consisting of the dictionary element, a keyword, a semantic element type (finding or location), and a variable number of necessary modifiers for that dictionary element.
2. Correctly associate the findings and locations present in the text by filling special-purpose memory structures which represent expectations about the stereotypic associations for findings and locations.
3. Resolve the problems related to incomplete or too general information.

Morphological analysis is not mentioned [95], but the use of a thesaurus should take care of terminological (and probably also morphological) variations. Incomplete or implicit information can be dealt with thanks to the hierarchical organization of the dictionary elements. The reasoning parts are expectation-driven; a context of frames built by the word-conversion routines determines the possible inference process.

When processing 209 chest radiographic reports, the most recent version of SPRUS is said to attain a correct-positive rate of 87% with an unreliable data rate of 5% for radiographic findings and for diagnostic interpretations a 95% and 6% rate, respectively. No data on the Symtext system are provided yet, but, instead, an evaluation (involving 10 chest X-ray reports) of various parsing strategies is available [93].

#### 4.2.6 Hélène-NLPAD-Ménélas

The Service d'Informatique Médicale (Assistance Publique, Hôpitaux de Paris and INSERM U194) has implemented several systems that aim at the linguistic analysis of complete patient discharge summaries and deep knowledge processing of their content (Hélène [55, 96], NLPAD [97] and Ménélas [98–100]. Whilst Hélène and NLPAD (Natural Language Processing of Patient Discharge Summaries) are limited to the domain of thyroid cancers [101], Ménélas (An Access System for Medical Records using Natural Language) covers the subdomain of cardiology and cardiac surgery [102].

Hélène and NLPAD are based on the sequence of four basic components (the pre-processor is specific for Ménélas):

1. The pre-processing module, to expand abbreviations, mark sentence boundaries, aggregate typical medical expressions and flag dates.
2. The morphological module, to determine the linguistic characteristics of a word.
3. The syntactic module, to determine the syntactic characteristics of the words in a sentence and to establish their interdependence.
4. The semantic module, to produce a conceptual interpretation of the sentences of a text.
5. The pragmatic module, to detect the relevant facts of the situation described in a text.

This approach is in contrast with the Geneva system (cf. section 4.2.3) that attributes the primordial role to semantics and with the LSP-MLP (cf. section 4.2.1). In the latter case, the semantic module is organized completely differently and the pragmatic component, strictly spoken, does not form part of the main NLP system. The Ménélas architecture clearly distinguishes the five phases due to the polylinguality of the system (French, English and Dutch) [100]. Specific for all these applications is the presence of a pragmatic module that handles discourse phenomena and deduces implicit knowledge [101, 103–105]. The Ménélas system has partly inherited this pragmatic component [106], while the French morphosyntactic (a simple word lexicon of more than 40,000 lemmas and a bottom-up chart parser with more than 300 DCG-like phrase structure rules) and semantic parts come from the IBM Kalipso system [107, 108]. The English Ménélas parts covering morphosyntax and semantics (some 3,312 unique head words and a GPSG-like grammar containing 535 ID-rules, 21 metarules, 76 propagation rules, 131 default rules, and 26 linear precedence rules) are mainly extensions of the Alvey Tools [109–112] Hélène and NLPAD use the Lexical Functional Grammar formalism [113] The overall aim of the Ménélas project and the Dutch parts in particular, will be discussed more extensively below (cf. section 4.4.1). Its main goal is to create and enhance user-level services (e.g., ICD-9 encoding [114] and information retrieval [115]) by extracting medical information from free text.

During the Ménélas project, extra tools were implemented to facilitate corpus analysis and the development and extension of the various knowledge bases [116]. More fundamentally, a new method has been developed to link semantic entries to concepts of an application domain [117, 118]. It consists in projecting linguistic relations to the normalized model of the domain and selecting the conceptual path that best links the corresponding notions. As a result, identical projections can be assigned to linguistic paraphrases, and lin

guistically similar expressions that have different meanings can be distinguished.

The pragmatic analyzer checks the redundancy and validity of the conceptual graphs delivered by the semantic component. These graphs can be adjusted (e.g., by changing the attachment point of prepositional phrases on grounds of the available domain knowledge) [119, 120]. The pragmatic analyzer uses a concept-type tree, over 500 reference models, a relational tree, and signatures (totalling more than 2,000 atomic types and more than 6,000 instances).

It comprises four main steps:

1. Semantic normalization (cf. supra) by which several "linguistic graphs" are obtained. They constitute concurrent interpretations of a sentence (or sentence segment). These CGs are "syntactically" cleared without the use of any domain-specific knowledge, and converted into "conceptual graphs" with the help of domain knowledge [121].
2. Synchronic inference by which concurrent CGs are completed in parallel, or rejected according to domain-specific knowledge sources.
3. Ambiguity resolution, a procedure heuristically based on the information content of a CG that chooses the best CG among the remaining fully completed ones.
4. Diachronic inferencing and integration that support temporal reasoning through date comparison and that store the selected CG in a list.

The French and Dutch language-specific components are implemented in Prolog by BIM; the English language-dependent parts in CMU LISP, and the language-independent parts in Prolog by BIM and Allegro Common LISP. The information retrieval system uses Prolog by BIM. The utilities and preprocessor are written in C and C++. The NLP system runs on a Sun Sparc station using X11R5 OSF Motif (the French linguistic components also on IBM RS/6000) while the consultation service communicates with the user by means of the MS-Windows graphical interface on a client PC).

#### 4.2.7 Aristote

Another system for the French language is Aristote, implemented in LPA Prolog at the Laboratoire d'Informatique Médicale (Faculté de Médecine in Montpellier) [122–124]. It deals with the creation of databases for research purposes of analyzed pathology reports with the associated radiology images. The processed texts belong to the domain of thyroid cancer.

The architecture of the Aristote system follows a more classical scheme:

1. The Reader processes words by isolating them without considering their functional role or their meaning.
2. The Syntactic-Semantic Parser processes sentences by working out the meaning of each sentence from the syntactic and semantic information in the words.
3. The Constructor processes the text by linking up and structuring the meaning of the text from the meaning of the sentences (paying attention to the reference resolution and discourse analysis).

A model-theoretical description of the application domain was needed. The representation formalism retained was – once again – the Conceptual Graphs formalism [79]. At present, Aristote has a rather limited coverage (4,000 lexical entries and 27 sentences from 3 texts can be correctly processed) [125]. Extension of the coverage is planned and in progress. Currently, an NL interface is being developed which converts user-defined queries into the conceptual graph form so that conventional graph projection-based query techniques can be applied [126].

#### 4.2.8 Rime

The Rime system (Recherche d'Informations Médicales) was developed at the Laboratoire Génie Informatique and the Service d'Informatique Médicale (Grenoble) [127–130]. Its purpose is to link an indexed textual description with the corresponding radiograph so that the user of the Rime system can retrieve the matching radiograph by an NL query. In essence, Rime is an indexing system that uses NLP to retrieve the stored information ("access by content").

With respect to the linguistic aspects, there are three processes involved:

1. The morphological process identifies the words and deduces the virtual attributes of the words through the consultation of a dictionary.
2. The syntactic process deduces the actual attributes of the words (using a precedence matrix and a list of ambiguous patterns), builds and nominates some specific syntactic syntagmas, indicates deletions, and validates the resolution of pronominal anaphoras.
3. The semantic process builds and nominates semantic structures according to the Rime semantic model and solves some interstructural tasks (e.g., nominal anaphoras).

A semantic model for radiology has been described as a set of internally related concepts using a self-defined Conceptual Language; 60 rules, 10 operators, and 20 types or classes are created. Once again, the "deep understanding" allows intelligent information retrieval. Many expressions are translated into the terms of the semantic model. A sentence is transduced into a semantic binary tree whose leaves are medical terms. The non-terminal nodes are semantic operators. Each tree is built according to the rules determined by the semantic model.

Basically, the query mechanism makes use of Boolean combinations of domain concepts. Retrieving the documents is done according to the similarity between the tree representation of the question and the ones of the stored texts. Specific for the Rime approach is that the syntactic knowledge is represented as a "precedence matrix" instead of a more classical context-free grammar. The precedence matrix contains the "allowed passes" from one syntactic category to another.

The Rime prototype [130] consists of some 5,000 lines of Prolog (for a Sun workstation). This accounts for a very basic morphological analyzer, a syntactic parser, and the complete semantic processor. Additional work on the interfaces for the lexical, syntactic and semantic knowledge extension tools is scheduled.

#### 4.2.9 MediTas

The Medical Text Analysis (MediTas) system is developed at the Department of Medical Informatics (Georg-August-Universität Göttingen). It is a system aiming at the development of intelligent full-text retrieval systems making use of NLP methods.

The MediTas system has four main levels:

1. Surface structuring. The structuring component decomposes recursively the continuous text of a document into coherent text segments. It is based on Augmented Transition Networks.
2. Syntactic analysis. Syntax analysis tries to discover the formal structure of an expression without explicitly taking into account aspects of meaning and situation context.
3. Semantic analysis. Medical reasoning is mainly concerned with semantic concepts and associations between them.
4. Pragmatic analysis. The consideration of the communication background (the so-called situation context such as medical knowledge, factual knowledge, experiences, and motivations of the communication partners) has to be taken into account and should result in an extended computer-processable description of facts related to a patient.

Only the pre-processor (level 1), the morphological and morphosyntactic modules (level 2) are completed [10, 131]. The syntactic parser (left to right, bottom up, and breadth first) is of Left Associative Grammar type (64 rules, 104 package rules and 1,000 auxiliary functions). The treated texts are cytopathological findings and reports (18,400 in total). The corpus contains 8,790 unique sentences (3,760 unique words), of which more than 92% are said to be correctly analyzed. The LAG parser can handle many linguistic phenomena but does not include coordination. According to the author, the MediTas system proves the feasibility and interest of complete syntactic analysis (as opposed to the more semantically-oriented systems). A pre-processor was implemented that expands the abbreviations and detects the sentence boundaries. The lexicon is a full-form dictionary. Further work on the automated systematized mapping (based on valency principles) of syntactic structures into a semantic representation using the Conceptual Graphs formalism [79] is undertaken [82, 132]. Special attention was paid to subtyping, negation and modality. The complete system is written in PC-Scheme (a LISP dialect) and runs on a PC. Without the semantic module, the MediTas system consists of some 4,000 lines of code.

#### 4.2.10 MeTexA

For German, another system exits: MeTexA (Medical Text Analysis). It was developed at the Department of Medical Informatics of the University of Hamburg. MediTas is able to analyze radiology reports of the thorax and to represent the contents by Conceptual Graphs [81]. Semantics and knowledge processing are the main issues of the MeTexA system. A module for speech processing is also foreseen [18].

The following components form the backbone of the system:

1. Lexicon. For each word, its base form, syntactic category and the lexical semantics are specified using references to the domain model and macro expansion.
2. Parser and grammar. The parser is a bottom-up parser augmented by a graph-unification mechanism that produces a feature matrix of attribute value pairs. However, this structure is not committed to any linguistic theory.
3. Semantic analysis. Each possible combination of syntactic constituents needs to be validated by a semantic relation between the corresponding main concepts.
4. Domain model. A type lattice, conceptual graphs and schemata are used.
5. Inferencing. It is an implementation of the resolution-based inferencing method [108], which infers new facts from the analyzed sentences.
6. Planning. Checklists, plans and scripts are used to generate expectations about the next incoming utterance.

Although German is a highly inflectional language, the MeTexA system does not use morpho-syntactic features, because the reports, written in a telegram style, often lack correct inflectional endings [133]. The radiology reports are also characterized by the frequent omission of verbs and the existence of coordination between nominal groups and prepositional groups. Therefore a principally semantically-driven approach (interleaved with syntactic analysis) was adopted favoring the analysis of noun phrases. One of the advantages of the system is its ability to process expressions that are agrammatical, but nevertheless used in the radiology reports. Currently, MeTexA is being applied to the analysis of clinical questions in the context of a clinical workstation.

Some theoretical notions of the semantic and knowledge inference levels correspond to similar methods used by the Ménélas project with respect to the semantic and pragmatic analyzers (cf. supra [118]). The MeTexA system is implemented in Prolog by BIM and runs on a Sun Sparc Station. The full-form dictionary contains some 1,000 forms (probably base forms); the grammar consists of 100 rules and there are about 400 concept types. The author reports that some 20–30 complete radiology reports can be processed, as well as many single sentences [134].

#### 4.2.11 Medi-cat

At the Division of Medical Informatics of the National University at Chiba (Japan), the Medi-cat system, implemented in MUMPS, indexes symptoms and diagnoses from discharge summaries using the SNOMED nomenclature as indexing lexicon [135]. The advantage is that the SNOMED code number can be the index instead of the complete language description. Another advantage is the nature of SNOMED; an internationally recognized standard that is regularly updated. However, an NLP step remains necessary (in particular concerning morphology and synonymy). After the morphological phase (with spell checking), all the possible inflected forms are combined with each other until there is a match with the longest consistent SNOMED paraphrase (= index). When this is not the case, semantic analysis is applied (e.g., transformation of adjectives into corresponding nouns, conversion of syno-

nyms) to arrive at a SNOMED expression. The Medi-cat system can also translate many SNOMED expressions into an ICD-9-CM expression [136].

Lately, the Medi-cat system has been changed into a complete NLP system that uses a specific medical semantic grammar to analyze diagnostic sentences, symptoms and some procedures from a volume of the New England Journal of Medicine. The sentence analyzer tries to fit the elements of the sentence into (medical) patterns using the (syntactic but predominantly semantic) labels attributed during the word analysis. These templates constitute the semantic grammar, which looks like a combination of the syntactic grammar and the co-occurrence patterns of the LSP-MLP (cf. supra). SNOMED III delivers the greater part of the classes of medical concepts and attributes needed for building the semantic grammar. The information is finally represented by means of database frames. The semantic categories function as slot labels, while the SNOMED codes constitute the contents of the slot. The database frames can be easily transformed into first-order logic formalisms and CGs. The database can be used for hospital practice (indexing), as well as scientific purposes (determination of disease profiles or supplying knowledge to expert systems) [8].

The Medi-cat system can be subdivided into five stages:

1. Morphological analysis; identification of components of words.
2. String labelling and classification to identify and label meaningful strings.
3. Parsing and pattern matching to identify meaningful units, phrases and complete sentences; disambiguation and regularization of the sentences.
4. Indexing to index the medical terms and modifiers using SNOMED.
5. Information formatting to transform the sentence into a frame structure in the database.

This system is comparable to the RECIT system concerning its domain-dependent and semantics-driven approach. Likewise, translation is attempted. However, the "core sentence analysis" follows the more classical grammar approach (such as the LSP-MLP) as opposed to the proximity processing of the RECIT system, but it differs from the LSP-MLP by its language-independent information representation.

Of the 79 test sentences (describing symptoms), 73.4% are correctly formatted which results in a 95.3% score concerning the indexing of medical terms [137]. The programs are written in M and run on a network of 12 SUN-4 computers under UNIX. The analysis of a sentence requires between 5 and 15 cpu-seconds. The database containing the dictionaries uses some 10 Mbytes. A bilingual English-Japanese dictionary requires some 50 Mbytes.

### *4.3 Partial NLP Systems*

#### 4.3.1 The "NIH" System

Although this application, developed at the National Institutes of Health (Bethesda), is probably out of use due to the recent development of UMLS and Specialist (cf. section 4.2.2), it represents a fundamental contribution to the field of medical morphological research. It primarily aimed at the automated decomposition of medical words into their composing morphemes [138, 139]. The underlying syntactic-semantic composition rules are described so that automated meaning analysis and paraphrasis can be done [140, 141]. This is useful for dictionary building and text-indexing. The algorithms were applied to pathology data (SNOP dictionary) and implemented in Prolog.

#### 4.3.2 The "Münster" System

At the Institut für Medizinische Informatik und Biomathematik of the University of Münster, much theoretical work on medical language processing, and in particular morphosemantics and automated indexing, was carried out [142, 143]. A segmentation program consisting of a segment dictionary (10,415 English and German segments), a rule system (256 rules in a hierarchical graph), and (statistically based) segmentation algorithms has been implemented [32, 144]. The goal is to represent the content of a medical document by means of the SNOMED nomenclature for automated indexing, automated abstracting and automated retrieval of medical information. As it concerns only medical expressions (nominal phrases), the semantics of the expressions are derived from the combination of the meaningful words and word parts of a sentence using SNOMED-descriptors and codes [145] (an idea re-used by almost all the "partial NLP systems" [cf. infra] and also retained by Sager for encoding purposes [7]). Therefore, sentence analysis is not done.

#### 4.3.3 SALBIDH

The SALBIDH system (System for Automated Lexicon-Based Indexing of Diagnoses Heidelberg) developed at the Department of Medical Informatics of the University of Heidelberg [146, 147] aims at the (semi-)automated indexing of medical diagnoses for retrieval purposes.

The system consists of three modules:

1. The pre-processor divides any input string into words, and replaces the abbreviations and corrects spelling errors (by string replacement).
2. The morphological analyzer decomposes medical compounds into their constituent parts, lemmatizes the inflected forms and discards word-(part)s that are irrelevant for indexing.
3. The semantic analyzer maps the remaining word(part)s into the most specific non-redundant SNOMED indices.

Some semantic reasoning takes place using the SNOMED nomenclature as underlying semantic model and its relationships as semantic associations. SALBIDH largely builds on results of the work of Wingert [145] (cf. supra). It is implemented in the 4th-generation language NATURAL2/ADABAS. What the authors call "lexicon" is not to be interpreted in a strictly linguistic sense, but more as a set of database relations.

#### 4.3.4 The "Gabrieli" System

Private companies have also entered the medical language-processing arena, namely the "Gabrieli Medical Information Systems" [148, 149]. This company has implemented a semantically-based NLP system (written in MUMPS) that is claimed to process a discharge sum-

mary in 15 minutes. Instead of full-fledged sentence analysis, the medically important terms are identified and transformed into a numerical code [150]. Subsequently, these codes (= the meaning) are combined (transcending the sentence boundaries). The various dictionaries are quite extensive (medical lexicon: 430,000, lexicon: 126,000, abbreviations: 13,000, synonyms: 36,000, spelling lexicon: 38,000, counter indication lexicon: 9,000; probably, this concerns the number of entries of full-form dictionaries).

A coding system has been developed to represent the meaning of the discharge summary. The Gabrieli system is able to fill (and print) a predetermined schematic representation of a medical document with facts appearing in an actual document. These "case descriptions" are also entered into a database for subsequent clinical research.

#### 4.3.5 The "Montpellier" System

At the Département de l'Information Médicale of the Hôpital Lapeyronie (Montpellier), a general indexing program has been developed. The SNOP-SNOMED nomenclature has been used as indexing vocabulary. A nominal group parser (based on Augmented Transition Networks) was implemented.

Another research item was the development of a system producing a tree, containing the exact text representation, this tree being the same as the tree the physician in charge of the patient could manually fill in during the patient visit [151]. The system includes dictionaries with the morphological and semantic attributes of the words, knowledge bases containing a description of the system's expectations, and a description of the facts about the patient whose discharge letter is currently analyzed.

More research on morphosemantic analysis of medical vocabulary and the medical language in general (knowledge modelling, medical dictionaries) is performed but without building a "real" system. The focus seems to be principally on the theoretical level [34, 152–154] in the style of Norton, Pacak and Dunham [140] and Wingert [145] (cf. supra).

### *4.4 Projects Involving Dutch*

#### 4.4.1 Ménélas

The primary objective of the Ménélas system (cf. section 4.2.6) was to make the information in a Patient Discharge Summary (PDS) available through advanced NLP and Knowledge Representation (KR) techniques [155, 156]. The information on the PDSs is stored in a logical representation (Conceptual Graph) that allows a rule-based inference engine to deduce and make implicit knowledge explicit. Information from the PDSs is retrieved in an intelligent way. Instead of using pattern recognition, the system compares the "meaning of the query" with the "meaning of the data". More complete and accurate answers to the queries are thus formulated. As a corollary, the Ménélas software can encode PDSs according to the ICD-9-CM nomenclature [114].

The Dutch modules are typically language-dependent: morphosyntactic analysis and specific parts of the semantic analysis (semantic dictionary and composition rules). With respect to the morphological level, there are currently some 100,000 full forms in the syntactic lexical database (which is some 8,000 non-inflected forms) [157, 158]. For the moment, the relational database (Sybase 4.9.1) contains mostly simple word forms. Neither complex word forms nor idiomatic expressions are yet handled in a conclusive manner. All the verb entries contain valency information. A recognizer characterizes the unknown word forms morphologically (cf. [159]). After the dictionary lookup (or activation of the morphological recognizer), all the nouns are checked on adjacent "noun neighbors" (for possible noun-noun compounding) [160]. Also, various contextual rules are activated so that the ambiguous morphological analyses can be made unique (or at least reduce the number of morphological readings) [161]. The syntactic analyzer for Dutch uses the Restriction Grammar (RG) as the underlying grammar formalism, which is the Prolog version of String Grammar [162]. RG is a logic grammar formalism (close to Definite Clause Grammar) that combines context-free rules with context-sensitive information [163]. The restrictions limit the combinatory of the context-free production rules by checking contextual information (e.g., agreement in number between the syntactic subject and the main verb of the sentence). A grammar (200 RG rules and 157 restrictions) for the Dutch medical language has been defined as the result of the study of some 50 sentences of PDSs [165]. The grammar was refined – but not exhaustively – after a test involving some extra 50 PDSs [99]. Basically, the RG parser does not differ substantially from a DCG parser. It is a top-down, left to right parser that returns only one complete parse tree. Each RG parse tree is transduced in an Annotated Parse Tree (APT) [157], which is the input format for the common semantic analyzer. The subsequent stages of the processing chain are language-independent (cf. 4.2.6).

#### 4.4.2 Multi-Tale

The Multi-Tale system (generation of MULTI-lingual specialized lexicons using augmented TAgger-LEmmatizers), is a syntactic-semantic tagger for medical sublanguages [165]. Prototypes are currently operational for full-text neurosurgical procedure reports in Dutch [166] or English [167]. Multi-Tale is a modular system in which pre-existing syntactic part-of-speech taggers (D-Tale for Dutch running under UNIX, and PC and Dilemma for English running on a PC [168]) have been augmented to produce semantic tags according to the CEN 1828:1995 ENV standard in medical informatics [169]. The English Multi-Tale tagger is embedded in Word for Windows and allows direct manipulation of the various lexicons and syntactic-semantic grammar. It operates in a bottom-up approach aggregating syntactically tagged words into phrases and complex noun groups with adjectives and prepositional attributes, using contextual information, but without full parsing. Validation showed that for the learning sample (5 reports), syntactic information has been correctly provided for most of the cases (recall: 93.3% and precision: 94.4%), while this is even better (recall and precision: 95.7%) for the testing sample (also consisting of 5 reports). Semantic tagging is also very successful

(recall: 91.3% and 89.3%, respectively, and precision: 92% and 94.8%, respectively) [170].

#### 4.4.3 Anthem

The main objective of the Anthem (Advanced Natural language interface for multilingual Text generation in Health care) project is to develop a portable, platform-independent prototype of a natural-language interface that accepts medical diagnoses in Dutch or French, and translates this input in Dutch, French or German [171]. At the same time, automatic encoding of diagnoses following the rules of standardized classification systems (ICD-10 and SNOMED International) is realized [172]. This requires the development of a generic, intelligent, multi-lingual, natural-language interface that can be integrated with heterogeneous medical applications [173]. Also, a formalized description of SNOMED International serving as interlingua is undertaken. Anthem is built around a semantic model that bridges the gap between purely linguistic semantics and conceptual semantics in the sublanguage of medical diagnostic expressions [174]. The semantic model has been designed according to well-defined principles [175]. It has been used to adapt the existing machine translation system CAT2 [176]. As a result, by using the same semantic structures, it is possible to produce translations, and to automatically encode diagnostic statements into ICD-10 [177].

#### 4.4.4 Dome

The Dome (DOcument Management in health care) project is a more global project in the area of document handling. The focus of the work is a document-management system for health-care applications in a hospital context. The Dome system, which is now at the specification stage, exploits state-of-the-art NLP technology and aims at striking a balance between practical and economic feasibility from the hospital's point of view, and completeness and robustness from the end-user's point of view. It is an attempt to integrate and enhance document-managing systems with existing NLP technology. The core Dome system is best described from a user point of view as a multimedia, hypertextual patient record [178], including dynamic content-based retrieval facilities. This system will offer access to patient information in a structured and flexible way in a clinical setting. The objectives of the Dome system are ambitious. They include report-based coding assistance, NLP-based indexing and retrieval (content-based document search), an expert-text component, statistics and data extraction components, multi-linguality (Dutch, English and French), authoring tools (quality assurance of input records, automatic formatting and link creation), and voice-based report entry and retrieval [179].

## 5. Discussion

Before discussing the different projects and approaches, some general remarks are made first. All the "complete" NLP systems fundamentally aim at two indispensable and complementary tasks, which enable subsequent information retrieval and processing, i.e., language analysis and knowledge representation. Depending on the focus of the different research groups (knowledge engineering or computational linguistics), they favor a language-independent/domain-dependent (LI/DD) or a language-dependent/domain-independent (LD/DI) approach. The latter implies full parsing strategies while the former considers the medical interlingua "obscured" by the syntactic representation. The view on the focus topic determines the manner to tackle the other topic: LD/DI sentence analysis leads to language-dependent knowledge representation, while LI/DD knowledge engineering evolves into domain-dependent language analysis. The LD/DI approach is better for sentence analysis but as no language-independent concepts are used, such systems (cf. the LSP-MLP) have more difficulties with the knowledge processing jobs (e.g., inference or translation). The LI/DD systems (cf. the extensions of Medi-cat), on the other hand, are better suited for information processing but the (semantic) grammars for linguistic analysis are more complex and require more efforts to be built (many semantic templates need to be established). We

**Table 1** Summary of the most characteristic features of the presented complete NLP applications.

| Name | Programming language | Linguistic theory | Hardware platform | Knowledge Representation | Knowledge Domain | Encoding | Natural language |
|---|---|---|---|---|---|---|---|
| LSP-MLP | Fortran | String Grammar | Sun | RDBMS tables | sublanguage patterns | SNOMED | Fr, Eng, Ger |
| Specialist | Prolog | Restriction Grammar | Sun | | UMLS | | English |
| Récit | Prolog | Proximity Processing | Sun/HP/PC | CG | GALEN | ICD-9-CM | Fr, Eng, Ger |
| Ménélas | Prolog, Lisp, C | DCG/GPSG RG | Sun | CG | "UMLS" + | ICD-9-CM | Fr, Eng, Dutch |
| Aristote | Prolog | DCG | PC | CG | | | French |
| Time | Prolog | "rime" | Sun | CG | "rime" | | French |
| Leditas | PC-scheme | LAG | PC | CG | | | German |
| MeTexA | Prolog | "metexa" | Sun | CG | "metexa" | "metexa" | German |
| Medi-cat | Mumps | "medi-cat" | Sun | "frames" | SNOMED | SNOMED | Eng, Jap |
| SymText | | ATN/TG | | frames | | | English |
| Medlee | | DCG (sem.) | | CG | UMLS | | English |

Eng = English, Fr = French, Ger = German, Jap = Japanese

are convinced that, in the end, results of syntactically driven (and, thus, language-dependent/domain-independent) sentence analysis have to be mapped as much as possible to a conceptual (and, thus, language-independent) representation (cf. Ménélas), which implies to a large extent integration of both approaches mentioned.

What can be learned from the comparison of the presented "complete" projects (cf. Table 1) ?

Many items are shared by the majority of the mentioned projects. Nearly all the "complete" NLP systems use Prolog (or Lisp[7]). These programming languages are, indeed, considered as the programming languages for AI and NLP "par excellence"[8] . Another salient item is the almost ubiquitous presence of the Conceptual Graph formalism as knowledge-representation language [79-81][9]. To be noted as well is the predilection for a (Sun) UNIX workstation as hardware platform, but some PC versions are available as well. Some consider that future widespread usage of NLP tools prohibits expensive workstations [180] and advocate that at least a run-time version for a PC be provided. Another interesting feature to be emphasized is the special appeal of radiology reports chosen as the test domain for many of these NLP systems. The structured form of these texts and their on-line availability are to a large extent responsible for their popularity for NLP purposes. More projects concern English (6) than French (5) and German (4). Only one project treats Dutch and one project Japanese.

An important point of difference on which we want to elaborate is the choice between a language-dependent/domain-independent versus a language-independent/domain-dependent approach for language analysis (cf. supra). The latter attributes a more prominent role to semantic processing. The syntactic analyzer is activated by the semantic processor and serves to confirm or reject hypotheses already made by the semantic processor. The former approach stays with the more classical cascaded architecture: the output of the syntactic processor is the input for the semantic analyzer. The two processes operate in a sequential way. This difference can be derived from Table 1 by the presence of an underlying formal linguistic theory (at the syntactic level) or not. The Rime, RECIT, MeTexA and Medi-cat projects use internally developed methods that strongly integrate syntax and semantics[10]. The other systems use or re-use and adapt existing linguistic theories (context-free grammars).

It is difficult to determine which of the two approaches is better. The obvious advantage of the semantically-driven approach is the robustness of the processor when confronted with agrammaticalities. A semantically-driven parser can more easily take care of agrammatical input thanks to the available knowledge of the domain model [181]. However, it is possible for a syntactically-driven approach (e.g., using a chart parser) to pass partially analyzed sentence chunks to the semantic analyzer that will try to aggregate them correctly. On the other hand, even for sentences that are correctly analyzed at the syntactic level, a number of "senseless parsing trees" (i.e., semantically invalid) can be passed on to the semantic processor. Semantically-driven analyzers immediately reject such meaningless cases. Another advantage claimed by the defenders of the language-independent/domain-dependent approach is the ease with which an already existing NLP application can be adapted to another language, since the concept system does not need to be altered [73]. Only some local syntactic rules (and the dictionary) have to be adapted. The corresponding disadvantage is the difficulty of porting the application (especially the semantic grammar) to another (sub)domain [183]. A new domain model must be defined, as well as new semantic rules for sentence analysis (together with an extension of the dictionary). Of course, when a syntactically-driven application is ported to a new domain, basically the same tasks need to be done, but they will not (or to a lesser degree) affect the interaction between syntax and semantics. Since the LI/DD or semantically-driven approach activates the syntactic level, the interaction between the two levels[11] needs to be redefined as well when moving to another domain.

A more fundamental problem is the absence of a uniform and general semantic model or concept system for medicine (although the Galen project – now called Galen-in-Use – is working on it [11]). That is the most important reason why we still prefer the LD/DI or syntactically driven method. Although many different linguistic theories and implementations exist, they basically share the same notions[12] (meta-level interlingua, such as subject, plural, transitivity, etc.) and strive to obtain the "same" result, i.e., a representation of the relations and functions of the different parts of a sentence. As is illustrated by some of the mentioned projects, it is feasible to have a conversion module, mapping the output of different language analyzers (same parser for different languages (LSP-MLP) and different parsers for different languages (Ménélas)) to a common (syntactic) representation, which is passed on to a common and partially language-independent semantic processor. This means that existing morphosyntactic processors can be re-used. No new set of "local syntactic rules interacting with the semantic level" has to be implemented. Of course, the time needed to implement the conversion module should also be taken into account. But this may prove to be a small inconvenience compared to the power and coverage of a large and comprehensive grammar that is at that moment ready to be

[7] PC-scheme being a Lisp dialect.

[8] The LSP-MLP is the exception. But when the LSP-MLP was implemented, Prolog hardly existed. However, Hirschmann and Puder ported parts of the LSP-MLP to Prolog [57].

[9] The LSP-MLP uses SQL (and the Relational Database paradigm) to represent the "knowledge" (cf. supra), the Conceptual Graph formalism not yet existing at that moment. The Medi-cat system allows the transformation of its frames into CGs [137].

[10] Aristote occupies an intermediate position: the syntactic level activates local semantic checks before passing the results of the sentence analysis to the semantic processor.

[11] We do not mean the interaction mechanism (i.e., the way the local syntactic rules are invoked), but the rules determining the moment of activation. Probably, they form part of the semantic grammar.

[12] At least when considering Indo-European languages.

re-used (together with its grammar and dictionary extension utilities). The example of porting the LSP-MLP to French [48, 49, 71] may be illustrative.

In the same spirit, it is possible to couple a morphosyntactic analyzer to another semantic processor (e.g., with another domain-modelling method). Here again, adaptations must be done (especially adaptations of semantic features in the lexicon). An illustration of this case is the combination of the Pundit system[13] and UMLS into the Specialist system [61, 62]. In principle, it would be possible to couple a language-dependent morphosyntactic parser to a language-independent semantic processor, whereas it would be hard to couple a domain-dependent system to a domain-independent or another domain-dependent application. Although this can seem senseless at first glance, it could be interesting to couple a domain-dependent application that does not handle discourse phenomena (such as reference resolution, causal and/or temporal reasoning) to another system that does process the mentioned items (= the pragmatic level of language understanding), depending on the similarity of the domain model and the knowledge-representation formalism.

The choice between a language-dependent/domain-independent versus a language-independent/domain-dependent method for language analysis is more important for domains with knowledge models consisting largely of language-independent concepts. Medical terminology and concepts are, indeed, highly standardized and, in general, can act as interlingua. However, even from the modelling perspective, any conceptual model cannot be totally language-independent as it necessarily reflects the judgment of a "categorizer" [183]. This remark echoes the theory of linguistic relativism proposed by Von Humboldt. Roughly summarized, it states that someone's knowledge and understanding of the world is related to his language. His ideas were later reworked and became known as the Sapir-Whorf hypothesis[14]. Sapir claims that the language habits of a community predispose certain choices of interpretation.

If one adds to this that a sublanguage, although it has a large idiosyncratic vocabulary (that in the case of medicine is fortunately rather conventional across languages), there still remains a set of general language words being either strictly general (and thus not medical at all) or possessing a medical meaning next to a general one [30, 37]. Although some words of the latter category have a medical meaning that is well circumscribed (e.g., "heart"), other words manifestly do not (e.g., "pain"). As these occur frequently[15] – otherwise these words would be strictly sublanguage words – our hypothesis is that these words will cause problems when applying an LI/DD approach to multilingual NLP applications. This is the reason why a "partial application" (cf. section 4.3) can be quite successful since only the strictly medical vocabulary is processed, avoiding the potentially problematic words. Probably, the most critical words are the verbs.

## 6. Conclusion

Briefly, as it is difficult to build large NLP applications for medicine (especially with respect to the semantic processing involving domain modelling, cf. [184]), we believe that the language-dependent/domain-independent approach for language analysis offers more flexibility due to the better separation between linguistic knowledge and domain knowledge. A syntax-driven approach with local domain-dependent semantic checks adds some of the advantages of the language-independent/domain-dependent to the LD/DI approach (robustness, early pruning of senseless parsing trees), but is harder to implement when more than one language is involved (and certainly when different parsing techniques are used). In addition to these pragmatic reasons, some acquired philosophical insights combined with results of sublanguage study support our point of view that it is better to start with a syntactically-driven approach when building multilingual NLP applications.

However, this is far from saying that the semantically-driven approach is senseless. Especially in medicine, where many concepts are language-independent, very good results are obtained by domain-dependent language analyzers – especially concerning agrammatical input (cf. sections 4.2.3 and 4.2.4). That is why we advocate the combination of a language-dependent syntactically-driven parser and a language-independent conceptual processor. NLP systems primarily based on domain knowledge inherently have some fundamental shortcomings, at least in our opinion. Of course, these disadvantages can be irrelevant for the intended purpose of the NLP application.

**Acknowledgments**

We wish to thank R. Baud and A.-M. Rassinoux for their remarks and suggestions.

REFERENCES[16]

1. McCray AT, Safran C, Chute C, Scherrer JR, eds. Natural Language and Medical Concept Representation. Meth Inform Med 1995: 34, 1/2 (special issue).
2. Scherrer JR, Coté R, Mandil S, eds. *Computerized Natural Medical Language Processing for Knowledge Engineering.* Amsterdam: North Holland Publ Comp, 1989.
3. Van Ginneken AM. Electronic Health Record (Synopsis). In: Van Bemmel JH, McCray AT, eds. *Yearbook of Medical Informatics 1994.* Stuttgart: Schattauer Verlag 1994: 173-5.
4. Scherrer JR, Revillard C, Borst F, Berthoud M, Lovis C. Medical office automation integrated into the distributed architecture of a hospital information system. Meth Inform Med 1994; 33: 174-9.
5. Baud R, Rassinoux AM, Scherrer JR. Natural Language Processing and Medical Records. In: Lun K, Degoulet P, Pierre T, Rienhoff O, eds. *MEDINFO 92.* Amsterdam: North Holland Publ Comp, 1992: 1362-7.
6. Rassinoux AM, Michel PA, Juge C, Baud R, Scherrer JR. Natural Language Process-

[13] The Pundit system originally processed US Navy repair reports before it was ported to the domain of air-travel planning information systems (ATIS) where it served as a natural-language interface to a relational database [56].

[14] Von Humboldt (1767–1835), Sapir (1884 to 1939) and Whorf (1897–1941) were language philosophers. After them, many philosophers have debated and polemized on this subject but without really refuting (or validating) the Sapir-Whorf hypothesis.

[15] Cf. Zipf's law that inversely correlates high frequency of use and preciseness of meaning of a word.

[16] Neither the bibliography nor the enumeration of NLP projects claim exhaustivity. Therefore the author welcomes all information about projects or articles that are not mentioned.

ing of Medical Texts within the HELIOS Environment. Comput Meth Prog Bio 1994; 45: 79-96.
7. Sager N, Lyman M, Nhan NT, Tick L. Medical Language Processing: Applications to Patient Data Representation and Automatic Encoding. Meth Inform Med 1995; 34: 140-6.
8. Do Amaral M, Satomura Y. Associating Semantic Grammars with the SNOMED: Processing Medical Language and Representing Clinical Facts into a Language-Independent Frame. In: Greenes R, Peterson H, Protti D, eds. *MEDINFO 95*. Edmonton: Healthcare Computing & Communications, 1995: 18-22.
9. McCray AT, Nelson S. The Representation of Meaning in the UMLS. Meth Inform Med 1995; 34: 193-201.
10. Pietrzyck P. A Medical Text Analysis System for German Syntax Analysis. Meth Inform Med 1991; 30: 275-283.
11. Rector AL, Solomon WD, Nowlan WA et al. A Terminology Server for Medical Language and Medical Information Systems. Meth Inform Meth 1995; 34: 147-57.
12. Sager N, Friedman C, Lyman M. *Medical Language Processing: Computer Management of Narrative Data*. Reading, Massachussets: 1987.
13. Zweigenbaum P, et al. MENELAS, an access system for medical records using natural language. Comput Meth Prog Bio 1994; 45: 117-20 .
14. Bernauer J. Conceptual Graphs as an Operational Model for Descriptive Findings. In: Frisse ME, ed. *SCAMC 92*. New York: McGraw Hill, Inc., 1992: 214-8.
15. Li PY, Evens M, Hier D. Generating Medical Case Reports with the Linguistic String Parser. In: *AAAI-86*. 1986: 1069-73.
16. Wagner J, Solomon W, Michel PA, et al. Multilingual Natural Language Generation as Part of a Medical Terminology Server. In: Greenes R, Peterson H, Protti D, eds. *MEDINFO 95*. Edmonton:Healthcare Computing & Communications, 1995: 100-4.
17. Moore G, Hutchins G, Boitnoit J, Miller R, Polacsek R. Word Root Translation of 45,564 Autopsy Reports into MeSH Titles. In: Stead WW, ed. *SCAMC 87*. Los Angeles: IEEE Computer Society Press, 1987: 128-32.
18. Schröder M. Supporting Speech Processing by Expectations: A conceptual Model of Radiological Reports to Guide the Selection of Word Hypotheses. In: Görz G. ed. *KONVENS 92*. Berlin: Springer-Verlag, 1992: 119-28.
19. Humphreys B, Lindberg D. The Unified Medical Language System Project: a distributed experiment in improving access to biomedical information. In: Lun K, Degoulet P, Pierre T, Rienhoff O, eds. *MEDINFO 92*. Amsterdam: North Holland Publ Comp, 1992: 1496-500.
20. Friedman C, Huff S, Hersh W, Pattison-Gordon E, Cimino J. The Canon Group's Effort: Working Toward a Merged Model. J Am Informatics Assoc, 1995: 2: 4-18.
21. Rossi Mori A. Cooperative Development of a shared Ontology for Medicine. In: *CEN/TC251/WG2*. Geneva: 1994.
22. White W, Barkman B, Bernier-Bonneville L, Cousineau L. A Method for Automatic Coding of Medical Information. Meth Inform Med 1977; 16: 1-10.
23. Zingmond D, Lenert L. Monitoring free-text data using medical language processing. Comput Biomed Res 1993; 26: 467-81.
24. Hersh W, Leone T. The SAPHIRE Server: A New Algorithm and Implementation. In: Reed M, Gardner M, eds. *SCAMC 95*. Philadelphia: Harley & Belfus Inc., 1995: 858-62.
25. Hersh W, Hickham D. Information Retrieval in Medicine: the SAPHIRE experience. In: Greenes R, Peterson H, Protti D, eds. *MEDINFO 95*. Edmonton: Healthcare Computing & Communications, 1995: 1132-7.
26. Evans D, Rothwell D, Monarch I, Lefferts R, Côté R. Towards Representations for Medical Concepts. Med Decis Making 1991; 11(supplement): S102-S8.
27. Evans D, Hersh W, Monarch I, Lefferts R, Handerson S. Automatic Indexing of Abstracts via Natural-language Processing Using a Simple Thesaurus. Med Decis Making 1991; 11(supplement): S108-S15.
28. Evans D. The Language of Medicine and the Meaning of Information. In: Evans D, Patel V, eds. *Advanced Models of Cognition for Medical Training and Practice*. Berlin: Springer-Verlag, 1992.
29. Morris Ch.W. *Writings of the General Theory of Signs*. The Hague: Mouton, 1971.
30. Wolff S. The use of morphosemantic regularities in the medical vocabulary for automatic lexical coding. Meth Inform Med 1984; 23: 195-203.
31. Rassinoux AM, Wagner J, Lovis Ch, Baud R, Rector A, Scherrer JR. Analysis of Medical Texts Based on a Sound Medical Model. In: Reed M, Gardner M, eds. *SCAMC 95*. Philadelphia: Harley & Belfus Inc., 1995: 27-31.
32. Wingert F. Morphologic Analysis of Compound Words. Meth Inform Med 1985; 24: 155-62.
33. Dujols P, Aubas P, Baylon C, Grémy F. Morphosemantic Analysis and Translation of Medical Compound Terms. Meth Inform Med 1991; 30: 30-5.
34. Rossi Mori A, Thornton A, Gangemi A. An Entity-Relationship Model for a European Machine-Dictionary of Medicine. In: *SCAMC 90*. SCAMC Inc., 1990: 185-9.
35. Nijholt A. *Computers and Languages, Theory and Practice (Studies in Computer Science and Artificial Intelligence 4)*. Amsterdam: North Holland Publ Comp, 1988.
36. Grishman R, Kittredge R. *Analyzing Language in Restricted Domains: Sublanguage Description and Processing*. Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1986.
37. Kittredge R, Lehrberger J. *Sublanguage*. Berlin: De Gruyter, 1982.
38. Chi E, Lyman M, Sager N, Friedman C, Macleod CA. Database of Computer-structured Narrative: Methods of computing complex relations. In: *SCAMC 85*. IEEE, 1985: 221-6.
39. Grishman R, Sager N, Raze C, Bookchin B. The Linguistic String Parser. In: *National Computer Conference*. AFIPS Press, 1973: 427-34.
40. Hirschman L, Grishman R, Sager N. From text to structured information - Automatic processing of medical reports. In: *National Computer Conference*. AFIPS Press, 1976: 267-75.
41. Sager N. The String Parser for Scientific Literature. In: Rustin R. ed. *Natural Language Processing*. New York: Algorithmics Press Inc., 1973: 61-87.
42. Sager N. *Natural Language Information Processing: a computer grammar of English and its applications*. Reading MA: Addison-Wesley, 1981.
43. Sager N, Friedman C, Chi E et al. The Analysis and Processing of Clinical Narrative. In: Salamon R, Blum B, Jørgensen M. eds. *MEDINFO 86*. Amsterdam: North Holland Publ Comp, 1986 Elsevier 1986: 1101-5.
44. Lyman M, Sager N, Friedman C, Chi E. Computer-structured Narrative in Ambulatory Care: Its Use in Longitudinal Review of Clinical Data. In: *SCAMC 85*. IEEE, 1985: 82-6.
45. Oliver N. *A sublanguage based medical language processing system for German*. New York: Dept. of Computer Science, New York University, 1992 [unpublished Ph.D. thesis].
46. Grishman R. Implementation of the String Parser of English. In: Rustin R, ed. *Natural Language Processing*. New York: Algorithmics Press Inc., 1973: 89-109.
47. Lyman M, Sager N, Chi E, Tick L, Nhàn NT, Su Y, Borst F, Scherrer JR. Medical Language Processing for Knowledge Representation and Retrieval. In: *SCAMC 89*. SCAMC Inc., 1989: 548-53.
48. Nhàn NT, Sager N, Lyman M, Tick L, Borst F, Su Y. A Medical Language Processor for Two Indo-European Languages. In: *SCAMC 89*. SCAMC Inc., 1989: 554-8.
49. Sager N, Lyman M, Tick L, Borst F, Nhàn NT, Revillard C, Su Y, Scherrer JR. Adapting a Medical Language Processor from English to French. In: *MEDINFO 89*. 1989: 795-9.
50. Sager N, Nhàn NT, Lyman M, Tick L. Computer Analysis of Clinical Narrative: why, how, what, when. In: *BIRA 95*. Gent: 1995: 22-53.
51. Hirschman L, Grishman R, Sager N. Grammatically-Based Automatic Word Class Formation. Inform Process Manag 1975 39-57.
52. Marsh E, Sager N. Analysis and Processing of Compact Text. In: Horecký J. ed. *Proceedings of COLING 82*. Amsterdam North Holland Publ Comp, 1982: 201-6.
53. Sager N. A Two-Stage BNF Specification of Natural Language. J Cybernetics 197 39-50.
54. Sager N. Syntactic formatting of science information. *Fall Joint Computer Conferen* 1972: 791-800.
55. Zweigenbaum P, Bachimont B, Bouaud J Cavazza M, Doré L. HELENE: Compréhension de comptes rendus d'hospitalisation. In: Degoulet P et al, eds. *Informatique et Santé*, vol 1: Informatiq et Gestion des Unités de Soins. 19 257-68.

56. Hirschman L, Palmer M, Dowding J, Dahl D, et al. The Pundit NLP System. In: *AI systems in Government Conference* 1989. IEEE Computer Society Press, 1989.
57. Hirschman L, Puder K. Restriction Grammar: a Prolog Implementation. In: van Caneghem M, Warren D, eds. *Logic Programming and its Applications.* Norwood, New Jersey: Ablex Publishing Corporation, 1985: 244-61.
58. Borst F, Lyman M, Nhan NT, Tick L, Sager N, Scherrer JR. Textinfo: A Tool for Automatic Determination of Patient Clinical Profiles Using Text Analysis. In: *SCAMC 91.* McGraw-Hill, 1991: 63-7.
59. Sager N, Lyman M, Nhàn NT, Tick L, Borst F, Scherrer JR. Clinical knowledge bases from natural language patient documents. In: Lun K, Degoulet P, Pierre T, Rienhoff O, eds. *MEDINFO 92.* Amsterdam: North Holland Publ Comp, 1992: 1374-81.
60. McCray AT, Sponsler J, Brylawski B, Browne A. The Role of Lexical Knowledge in Biomedical Text Understanding. In: *SCAMC 87.* IEEE Computer Society Press, 1987: 103-7.
61. McCray AT. Natural language processing for intelligent information retrieval. In: Nagel J, Smith W, eds. *IEEE 1991.* Orlando, 1991: 1160-1.
62. McCray AT. Extending a Natural Language Parser with UMLS knowledge. In: Clayton P, ed. *SCAMC 91.* McGraw Hill, 1991: 194-8.
63. McCray AT, Srinivasan S, Browne A. Lexical Methods for Managing Variation in Biomedical Terminologies. In: *SCAMC 94.* 1994: 235-9.
64. McCray AT, Razi A. The UMLS Knowledge Source Server. In: Greenes R, Peterson H, Protti D, eds. *MEDINFO 95.* Edmonton: Healthcare Computing & Communications, 1995: 144-7.
65 UMLS *Knowledge Sources 6th Experimental Edition.* National Library of Health, 1995.
66. McCray AT, Srinivasan S. Automated Access to a Large Medical Dictionary: Online Assistance for Research and Application in NLP. Comput Biomed Res 1990; 23: 179-98.
67. Rindflesch T, Aronson A. Semantic Processing in Information Retrieval. In: Safran Ch, ed. *SCAMC 93.* McGraw Hill, 1993: 611-5.
68. McCray AT. Inferencing in Information Retrieval. In: *DARPA 92.* 1992: 218-23.
69. Baud R, Rassinoux AM, Scherrer JR. Natural Language Processing and Semantical Representation of Medical Texts. Meth Inform Med 1992; 31: 117-25.
70. Borst F, Wehrli E, Scherrer JR. MEDIAL, a Natural Language Processing System for Medical Records. In: Roger F, Willems JL, O'Moore R, Barber B, eds. *MIE 84.* Berlin: Springer-Verlag, 1984: 128-33.
71. Borst F, Sager N, Nhàn NT et al. Analyse automatique de comptes rendus d'hospitalisation. In: Degoulet P et al, eds. *Informatique et Santé, vol. 1: Informatique et Gestion des Unités de Soins.* France: Springer Verlag, 1989; Vol. 1: 246-56.
72. Baud R, Rassinoux AM, Scherrer JR. Knowledge Representation of Discharge Summaries. In: Stefanelli M, Hasman A, Fieschi M, Talmon J, eds. *AIME 91.* Springer-Verlag, 1991: 173-82.
73. Baud R, Alpay L, Lovis C. Let's Meet the Users with Natural Language Understanding. In: Barahona P, Christensen JP, eds. *Knowledge and Decisions in Health Telematics.* Amsterdam: IOS Press, 1994: 103-8.
74. Rassinoux AM, Baud R, Scherrer JR. Conceptual graphs model extension for knowledge representation of medical texts. In: Lun K, Degoulet P, Pierre T, Rienhoff O, eds. *MEDINFO 92.* Amsterdam: North Holland Publ Comp, 1992: 1368-74.
75. Morel-Guillemaz [Rassinoux] AM, Baud R, Scherrer JR. Proximity Processing of Medical Text. In: O'Moore R, Bengtsson S, Bryant J, Bryden J, eds. *MIE 90.* Springer-Verlag, 1990: 625-30.
76. Rassinoux AM, Juge C, Michel PA, eds. Analysis of Medical Jargon: the RECIT system. In: Barabhona P, Stefanelli M, Wyatt J, eds. *AIME 95.* Springer-Verlag, 1995: 42-52.
77. Wagner J, Baud R, Scherrer JR. Using the Conceptual Graph Operations for Natural Language Generation in Medicine. In: *ICCS95.* 1995.
78. Rassinoux AM. *Extraction et Représentation de la Connaissance tirée des Textes Médicaux.* Département d'Informatique, Université de Genève, 1994 [unpublished Ph.D. thesis].
79. Sowa J. *Conceptual Structures: Information Processing in Mind and Machine.* Reading, MA: Addison-Wesley, 1984.
80. Sowa J. *Principles of Semantic Networks.* San Mateo: Morgan Kaufmann, 1991.
81. Sowa J. Conceptual Analysis for Knowledge Base Design. Meth Inform Med 1995; 1/2: 165-71.
82. Pietrzyck P. Free text analysis. Int Journal Biomed Comput 1995; 39: 139-44.
83. Baud R, Lovis C, Alpay L. Modelling for Natural Language Understanding. In: Safran Ch, ed. *SCAMC 93.* New York: McGraw Hill, 1993: 289-93.
84. Rector A. Coordinating Taxonomies: Key to Re-usable Concept Representations. In: *AIME 95.* Springer Verlag, 1995: 17-28.
85. Wagner J, Baud R, Scherrer JR. Generating Noun Phrases from a Medical Knowledge Representation. In: Barahona P, Veloso M, Bryant T, eds. *MIE 94.* Lisbon: 1994: 218-23.
86. Lovis C, Michel PA, Baud R, Scherrer JR. Word Segmentation Processing: A way to Exponentially Extend Medical Dictionaries. In: Greenes R, Peterson H, Protti D, eds. *MEDINFO 95.* Edmonton: Healthcare Computing & Communications, 1995: 28-32.
87. Friedman C, Cimino J, Johnson S. A Conceptual Model for Clinical Radiology Reports. In: *SCAMC 93.* New York: McGraw Hill, 1993: 829-33.
88. Friedman C, Johnson S, Forman B, Starren J. Architectural Requirements for a Multipurpose Natural Language Processor in the Clinical Environment. In: Reed M, Gardner M, eds. *SCAMC 95.* Philadelphia: Harley & Belfus Inc., 1995: 347-51.
89. Friedman C, Alderson P, Austin J, Cimino J, Johnson S. A General Natural-language Text Processor for Clinical Radiology. J Am Med Informatics Assoc 1994; 1: 161-74.
90. Cimino J, Clayton P, Hripcsak G, Johnson S. Knowledge based approaches to the maintenance of a large controlled medical terminology. J Am Med Informatics Assoc 1994; 1: 35-40.
91. Hripcsak G, Friedman C, Alderson P et al. Unlocking Clinical Data from Narrative Reports: A Study of Natural Language Processing. Ann Intern Med 1995; 9: 681-8.
92. Johnson S, Aguirre A, Peng P, Cimino J. Interpreting Natural Language Queries using the UMLS. In: *SCAMC 93.* New York: McGraw Hill, 1993: 294-8.
93. Haug P, Koehler S, Lau ML, Wang P, Rocha R, Huff S. Experience with a Mixed Semantic/Syntactic Parser. In: Reed M, Gardner M, eds. *SCAMC 95.* Philadelphia: Harley & Belfus Inc., 1995: 284-8.
94. Haug P, Ranum D, Frederick P. Computerized Extraction of Coded Findings from Free-Text Radiologic Reports. Radiology 1990; 174: 543-8.
95. Ranum D. Knowledge-based understanding of radiology text. Comput Meth Prog Bio 1989; 30: 209-15.
96. Zweigenbaum P, Cavazza M. Deep sentence understanding in a restricted domain. In: *COLING 90.* Helsinki: 1990: 82-4.
97. Zweigenbaum P, Cavazza M, Doré L, Bouaud J, Sedlock D. Natural Language Processing of Patient Discharge Summaries (NLPAD) - Extraction Prototype. In: Nothoven van Goor J, Christensen JP, eds. *AIM 92.* Amsterdam: IOS Press, 1992: 215-22.
98. Zweigenbaum P, Bachimont B, Bouaud J, Charlet J, Boisvieux JF. Issues in the Structuration and Acquisition of an Ontology for Medical Language Understanding. Meth Inform Med 1995; 1/2: 15-24.
99. Zweigenbaum P et al. *MENELAS,* The Final Report. Menelas Deliverable #17, Paris: 1995.
100. Zweigenbaum P et al. MENELAS, Coding and Information Retrieval from Natural Language Patient Discharge Summaries. In: Laires M, Ladeira M, Christensen J, eds. *Health in the New Communication Age.* Amsterdam: IOS Press, 1995: 82-9.
101. Zweigenbaum P, Cavazza M. Extracting Implicit Information from Free Text Technical Reports. *Information Processing and Management 1992.*
102. Volot F, Zweigenbaum P, Bachimont BS et al. Structuration and Acquisition of Medical Knowledge (using UMLS in the Conceptual Graph Formalism). In: *SCAMC 93.* New York: McGraw Hill, 1993: 710-4.
103. Bouaud J, Zweigenbaum P. A reconstruction of conceptual graphs on top of a production system. In: *7th Annual Workshop on Conceptual Graphs.* Las Cruces: 1992.
104. Bouaud J. TREE: the Heuristic Join Strategy of a RETE-*like* Matcher. In: *IJCAI93.* Chambéry: IJCAI, 1993: 496-502.
105. Bouaud J. Un système de production à base de graphes conceptuels: application dans un système de compréhension de textes. In: *Actes de la Journée PRC-GDR IA Graphes Conceptuels 94.* Marseille: LIRMM, 1994.

106. Zweigenbaum P et al. *MENELAS Linguistic and Conceptual Knowledge Version 1.* Menelas Deliverable #9. Paris: 1993.
107. Bérard-Dugourd A, Fargues J, Landau MC, Rogala JP. Un système d'analyse de texte et question/réponse basé sur les graphes conceptuels. In: Degoulet P et al., eds. *Informatique et Santé, vol 1: Informatique et Gestion des Unités de Soins.* France: Springer Verlag, 1989; Vol.1: 223-33.
108. Fargues J, Landau MC, Dugourd A, Catach L. Conceptual Graphs for semantics and knowledge Processing. IBM J Res Dev 1986; 30: 70-9.
109. Grover C, Caroll J, Briscoe T. *The Alvey Natural Language Tools Grammar.* Cambridge: University of Cambridge, 1992: (4th release).
110. Mikheev A, Moens M. KADS methodology for Knowledge Based Language Processing Systems. In: *8th Banff Knowledge Acquisition for Knowledge-Based Systems Workshop.* Canada: 1994: 5.1-5.17.
111. Mikheev A, Moens M. Acquiring and Representing Background Knowledge for a Natural Language Processing System. In: *AAAI 94 Fall Symposium: KR for NLP in Implemented Systems.*
112. Whittemore G. The MENELAS English Natural Language Understander: Natural Language Understanding in the medical domain. In: *The First World Congress on Computational Medicine, Public Health and Biotechnology.* Austin, Texas: 1994.
113. Kaplan R, Bresnan J. Lexical-Functional Grammar: a formal system for grammatical representation. In: Bresnan J, ed. *The Mental Representation of Grammatical Relations.* Cambridge, Mass.: MIT Press, 1982: 173-81.
114. Delamarre D, Burgun A, Seka LP, Le Beux P. Automated coding system of patient discharge summaries using conceptual graphs. Meth Inform Med 1995; 34: 345-51.
115. Nangle B, Keane M. Effective retrieval in Hospital Information Systems: The use of context in answering queries to Patient Discharge Summaries. Artif Intell Med 1994; 6: 207-27.
116. Bouaud J, Bachimont B, Charlet J, Zweigenbaum P. Acquisition and structuring of an ontology within conceptual graphs. In: *ICCS94 Workshop on Knowledge Acquisition using Conceptual Graph Theory,* 1994: 1-25.
117. Bouaud J, Bachimont B, Charlet J, Zweigenbaum P. Methodological principles for structuring an "ontology". *IJCAI95 Workshop on "Basic Ontological Issues in Knowledge Sharing".*
118. Zweigenbaum P, Bachimont B, Bouaud J, Charlet J, Boisvieux JF. A Multi-lingual Architecture for Building a Normalised Conceptual Representation from Medical Language. In: Reed M, Gardner M, eds. *SCAMC 95.* Philadelphia: Harley & Belfus Inc., 1995: 357-61.
119. Cavazza M, Doré L, Zweigenbaum P. Model-based Natural Understanding in Medicine. In: *MEDINFO 92.* Amsterdam: North Holland Publ Comp, 1992: 1356-61.
120. Doré L, Cavazza M, Zweigenbaum P, Boisvieux JF. Analyse Pragmatique pour la compréhension de comptes rendus d'hospitalisation. In: Degoulet P et al., eds. *Informatique et Santé, vol. 5: Nouvelles Méthodes de Traitement de l'Information Médicale.* France: Springer Verlag, 1992; Vol. 5: 139-52.
121. Bouaud J, Bachimont B, Zweigenbaum P. Processing Metonymy: a Domain-Model Heuristic Graph Traversal Approach. In: *COLING 96* (in press).
122. Ledoray V, Pellegrin L, Guisano B, Roux M. Système de compréhension de comptes rendus médico-techniques: architecture, connaissances nécessaires et résultats. In: Degoulet P et al., eds. *Informatique et Santé.* France: Springer Verlag, 1992: 111-25.
123. Pellegrin L, Bastien C, Roux M. Representation of Medical Concepts of the Thyroid Gland by Physicians in Anatomy and Pathology. Meth Inform Med 1994; 33: 382-9.
124. Roux M, Giusiano B. A propos d'une application de l'intelligence artificielle à la médecine: l'analyse automatique des comptes rendus médico-techniques. Pathologie Biologie 1990; 38: 626-33.
125. Ledoray V, Guisano B, Roux M. A system for understanding medical reports: architecture and knowledge required. In: *MEDINFO 92.* Amsterdam: North Holland Publ Comp, 1992: 1389-94.
126. Smart J, Roux M. A model for representing medical knowledge: application to the analysis of medical reports written in natural language. In: *AIME 95.* Springer Verlag, 1995.
127. Berrut C. *Une méthode d'indexation fondée sur l'analyse sémantique de documents specialisés. Le prototype RIME et son application à un corpus médical.* Grenoble: Laboratoire de Genie Informatique – IMAG – l'Université Joseph Fourier, Grenoble I, 1988: [unpublished Ph.D. thesis].
128. Berrut C, Cinquin P. Natural language understanding of medical reports. In: Scherrer JR, Coté R, Mandil S, eds. *Computerized Natural Medical Language Processing for Knowledge Engineering.* Amsterdam: North Holland Publ Comp 1989: 129-37.
129. Chiamarella Y, Jianyun N. A Retrieval Model based on an Extended Modal Logic and its Application to the RIME Experimental Approach. J ACM 1990; 9: 25-43.
130. Berrut C. Indexing Medical Reports: The RIME Approach, Inform Process Manag 1990; 26: 93-109.
131. Pietrzyck P. Survey of the Goettingen Medical Text Analysis System. In: *Medical Informatics in Europe 88.* Springer-Verlag, 1988: 128-32.
132. Pietrzyck P. Literal meaning of sentences from the medical free text. In: *Medical Informatics in Europe 93.* London: Freund Publishing House, 1993: 64-7.
133. Schröder M. Knowledge Based Analysis of Radiology Reports Using Conceptual Graphs. In: Pfeiffer HD, ed. *Seventh Annual Workshop on Conceptual Graphs.* 1992: 213-22.
134. Schröder M. Knowledge Based Processing of Medical Language: A Language Engineering Approach. In: Ohlbach HJ, ed. *Sixteenth German Workshop on AI (GWAI 92).* Berlin: Springer-Verlag, 1992: 221-34.
135. Satoruma Y, Kaihara S, Oikawa A, Imanishi K. A Medical Computer Dictionary and an application to the discharge summary. In: Scherrer JR, Coté R, Mandil S, eds. *Computerized Natural Medical Language Processing for Knowledge Engineering.* Amsterdam: North Holland Publ Comp 1989: 263-8.
136. Satoruma Y, Do Amaral M. Automated diagnostic indexing by Natural Language Processing. Med Inform 1992; 17: 149-63.
137. Do Amaral M, Satomura Y. Structuring medical information into a language-independent database. Med Inform 1994; 19: 269-82.
138. Pacak M, Pratt A. Identification and Transformation of Terminal Morphemes in Medical English (part 1). Meth Inform Med 1969; 8: 84-90.
139. Pacak M, Pratt A. Identification and Transformation of Terminal Morphemes in Medical English (part 2). Meth Inform Med 1978; 17: 95-100.
140. Norton L, Pacak M. Morphosemantic Analysis of Compound Word Forms Denoting Surgical Procedures. Meth Inform Med 1983; 22: 29-36.
141. Pacak M, Norton L, Dunham G. Morphosemantic of -ITIS Forms in Medical Language. Meth Inform Med 1980; 19: 99-105.
142. Wingert F. In: Reichertz P. Goos G, eds. *Informatics and Medicine, an advanced course.* Springer Verlag, 1977: 579-646.
143. Wingert F. Medical Linguistics: a Review. In: *MEDINFO 80.* 1980: 1321-31.
144. Wingert F. Morphological Analysis of Medical Compound Word Forms. In: Schneider W, Sågvall Hein AL, eds. *Computational Linguistics in Medicine.* 1977: 79-89.
145. Wingert F, Rothwell D, Côté R. Automated Indexing into SNOMED and ICD. In: Scherrer JR, Coté R, Mandil S, eds. *Computerized Natural Medical Language Processing for Knowledge Engineering.* Amsterdam: North Holland Publ Comp 1989: 5.1-5.38.
146. Brigl B, Mieth M, Haux R, Glück E. The LBI-method for automated indexing of diagnoses by using SNOMED. Part 1: Design and Implementation. Int J Biomed Comput 1994; 37: 237-47.
147. Brigl B, Mieth M, Haux R, Glück E. The LBI-method for automated indexing of diagnoses by using SNOMED. Part 2: Evaluation. Int J Biomed Comput 1995; 38: 101-8.
148. Gabrieli E, Speth D. Computer Processing of Discharge Summaries. In: *SCAMC 87.* /EEE Computer Society Press, 1987: 137-40.
149. Gabrieli E, Speth D. Automated analysis of natural language medical text. *Proceedings AAMS* 1988; 17: 66-75.
150. Gabrieli E. Computer-based Medical Terminology and Knowledge Representation. In: De Moor G et al., eds. *Progress in Standardisation in Health Care Informatics.* Amsterdam: 1993: 81-5.
151. Dujols P. How to extract relevant information from well kept discharge summaries. In: Scherrer JR, Coté R, Mandil S, eds. *Computerized Natural Medical Language Processing for Knowledge Engineering.*

Amsterdam: North Holland Publ Comp 1989: 47-56.
152. Dujols P, Aubas P, Romero M. Saisie et communication de dossiers médicaux en langage clair. In: Degoulet P et al., eds. *Informatique et Santé, vol. 1: Informatique et Gestion des Unités de Soins.* France: Springer Verlag, 1989; Vol.1: 234-45.
153. Dujols P, Baylon Chr, Chein M. Projet LIME: Linguistique et Langage Médical. In: Degoulet P et al., eds. *Informatique et Santé, vol 5: Nouvelles Méthodes de Traitement de l'Information Médicale.* France: Springer Verlag, 1992; Vol. 5: 126-38.
154. Dujols P. Langage, continuité ou retour. In: Degoulet P et al., eds. *Informatique et Santé, vol. 7: Informatisation de l'Unité de Soins du Futur.* France: Springer Verlag, 1994; Vol.7: 89-110.
155. Spyns P, Willems JL. Dutch Medical Language Processing: discussion of a prototype. In: *MEDINFO 95.* Amsterdam: North Holland Publ Com, 1995: 37-40.
156. Spyns P, Zweigenbaum P, Willems JL. Representation and Extraction of Information from Patient Discharge Summaries by means of Natural Langage Processing. In: Ten Hoopen A, Hofdijk W, Beckers W, eds. *MIC 92.* Rotterdam: Publicon Publishing, 1992: 309-16 [in Dutch].
157. Spyns P, Dehaspe L, Willems JL. *The Menelas Syntactic Analysis Component for Dutch.* Menelas Delivrable #6. Leuven: 1993.
158. Spyns P. *A Tagger/Lemmatizer for Dutch medical language.* Technical report 94-002, K.U. Leuven, Dept. of Medical Informatics. Leuven: 1994.
159. Spyns P. A robust Category Guesser for Dutch Medical Language. In: *4th Conference on Applied Natural Language Processing.* ACL, Morgan Kaufmann, 1994: 150-5.
160. Spyns P, De Wachter L. Morphological Analysis of Dutch Medical Compounds and Derivations. ITL Review of Applied Linguistics, 1995; 109-10: 19-35.
161. Spyns P, De Moor G. A Dutch Medical Language Processor. J Biomed Eng 1996 [in press].
162. Spyns P, Adriaens G. Applying and Improving the Restriction Grammar Approach for Dutch Patient Discharge Summaries. In: *14th International Conference on Computational Linguistics.* Nantes: ACL, 1992: 1164-8.
163. Hirschman L, Dowding J. Restriction Grammar: a logic grammar. In: Saint-Dizier P, Szpakowicz S, eds. *Logic and Logic Grammars for Language Processing.* Ellis Horwood, 1990: 141-67.
164. Spyns P. A *Prototype for semi-automatic encoding.* Leuven, 1991 [unpublished Ph. D. thesis in Dutch].
165. Mommaerts J, Ceusters W, Deville G. Zijn taggers voor algemene taal bruikbaar voor medische subtalen? In: Beckers WPA, ten Hoopen AJ, eds. *MIC 94.* Rotterdam: VMBI/TMI, 1994: 283-90 [in Dutch].
166. Maks I, Martin W. MULTI-TALE: Linking Medical Concepts by means of Frames. In: *COLING96* (in press).
167. Ceusters W. *The generation of multi-lingual specialised lexicons by using augmented lemmatizer-taggers.* Delivrable report Multi-TALE #1. Gent: 1994.
168. Paulussen H, Martin W. Dilemma-2: a Lemmatizer-Tagger for medical abstracts. In: *3th Conference on Applied Natural Language Processing.* Trento: ACL, 1992: 141-6.
169. De Moor G. Standardisation in Health Care Informatics and Telematics in Europe: CEN TC 251 Activities. In: De Moor G et al., eds. *Progress in Standardisation in Health Care Informatics.* Amsterdam: 1993: 1-13.
170. Ceusters W, Deville G, De Moor G. Automated extraction of neurosurgical procedure expressions from full text reports: the Multi-TALE experience. In: *MIE 96.* 1996 (in press).
171. Ceusters W, Deville G, Streiter O, Herbigniaux E, Devlies J. A Computational Linguistic Approach to Semantic Modelling in Medicine. In: Beckers WPA, ten Hoopen AJ, eds. *MIC 94.* Rotterdam: VMBI/TMI, 1994: 311-9.
172. Ceusters W, Devlies J. The Anthem Representation Formalism for the Alphabetic Index of ICD. In: Greenes R, Peterson H, Protti D, eds. *MEDINFO 95.* Edmonton: Healthcare Computing & Communications, 1995: 113-6.
173. Mousel P, Thienpont G. Integrating Natural Language Technology in Medical Information Systems: The Anthem Approach. In: *BIRA 95.* Gent: 1995: 82-90.
174. Deville G, Herbigniaux E. Natural Language Modeling in a Machine Translation Prototype for Healthcare Applications: a Sublanguage Approach. In: *Proceedings TMI 95.* Leuven: 1995: 142-57.
175. Ceusters W, Deville G. A Multi-Dimensional View on Natural Language Modelling in Medicine: Identifying Key-features for successful Applications. Meth Inform Med 1995; 1/2 (supplementary paper).
176. Deville G. When linguistics meets medical knowledge engineering: an interdisciplinary approach to Machine Translation in Healthcare. In: *BIRA 95.* Gent: 1995: 54-81.
177. Ceusters W, Deville G, Devlies J, Gérardy C, Mousel P, Streiter O, Penson D. The ANTHEM Project: when machine translation meets automatic encoding. In: *Language Engineering Convention.* Paris, France: 1994: 25-32.
178. Bouaud J, Séroussi B. An experiment towards a Document-Centered Hypertextual Computerised Patient Record. In: *MIE 96.* Amsterdam: IOS Press, 1996 (in press).
179. Spyns P, Ceusters W. Document Management in Health Care: Presentation of the DOME project. In: van der Lei J, Beckers WPA, eds. *AMICE 95.* Rotterdam: VMBI/TMI, 1995: 359-68.
180. Baud R, Lovis C, Rassinoux AM, Scherrer JR. Tendances en traitement du langage naturel. In: Duserre L, Goldberg M, Salamon R, eds. *Informatique et Santé, vol 8: Information Médicale: Aspects Déontologiques, Juridiques et de Santé Publique.* France: Springer Verlag, 1996; Vol. 8: 111-9.
181. Baud R, Rassinoux AM, Wagner J, Lovis C et al. Representing Clinical Narratives Using Conceptual Graphs. Meth Inform Med 1995; 1/2: 176-86.
182. Friedman C, Johnson S. Medical Text Processing: Past achievements, future directions. In: Ball M, Collen M, eds. *Aspects of the Computer-based Patient Record.* Springer Verlag, 1992: 212-28.
183. Ceusters W, Deville G, Buekens F. The Chimera of Purpose and Language Independent Concept System in Health Care. In: Barahona P, Veloso M, Bryant T, eds. *MIE 94.* Lisbon: 1994: 208-12.
184. Baud R, Lovis C, Rassinoux AM et al. Towards a Medical Linguistic Knowledge Base. In: Greenes R, Peterson H, Protti D, eds. *MEDINFO 95.* Edmonton: Healthcare Computing & Communications, 1995: 13-7.
185. Barahona P, Veloso M, Bryant T, eds. *MIE 94.* Lisbon: 1994.
186. Barahona P, Stefanelli M, Wyatt J, eds. *AIME 95.* Springer Verlag, 1995.
187. Degoulet P et al, eds. *Informatique et Santé, vol. 1: Informatique et Gestion des Unités de Soins.* France: Springer Verlag, 1989.
188. Degoulet P et al, eds. *Informatique et Santé, vol. 5: Nouvelles Méthodes de Traitement de l'Information Médicale.* France: Springer Verlag, 1992.
189. Degoulet P et al, eds. *Informatique et Santé, vol. 7: Informatisation de l'Unité de Soins du Futur.* France: Springer Verlag, 1994.
190. De Moor G et al, eds. *Progress in Standardisation in Health Care Informatics.* Amsterdam: IOS Press, 1993.
191. Greenes R, Peterson H, Protti D, eds. *MEDINFO 95.* Alberta: HC & CC, 1995.
192. Lun K, Degoulet P, Pierre T, Rienhoff O, eds. *MEDINFO 92.* Amsterdam: North Holland Publ Comp, 1992.
193. Reed M, Gardner M, eds. *SCAMC 95.* Philadelphia: Harley & Belfus Inc., 1995.
194. Safran Ch, ed. *SCAMC 93.* New York: McGraw-Hill Inc., 1993.

Address of the authors:
Peter Spyns,
Afdeling Medische Informatica,
Universitair Ziekenhuis Gent,
De Pintelaan 185 (5K3)
B-9000 Gent, Belgium
e-mail: Peter.Spyns@rug.ac.be