Natural Language Modeling in a Machine Translation Prototype for Healthcare Applications: a Sublanguage Approach

G. Deville & E. Herbigniaux

Ecole de Langues Vivantes - Facultés Universitaires de Namur
61 rue de Bruxelles - B-5000 Namur - Belgium

Abstract

This paper discusses methodological issues related to natural language modeling in the framework of the LRE project ANTHEM 1. The objective of ANTHEM is to develop a portable prototype of a multilingual natural language interface that allows users of Healthcare Information Systems to enter diagnostic expressions using their own natural language, and to have this input translated in whatever formal, structured or natural language. We describe the approach adopted in the modeling of the ANTHEM application language in the prospect of the ANTHEM lingware development. This approach is motivated by the notion of sublanguage.

1 INTRODUCTION

The computational intractability of the natural language phenomenon in its most general usage has become very obvious since computer specialists started to use their instruments for non-mathematical applications. As a result, even the computer of the latest generation, with its colossal storage and computing capacities, is so far unable to treat language in its entirety. The main reason for this unanswered challenge is that a computer needs all kinds of explicit information, even when this information is too much of a truism to the human mind (e.g. polysemies, implication of the communication content, intentions, presuppositions of the speakers and encyclopedic or world knowledge). If this linguistic (including pragmatic) knowledge could be kept within such boundaries, so as to allow the computer to store it explicitly, automatic natural language understanding would become feasible within this limited world. When referring to such a world, one uses a sublanguage that shows prototypical properties in mainly lexical, syntactic, semantic and pragmatic aspects.

The present paper substantiates the claim that the notion of sublanguage can be successfully applied to the domain of written medical diagnostic expressions, and that such a methodological approach enables us to define this application language upon computationally tractable criteria for the development of a machine translation (MT) prototype for healthcare applications (ANTHEM project).

In section 2, a tentative definition of a sublanguage is first proposed in the light of a short survey of the relevant literature. The modeling of the ANTHEM application sublanguage in the prospect of the ANTHEM lingware development (multilingual electronic lexicons and grammars) necessitated an extensive study of large representative corpora of medical diagnostic expressions. Section 3 discusses the medical corpus collection, formatting, sampling and analysis. In section 4 the sublanguage of medical diagnostic expressions will then be explored at various linguistic levels, and illustrated by typical examples taken from the above-mentioned language data. Section 5 briefly expounds the semantic representation model developed in the context of the ANTHEM project.

[1] The ANTHEM project: “Advanced Natural Language Interface for Multilingual Text Generation in Healthcare” (LRE 62-007) is co-financed by the European Union within the “Linguistic Research and Engineering” program. The ANTHEM consortium is coordinated by W. Ceusters of RAMIT vzw (Ghent University Hospital) and further consists of the Institute of Modern Languages of the University of Namur (G. Deville), the IAI of the University of Saarbrücken (O. Streiter), the CRP-CU of Luxembourg (P. Mousel), the University of Liège (C. Gérardy), Datasoft Management nv - Oostende (J. Devlies) and the Military Hospital in Brussels (D. Penson).

2. TENTATIVE DEFINITION OF A SUBLANGUAGE

In the literature (e.g. [KITTREDGE & LEHRBERGER 82]), various terms refer to the notion of sublanguage, reflecting not only the historical context in which this concept appeared and evolved, but also the perspective in which the sublanguage phenomenon was described (some authors insisting on the structural aspects of the linguistic productions in a limited domain, others privileging the functional dimension of these productions).

Harris was the first linguist who introduced the concept of sublanguage in the precursory terms of the transformational generative paradigm, where in his analysis of scientific writings, he claims that certain proper subsets of the sentences of a sublanguage may be closed under some or all the operations defined in the language, and thus constitute a subsystem of it [HARRIS 68]. A similar notion appeared in applied linguistics, where the international dimension of scientific and technological information exchange resulted into a growing need for conceiving foreign language course materials targeted at specialists in a given domain, hence the concept of 'languages for specific purposes' or 'LSP' [SAGER e.a. 80]. In the field of terminology, the notion of sublexica has been recently examined by [MARTIN & TEN PAS 91], in the light of an E.C. study by [McNAUGHT e.a. 91] on the feasibility of standards for terminological description of lexical items in an automatisation perspective.

On the other hand, a large proportion of European and overseas research projects in language engineering have extensively studied specific application languages in the sublanguage perspective, namely in the fields of automatic parsing of medical texts for information formatting purpose [SAGER 82], automatic tagging of medical abstracts [PAULUSSEN & MARTIN 92], and automatic translation of weather forecasts [CHEVALIER 78]. [ADRIAENS 93] is currently developing a simplified English grammar and style checker for the automatic correction of texts in the field of telecommunications.

In its oral and dialogic form, the operative dimension of a sublanguage has been convincingly demonstrated by [FALZON 86] in a study of the language used in air-traffic control, and [DEVILLE 89] appeals to the notion of sublanguage as a methodological principle to modelize the language of requests for administrative information in a task-oriented voice-driven man-machine dialogue system.

Although the systematic study of sublanguage as a working hypothesis is lacking a long tradition, we will define a sublanguage as *'a set of utterances referring to a limited and well-defined application domain and used for a specific function'*.

Such utterances are generated by a specific grammar with a specific vocabulary. Our definition views a sublanguage in terms of limitations of the application domain referred to, rather than mere restrictions of the language forms. This view is expressed in our description of a sublanguage as the merging of a subset of the general language with a set of specific elements of its own. In this perspective, the sublanguage elements are described as a set of elements placed in a preferential order, depending on the communicative situation and the domain of application, and the restricted character of a sublanguage is not an exclusion but a preferential hierarchy. We come back to this point in section 4.

3. A CORPUS-BASED ANALYSIS

Aiming at a representative sample of the ANTHEM application language, three corpora of medical diagnostic expressions have been collated in different clinical environments. One source of language data in the ANTHEM project is the corpus of diagnostic expressions from the Belgian Army (referred to as the ABL corpus). The ABL corpus contains a total of 227.900 diagnostic expressions in French and Dutch, written by military doctors from 1970 to 1993, during consultations with professional servicemen and civilians in national service. In parallel, a corpus of 12.671 Dutch and French diagnostic expressions (referred to as the MEDIDOC corpus) covering the same period as the ABL corpus, has been collected from various civilian doctors in the framework of a private consulting practice, re-using the relevant parts of the electronic medical records produced by an existing medical data managing software (MEDIDOC) that has been designed by Datasoft Management nv. Finally, a corpus of 500 German diagnostic expressions (referred to as the HOMBURG corpus) has been collected from medical reports written at the Universitatsklinik der Universitat des Saarlandes, Homburg. All the expressions of the three corpora have been tagged with information on their origin, year and language. Furthermore, each expression is identified by a sequentially allocated number, as illustrated in the following examples[2]:

ABL_85_F_12384	RUPTURE MENISQUE INTERNE GENOU DROIT.
ABL_88_N_11224	RECHTER ACHILLESPEES TENDINITIS.
MDD_93_F_00172	CONTUSION GENOU DROIT.
MDD_93_N_01176	WONDE LINKER SCHEENBEEN.
HOM_XX_D_00488	Artérielle Hypertonie.

As the amount of 241.071 expressions is not manageable for linguistic analysis and modeling, a first sample of 2.343 expressions has been randomly selected on the basis of the ABL and MEDIDOC corpora. As the development of German lingwares was not foreseen in this stage of the project, the HOMBURG corpus has not been sampled so far. The linguistic and medical validity of these expressions was then checked respectively by linguists and medical practitioners, and a final sample of 1.362 valid expressions has been set up (i) for the elaboration of the lingwares as well as (ii) for the testing of the ANTHEM prototype. A detailed account of the methodological principles for the elaboration of the corpora of diagnostic expressions in the ANTHEM project is given in [DEVILLE & HERBIGNIAUX 94].

4. THE SUBLANGUAGE OF MEDICAL DIAGNOSTIC EXPRESSIONS

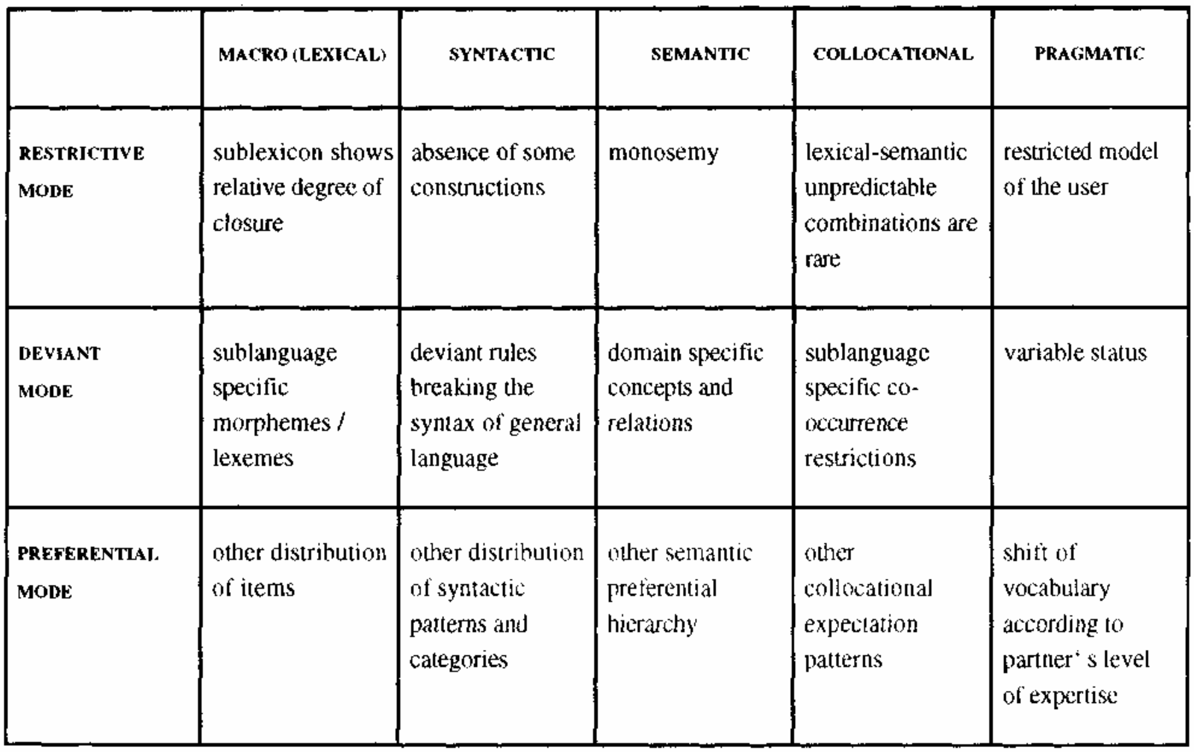
A sublanguage differs from the general language in its pragmatic dimensions : it is used in a limited communicative context and refers to a specific domain of knowledge. In order to further define the status of a sublanguage vs. the general language, we propose to structurally define its properties at each linguistic level (lexical, syntactic, semantic and pragmatic) according to three complementary modes :

- the restrictive mode : by excluding certain features of the general language, a sublanguage can be described as a restricted form of language;
- the deviant mode : a sublanguage can show specific features which are not found in general language and therefore can be considered as a deviant form of language;

[2] As most of the HOMBURG corpus expressions revealed impossible to date with certainty, the year tag has been systematically filled with the string “XX”.

- the preferential mode : this approach of sublanguage phenomena is complementary to the restrictive and deviant modes, and is expressed in terms of statistical preference : specific features of the general language have a low probability of occurrence in a given sublanguage, though they cannot be totally excluded from that sublanguage. Conversely, some words, syntactic structures and categories occur more frequently in a given sublanguage than in general language.

We will now apply this descriptive matrix to the language of medical diagnostic expressions, and discuss its prototypical sublinguistic characteristics in the light of examples from the literature. Our discussion is based on [DEVILLE 89] & [MARTIN & TEN PAS 91]. The lexical, syntactic, semantic and pragmatic characteristics of a sublanguage are summarized in Table 1 below.



	MACRO (LEXICAL)	SYNTACTIC	SEMANTIC	COLLOCATIONAL	PRAGMATIC
RESTRICTIVE MODE	sublexicon shows relative degree of closure	absence of some constructions	monosemy	lexical-semantic unpredictable combinations are rare	restricted model of the user
DEVIANT MODE	sublanguage specific morphemes / lexemes	deviant rules breaking the syntax of general language	domain specific concepts and relations	sublanguage specific co-occurrence restrictions	variable status
PREFERENTIAL MODE	other distribution of items	other distribution of syntactic patterns and categories	other semantic preferential hierarchy	other collocational expectation patterns	shift of vocabulary according to partner' s level of expertise

Table 1. Lexical, syntactic, semantic, collocational and pragmatic sublanguage definition.

4.1. Lexical aspects

A sublanguage is characterized by the relative degree of closure of its lexicon, which implies that, in contrast to general language, the vocabulary of a sublanguage is virtually finite. Lexical closure is thus to be defined as a relative concept rather than as a binary property, as defined by [MOSKOVICH 82] :

> "if we have a finite set of texts belonging to a certain sublanguage, we can determine its vocabulary and inventory of grammatical constructions. Let us add to this set new texts of the same sublanguage. If the vocabulary and the inventory of grammatical constructions remain unchanged, the sublanguage is called a closed one - in the opposite case, it is called an open one."

[KOCOUREK 82] comes to an apparently opposite conclusion when he observes that specialized language lexicons are quite extensive as opposed to usual and symbolic languages. [MARTIN & TEN PAS 91] consider that (i) in the latter case [KOCOUREK 82] focuses on the in-depth extension of specialized lexicons that expresses increasingly specialized knowledge as it is the case in organic chemistry (for example [MOSKOVICH 82]), and that (ii) in his approach, [DEVILLE 89] focuses on the in-breadth limitation of the sublanguage lexicon due to the restricted pragmatic characteristics of that sublanguage.

The extension of the lexicon of medical diagnostic expressions is characterized more by its in-depth than in-breadth dimension. The pragmatic limitations (i.e. in-breadth dimension) of the ANTHEM sublanguage come down to the factual observation of physical / physiological / psychological changes at a patient in a given time lapse. The domain of knowledge that is to be covered by the sublanguage lexicon (i.e. its in-depth dimension) is reflected by the size of the lexicon of the SNOMED nomenclature (Systematized Nomenclature of Human and Veterinary Medicine) used in ANTHEM as one of the elements for designing the semantic representation of medical diagnostic expressions [COTE e.a. 93]. The entire SNOMED nomenclature amounts to 132,641 terms or expressions. Table 2 gives the size of the modules (in terms / expressions) covering each of the twelve SNOMED types.

Broken down into its individual terms, the SNOMED nomenclature comes down to a lexicon of 57.000 words, or so. We are currently examining the degree of coverage of the SNOMED lexicon on the restricted sample of the ABL corpus (French and Dutch parts). The hypothesis we want to test is whether the lexicon of the medical diagnostic sublanguage shows a high degree of closure in its in-depth dimension, and whether the reduced SNOMED lexicon at atomic conceptual level is a relevant yardstick to test the significance of this property. We discuss below the lexical properties of the sublanguage under study in a preferential mode.

Interestingly, [MARTIN & TEN PAS 91] also consider the closure property of sublexicons from a morphological viewpoint : the generation/formation of new terms in the specialized language is not only semantically limited, but also meets systematic and controlled rules, compared to word-formation in general language. This morphological phenomenon systematically occurs in the sublanguage of medical diagnostic expressions, as illustrated by the following suffix examples :

- the suffix *-algie* refers to the concept of pain :

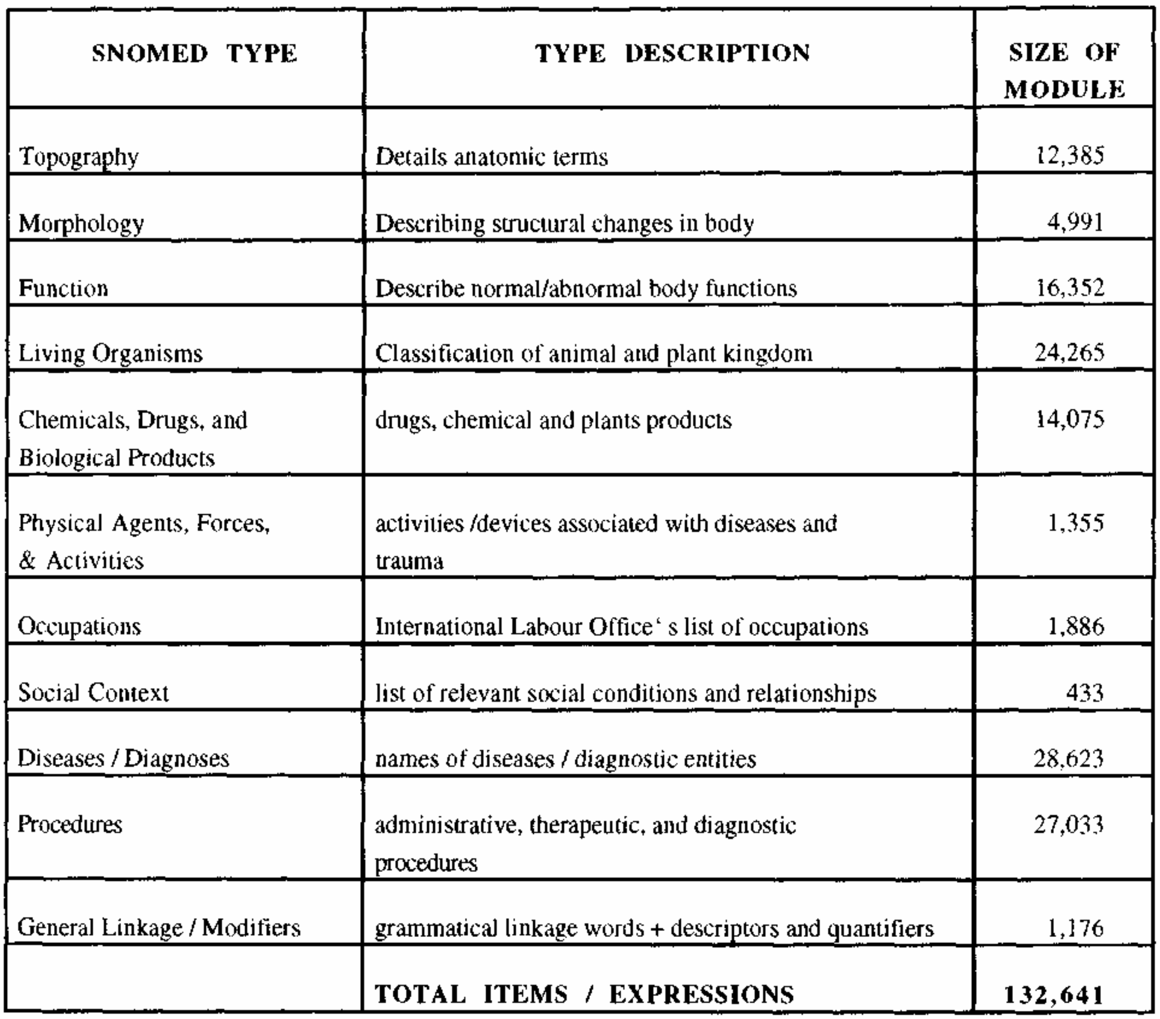
ABL_78_F_00943	TALALGIE DR
ABL_91_N_06113	LUMBO ISCHIALGIE

MDD_01_N_00087	0 1 06 OTALGIE REFERRED PAIN
MDD_01_N_02027	0 1 02 CERVICALGIE

- the suffix *-alis/-aal/-ale* means ‘concerning (a body part/ ethiological agent)’ :

ABL_73_F_00818	PROLAPSUS HEMORROIDAL
ABL_86_N_17868	ZONA THORACALIS

MDD_01_N_01011	0 1 06 SPONDYLOSE THORACAAL ERNSTIG
MDD_01_N_11043	0 0 02 FLUXIO HEMORROIDALIS



SNOMED TYPE	TYPE DESCRIPTION	SIZE OF MODULE
Topography	Details anatomic terms	12,385
Morphology	Describing structural changes in body	4,991
Function	Describe normal/abnormal body functions	16,352
Living Organisms	Classification of animal and plant kingdom	24,265
Chemicals, Drugs, and Biological Products	drugs, chemical and plants products	14,075
Physical Agents, Forces, & Activities	activities /devices associated with diseases and trauma	1,355
Occupations	International Labour Office' s list of occupations	1,886
Social Context	list of relevant social conditions and relationships	433
Diseases / Diagnoses	names of diseases / diagnostic entities	28,623
Procedures	administrative, therapeutic, and diagnostic procedures	27,033
General Linkage / Modifiers	grammatical linkage words + descriptors and quantifiers	1,176
	TOTAL ITEMS / EXPRESSIONS	**132,641**

Table 2. SNOMED types, type descriptions and size of the modules.

- the suffix *-ase / -asis* refers to the presence of a physical agent / living thing) :

ABL_81 _F_02305	LITHIASE RECIDIVANTE URINAIRE
MDD_01_N_01069	0 0 06 CHOLECYSTOLITHIASIS

- the suffix *-emie* refers to the blood :

ABL_73_F_01855	CIRHOSE ETHYLIQUE AVEC LEGERE ANEMIE HYPERCHROME
ABL_75_N_04207	HYPERLIPOPROTEINEMIE 2A. NEURITIS RE ULNARIS. PREDIABETES
MDD_01_N_00060	0 1 06 GLYCEMIE NUCHTER VERHOOGD
MDD_01_N_01063	0 1 06 HYPERURICEMIE

- the suffix *-ite / -itis* refers the the concept of inflammation :

ABL_78_F_01420	APPENDICITE. ADENITE MESENTERIQUE
ABL_78_N_01829	NASOPHARYNGITIS
MDD_01_F_13084	1 0 02 OTITE SEREUSE A MINIMA G
MDD_01_N_05032	0 0 02 TENDINITIS VAN DE ARMSPIEREN

- the suffix *-lyse /-lysis* refers to the physical concept of ‘solution’ :

ABL_86_N_03227	BILATERALE SPONDYLOLYSE.
ABL_87_N_15719	EPIFYSIOLYSE LINKER HEUP.
ABL_89_F_12335	EXCERESE CORPS ETRANGER + NEUROLYSE 3E RAYON MAIN GAUCHE

MDD_01_N_11038	0 1 02 TENOLYSE FLEXOREN RE WIJSVINGER + SYNOVECTOMIE

- the suffix *-oom / -ome / -oma* refers to a protuberance :

ABL_79_F_04560	GRANULOME POST-VASECTOMIE
ABL_79_N_00164	LIPOOM RELIES
ABL_86_F_07054	CONDYLOMES ACUMINES ANAUX.
ABL_91 _N_07468	HAEMATOMA LI OOGLID, GEEN FRAKTUUR ORBITA

MDD_01_N_01026	0006 HEMATOOM SUBUNGEAAL
MDD_01 _N_02002	0 1 02 LIPOOM GOEDAARDIG LINKER BOVENARM
MDD_01 _N_02060	0 1 02 CARCINOMA SIMPLEX KWAADAARDIG ANUS
MDD_01_N_1201 2	0 0 06 MULTIPLE FIBROMEN IN HUIDPLOOIEN

- the suffix *-pathie* means ‘suffering’ / ‘affection’ :

ABL_87_F_01275	LUMBAGO SUR DISCOPATHIE L5 S1.
ABL_89_N_00438	ISCHEMISCHE CARDIOPATHIE.

MDD_01_N_09011	0 1 02 FIBROCYSTISCHE MASTOPATHIE+PAPI
MDD_01_N_09061	0 1 02 TENDINOPATHIE RE LIES

- the suffix *-plegie* refers to a paralysis :

ABL_77_N_03033	FRACTUUR LUXATIE D12 PARAPLEGIE
ABL_89_N_24268	1. TETRAPLEGIE. 2. URETEROHYDRONEFROSE. 3. STUITWONDEN

One deviant feature of the sublanguage lexicon is the use of specific morphemes or lexemes (jargon words) related to the application domain, that are not found in the general language. This indicates that the lexicon of a sublanguage is not merely a subset of the lexicon of the general language. Examples of jargon words that are specific to the sublanguage of medical diagnostic expressions are :

ABL_75_N_02321	MITRALISTENOSE. LICHTEINSUFF NODALE EXTRASYSTOLIE
ABL_79_N_00170	PRECORDIALGIEN. CERVICALE ARTHROSE
ABL_80_F_00258	CHONDROMALACIE DE LA ROTULE *G*
ABL_86_N_03227	BILATERALE SPONDYLOLYSE.

MDD_01_N_00086	0 1 06 GLOMERULONEFRITIS ACUUT SNEL PROGRESSIEF
MDD_01_N_01017	0 1 06 SUPRASPINATUSSYNDROOM
MDD_01_N_01038	0 1 06 BRADYCARDIE SINUSAAL + ARRESTEN
MDD_01_N_08042	0 I 02 NEFROURETEROLITHIASE

Expressed in a preferential mode, a sublanguage includes a significant proportion of frequent domain dependent words which are subject to another distribution than in the general language. This phenomenon has been observed by [KELLY 77], who studied the influence of restricted vocabulary on performance in problem-solving tasks, and by [FALZON 84] who noticed in the field of task-oriented dialogues a more frequent use of a subset of a restricted sublanguage lexicon than of the rest of that lexicon. [MICHAELIS e.a. 77] showed that this kernel lexicon consists of words which are rare on the basis of frequency tables of the general language (in written and oral forms). These results obtained under experimental conditions corroborate with a study of real-scale communication situations in the field of oral requests for administrative information made by [DEVILLE 89].

A first observation of the ABL and MEDIDOC corpora indicates that the sublanguage of medical diagnostic expressions consists of three classes of words with a specific morphological and lexical distribution for each class, namely (i) function words, (ii) domain-dependent words and (iii) domain-independent words.

(i) Function words

The class of function words does not only include words that are traditionally labelled as syntactic functional words (articles, prepositions, conjunctions), but extends to the classes of words related to units of measure : time (heure, jour, mois, année), months (Janvier, février, ...), days (lundi, ...), numerals (including hours, dates, figures, ...), adverbs of time, place, frequency and modality. Most of these words are classified under the SNOMED type 'General Linkage / Modifiers' . This first class of words is closed and independent from the application, as they express universal semantic categories of time, space, numerals, etc. in a given culture.

(ii) Domain-dependent words

To this class correspond the words specifically related to the application domain of the medical diagnostic expressions. This class consists of a kernel lexicon made of words that are prototypically used to refer to the diagnostic expression as such, according to a typology of 'affections' (ie. essentially the SNOMED types of Disease / Diagnosis, Morphology and Functions) and associated concepts (ie. essentially the SNOMED types of Topography and Procedures, and some subclasses of the Living Organisms, Chemicals / Drugs and Biological Products, Physical Agents, Forces and Activities). Note that most of these words belong to the jargon specific to that domain. The class of domain-dependent words is virtually closed.

(iii) Domain-independent words

This class consists of all non-function and non-specialized words (i.e. that can be found in the general language) that are used in the application sublanguage. This is the case, for example, with terms referring to household appliances, devices and tools, transport vehicles, clothing materials (SNOMED type Physical Agents, Forces and Activities). Other examples are terms referring to technical, administrative and managerial workers, sales and service workers (SNOMED type Occupations), and terms referring to life style, religion, philosophy, economic status (SNOMED type Social Context); words referring to basic chemical compounds, industrial products, foods (SNOMED type Chemicals, Drugs and Biological Products), reptiles, fishes, plants (SNOMED type Living Organisms). The class of domain-independent words shows a relative degree of closure within a given family of applications. It is interesting to note that most of these words are terms from the general language which are monosemous in the application domain. Therefore, this class of words shows a lower degree of closure than the domain-dependent lexicon.

As a result, the sublanguage lexicon of medical diagnostic expressions can be viewed as a 'multi-layered' structure, in which each stratum has a specific morphological and lexical distribution, and meets the closure property to a certain degree.

4.2. Syntactic aspects

At the syntactic level, a sublanguage can be described (i) by using a smaller number of rules than are necessary for the general language (this is for example the case in the sublanguage of weather bulletins, that is exempt of direct questions and imperative forms [CHEVALIER 78]), and (ii) includes constructions that are ungrammatical in the general language (another example is the sublanguage of aircraft hydraulic maintenance, that is characterized by the absence of articles and subordinate conjunctions ('that' complementizer [LEHRBERGER 82]).

Note that these syntactic phenomena can be described in preferential terms : syntactic constructions and categories of a given sublanguage show another distribution than in the general language. The sublanguage of science and technology, for example, is characterized by the overuse of some morpho-syntactic structures such as nominalized forms and passive constructions, and differ from the general language in the distribution of word categories (nouns, verbs, adjectives, pronouns, etc.) [SAGER e.a. 80].

The sublanguage of medical diagnostic expressions is highly marked in its syntactic structures and categories, as illustrated by the following examples :

1 . absence of finite / infinite verb forms and overuse of nominalized forms

ABL_73_F_02860	LESION MENISQUE INTERNE GAUCHE
ABL_77_N_01379	CONTUSE RE ENKEL
MDD_01_N_02017	0 1 02 FRAKTUUR SCHEDELDAK GESLOTEN PARIETAAL
MDD_01_N_03011	0 0 23 INVERSIETRAUMA LI ENKEL

2. reduced use of articles and of some prepositions

ABL_80_F_01707	DECHIRURE PARTIELLE LIGAMENT LATERAL INTERNE GENOU DR
ABL_91_F_08952	AGRESSION BARRE DE FER - CONTUSION GENOU.
MDD_01_N_00055	0 1 06 CONTUSIE VINGER WIJSVINGER RE
MDD_01_N_00060	0 1 06 GLYCEMIE NUCHTER VERHOOGD

In the canonic structure of the noun group in Dutch the adjective (group) precedes its head (noun). This patterns is sometimes broken in Dutch diagnostic expressions, where the adjective is placed after the noun which it refers to. This is illustrated by the following examples :

ABL_73_N_04178	OTITIS BILATERAAL
ABL_91_N_00294	CORNEA EROSIE LI OOG, ONBEPAALD, MEDISCH
MDD_01_N_00094	0 1 06 PROSTATITIS CHRONISCH ABACTERIEEL
MDD_01_N_01011	0 1 06 SPONDYLOSE THORACAAL ERNSTIG

For operative reasons, the preferred structure of the noun group in the ANTHEM language representation model is ADJ + NOUN, whereas the pattern NOUN + ADJ will be considered as to be deviant, and will not generate an equivalent structure after translation.

4. absence of subordinate clause

ABL_77_F_04950	FRACT MALLEOLE EXT DR RUPTURE LIGAMENT INT GENOU DR SUSPICION DECHIRURE LIGAMENT CROISE ANTERIEUR
ABL_87_F_06938	CONTUSION POIGNET GAUCHE SUSPICION ATTEINTE DU CUBITAL
MDD_01_N_11006	0 1 06 RX. MOGELIJKHEID KLEIN JUXTA-BULBAIR ULC

5. overuse of embedded prepositional phrases (with optional use of preposition)

ABL_86_N_04442	CONTUSIO 3E TEEN RECHTER VOET
ABL_90_F_01311	FRACTURE ECRASEMENT 3E PHALANGE 5E DOIGT DROIT.
MDD_01_N_00002	0 1 06 FRACTUUR DISTAAL STUK VAN ACROMION LI
MDD_01_N_00055	0 1 06 CONTUSIE VINGER WIJSVINGER RE

Note that the absence of prepositions in these 'deviant' constructions might cause the ANTHEM prototype to overgenerate ambiguous representations/translations in some cases. Note however that examples like 'ABL_86_F_15447 DEPRESSION REACTIONNELLE DECES DE SA MERE' might be then considered as ill-formed diagnostic expressions.

To conclude, the medical diagnostic expression is prototypically made of a nominal syntagm that consists of a noun preceded or followed by an adjective /adjective group and optionally followed by a prepositional phrase (preposition followed by a nominal syntagm) in a recursive structure.

4.3. Semantic aspects

Not only the size of a sublanguage lexicon is restricted. Also the potential meanings per words are limited, due to the restricted domain of application. The vocabulary of a sublanguage includes non-specific terms (present in the general language) which, in most cases, are monosemic. Conversely, as [GUILBERT 73] notes, there is an accentuation of polysemy in general language use. The author explains that this univocal characteristic of sublanguage terms is not inherent to the word-form itself, but to its usage in a specific communication context, and to the implied reference.

In the ANTHEM application sublanguage, the domain independent words we discussed in the previous section show a high degree of monosemy, as illustrated by the following examples (cfr. 'verre', 'balle', 'werk', and 'scheiding') :

ABL_77_F_01786	1) PERFORATION OEIL DR (CORNEENNE AVEC INCRUSTATION DE VERRE) 2)KERATITE TRAUMATIQUE O G 3)CICATRICES DE PLAIES CUTANEES
ABL_88_F_05395	PLAIE PAR BALLE CUISSE DROITE.
MDD_01_N_09038	0 1 02 ALLERGISCHE ALVEOLITIS OP STOFWERK
MDD_01_N_12084	0 0 02 SCHEIDING

If we consider sublanguage according to a purely restrictive mode, we can state the following principle : the reduction in the number of potential meanings of a given word which is used in a sublanguage context, is proportional to the 'rate' of monosemy (expressed in N° occurrences of a word with meaning A / total number of its potential meanings) of that term in the sublanguage in question. The major drawback of this view of sublanguage is that a semantic reduction principle of this kind is too rigid, which is due to its strict adherence to a corpus-oriented analysis.

Though this semantic reduction process may be promising for computational applications, we prefer to follow [MARTIN & TEN PAS 91] and tackle this point in terms of preferential rules, indicating that the basic syntactic / semantic categories of a sublanguage term are ordered differently from the categories of the same term in general language. So we are not talking of restriction in terms of exclusion but in terms of preferential hierarchy.

Note, however, that the current version of the ANTHEM prototype is expected to generate all possible alternative representations of a diagnostic expression, as its lexicons are not built up on semantic preferential bases.

Interestingly, [MARTIN & TEN PAS 91] note another semantic characteristic of sublanguages : in a given domain of application, some salient concepts and concept types are distinguished, that stand in specific relation between themselves and to other concepts, so as to form the starting point for the semantic description of that domain. This means that sublanguages do not necessarily appeal to new concepts and relations, but that the underlying conceptual structure of sublanguages is organized differently from that of the general language.

This is the case with the sublanguage of medical diagnostic expressions, where only three basic concept types referring to the 'diseases' (Diseases/Diagnoses, Morphology and Functions) and one to body parts (Topography) are central in the description of the domain (i.e. they prototypically fulfill predicate slots). Other concepts like 'patient' - or more generally all ‘animated agents’- have a peripheral role in the ANTHEM application language. Conversely, in the general language one expects that objects like animated agents fulfill a central function in many ‘States of Affairs’. The specific conceptual organisation of the domain of medical diagnostic expressions is reflected in the semantic model discussed in section 5.

4.4. Collocational aspects

Besides the observation that some collocations of the general usage rarely appear in a given sublanguage, and that, on the other hand, there are sublanguage specific collocations / co-occurrence patterns [HIRSCHMAN & SAGER 82], we would like to examine the status of collocations in a sublanguage in the light of their formation process. A collocation is generally defined as a word combination that (i) consists of two words : one dependent and one independent lexeme, (ii) is generated by specific syntactic rules, (iii) expresses and is based on a semantic relation that is mostly unpredictable, and (iv) that is lexicalized by unpredictable and conventional means.

Following [MARTIN & TEN PAS 91], it is worth noticing that the syntactic-lexical unpredictability of collocations in a general usage are to be explained by the synonymic dimension of the language, that triggers several word candidates in order to express a given conceptual combination (e.g. *fluent speech / language; café fort / léger).* This conventional choice in the lexical variation of a collocation is strongly reduced in the case of a sublanguage, as word co-occurrences are mainly motivated here by a conceptual combinative dimension. In the perspective of automatic semantic parsing, the question will be how these sublanguage word co-occurrence patterns can be coded in the lingware so as to reflect their corresponding conceptual combinative dimension.

This phenomenon is observed in the ANTHEM sublanguage, where the co-occurrence of the adjectival modifiers like ‘chronique / chronisch’ and their preceding / following modified head illustrates the conceptual link between this SNOMED type of Modifier and its related subset of SNOMED type of Diseases / Diagnoses, as in the following examples :

ABL_73_N_01591	GASTRITIS. CHRONISCHE BRONCHITIS. CHRONISCH ALKOLISME
ABL_73_F_03656	ETHYLISME CHRONIQUE
ABL_80_F_02857	APPENDICITE CHRONIQUE. HYPERTHYROIDIE. DEPRESSION NERVEUSE
ABL_80_N_02261	DEPRESSIE. CHRONISCHE BRONCHITIS

MDD_01_N_00071	0 1 06 BRONCHITIS CHRONISCH EMFYSEMATEUS
MDD_01_N_00094	0 1 06 PROSTATITIS CHRONISCH ABACTERIEEL
MDD_01_N_01077	0 1 06 HYPOTENSIE CHRONISCH
MDD_01_N_02010	0 1 02 TRACHEITIS CHRONISCH

4.5. Pragmatic aspects

Sublanguage use is characterized by lexical and syntactic variations which result from the adaptation with respect to the speaker and his ‘level of expertise’ in a given communicative situation. [MARTIN & TEN PAS 91] distinguish four basic types of situation with related use of sublanguage types : from the communication between experts to the communication between non-experts through interdisciplinary communication and communication between expert and layman. These four types of sublanguage can be placed on a continuum, where each change of situation is reflected by a gradual shift in the use of the vocabulary. To a situation where the dialogue partners are both experts in the domain, corresponds a typical sublanguage use, with monosemous words and jargon terms, whereas in the sublanguage used in a situation where both partners are laymen the terms related to the specific domain are in a smaller proportion, and replaced by synonymous terms from the general language.

If we move along this continuum, we notice that speaker and receiver are faced with lexical choices in order to ‘tune up’ the communication. On the basis of his level of knowledge in the domain, a speaker will adapt his sublanguage to his opponent's -supposed or real - level of expertise in that same field. Hence the preferential use of synonymous words according to the communication situation (from typical sublanguage use to general language use).

In ANTHEM, the sublanguage of medical diagnostic expressions reflects a communicative situation of the first type (expert /expert) where the clinical description of the patient is in first instance meant for other specialists rather than for the patient himself. Hence the syntactic and semantic properties of the sublanguage as described above. We briefly describe in the next section how some of these sublanguage features have been modeled and implemented in the context of the ANTHEM prototype development.

5. THE ANTHEM REPRESENTATION MODEL

Following [DIK 89] we refer to the underlying semantic representation of a diagnostic expression as a predication. A predication is a structure that consists of a predicate with an adequate number of terms functioning as arguments of that predicate. Terms are phrases (i.e. noun phrases or prepositional phrases) used to refer to entities or to a set of entities in the conceptual world of a given sublanguage (a sublanguage being defined here as 'a set of expressions referring to a limited and well-defined application domain and used for a specific function'). A predicate (or head) is a noun phrase (i.e. noun or adjective) capturing semantic properties of — or semantic relations between — its arguments. A predication refers to a Sublanguage world Configuration. A Sublanguage world Configuration (hence SwC) is a cluster of sublanguage conceptual entities that are expressed in terms of their mutual relationships [DEVILLE 89].

In the case of ANTHEM, the sublanguage conceptual entities are defined as minimal units that not only refer to atomic (e.g. arm) or complex objects (e.g. left hand palm), but also mainly to states (e.g. inflammation), relations (e.g. part of) or to a lesser extent actions (e.g. ingestion) and processes (e.g. to fall). More precisely, the terms of the ANTHEM application sublanguage refer to objects, and predicates to states, relations and actions/processes. Predicates (or heads) are selected from a finite set of semantic types. A semantic type captures the prototypical semantic and combinatorial properties shared by a set of predicates. Most of the semantic types used in the ANTHEM representation model are inherited from the Systematized Nomenclature of Human and Veterinary Medicine (SNOMED) [COTE e.a. 93].

In a predication, the relation between the predicate and its argument(s) is specified by means of a case. A case is the expression of a prototypical semantic function or role fulfilled by a predicate's term with regard to the semantic class from which that predicate derives. The case frame of a semantic type is the sequence of required cases for the definition of the set of SwCs represented by that semantic type. As a predicate and its arguments refer to a particular SwC, a semantic type with its associated case frame refers, on a more prototypical conceptual level, to a class of SwCs. Such a higher level structure specifies the semantic roles of the arguments of the derived predicate, in relation to its corresponding semantic type.

A predication can be extended by means of one or more peripheral arguments. Peripheral arguments are not constitutive of the definition of a SwC but express (i) the spatio-temporal setting of that SwC, (ii) the secondary entities participating in the SwC or (iii) give information on the manner or conditions in which the SwC takes place. As opposed to central arguments, the semantic functions of peripheral arguments are not necessary to define a set of SwCs in terms of a semantic type with its associated case frame. Figure 1 illustrates the representation of the expression "RUPTURE MENISQUE INTERNE GENOU DROIT" (Rupture medial meniscus right knee) in terms of the model described above.

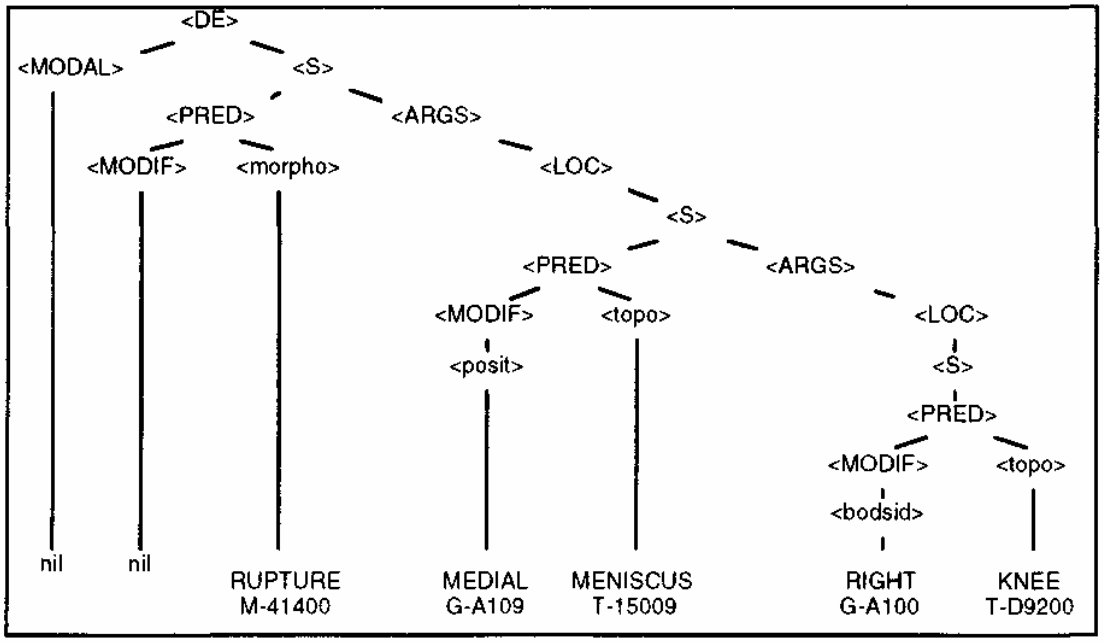


Figure 1 . Representation of the expression "RUPTURE MENISQUE INTERNE GENOU DROIT" (Rupture medial meniscus right knee).

The ANTHEM representation model is being implemented in the form of unification based rules. These rules form one part of the lingwares, referred to as the ANTHEM grammars (French, Dutch and German). The ANTHEM prototype contains one grammar module for each language. The grammar rules are applied to objects which are the lexical entries corresponding to the terms used in the ANTHEM application domain. These lexical entries form the second part of the lingwares, referred to as the ANTHEM lexicons. The ANTHEM prototype includes one lexicon for each language (here French, Dutch and German). As it is inappropriate to further describe the technical aspects of the ANTHEM lingwares in the limited context of this paper, we refer to [DEVILLE e.a. 94] for a detailed account of the specificity of the ANTHEM representation model from the implementational point of view.

6. CONCLUSION

This paper advocates the methodological viewpoint that the study of natural language in an automatic processing perspective is motivated by the notion of sublanguage. A definition of sublanguage is proposed in the light of mainstream literature. In the context of the development of an MT prototype for healthcare applications (the ANTHEM project), we stress the need to define the language of medical diagnostic expressions on the basis of representative corpora generated in real scale clinical environments. In order to illustrate our methodological hypothesis, we have scanned the most prototypical language data of medical diagnostic expressions by means of a sublanguage grid. We then briefly sketched the model in which these sublanguage features have been formalized and implemented in the operative context of the ANTHEM prototype development.

7. REFERENCES

[ADRIAENS 93]
ADRIAENS G. : *A Simplified English Grammar and Style Checker / Corrector (SECC),* LRE-62-064 project.

[CHEVALIER 78]
CHEVALIER M. : *TAUM-METEO : description du système.* Groupe de Recherche en Traduction Automatique, University of Montreal, 1978.

[COTE e.a. 93]
The Systematized Nomenclature of Human and Veterinary Medicine. SNOMED International. Introduction. Côté RA, Rothwell DJ, Beckett RS, Palotay JL College of American Pathologists, Northfield, 1993.

[DEVILLE e.a. 94]
DEVILLE G., HERBIGNIAUX E. & STREITER O. : *The CAT-2 Interface Structure Applied to the Medical Diagnose Sublanguage,* ANTHEM LRE-Project 62-007, Deliverable D2-1, 1994.

[DEVILLE & HERBIGNIAUX 94]
DEVILLE G. & HERBIGNIAUX E : *Methodological Principles for the Elaboration of Multilingual Corpora of Medical Diagnostic Expressions.* ANTHEM LRE-Project 62-007, Deliverable D2-2, Part I, 1994

[DEVILLE 89]
DEVILLE G. : *Modelization of Task-Oriented Utterances in Man-Machine Dialogue System.* Ph.D. Thesis, University of Antwerp, Antwerpen, 1989.

[DIK89]
DIK S. : *A Theory of Functional Grammar.* Dordrecht : Foris, 1989.

[FALZON 84]
FALZON P. : Les langages opératifs. In PIERREL J.M., CARBONELL N., HATON J.P. & NEEL F. (eds.) : *Dialogue homme-machine à composante orale. Actes du Séminaire GRECO Communication Parlée - CNRS,* Nancy, pp. 364-383, October 1984.

[FALZON 86]
FALZON P. : *Langages opératifs et compréhension operative.* Ph.D. Thesis in Social Sciences, University of Paris V - Sorbonne, Paris, 1986.

[GUILBERT 73]
GUILBERT L. : La spécificité du terme scientifique et technique, in *Langue Française,* 17, pp. 5-17, February 1973.

[HARRIS 68]
HARRIS Z. : *Mathematical Structure of Language.* John Wiley & Sons, New-York, 1968.

[HIRSHMAN & SAGER 82]
HIRSHMAN L. & SAGER N. : Automatic Information Formatting of a Medical Sublanguage. In KITTREDGE R. & LEHRBERGER J. (eds.) : *Sublanguage,* de Gruyter, Berlin, pp.27-80, 1982.

[KELLY 77]
KELLY M.J. & CHAPANIS A. : Limited vocabulary natural language dialogue. *International Journal of Man-Machine Studies,* 9, pp. 479-501, 1977.

[KITTREDGE & LEHRBERGER 82]
KITTREDGE R. & LEHRBERGER J. (eds.) : *Sublanguage : studies of language in restricted semantic domain,* de Gruyter, Berlin, 1982.

[KOCOUREK 82]
KOCOUREK R. : *La langue française de la technique et de la science,* Brandstetter Verlag, 1982.

[LEHRBERGER 82]
LEHRBERGER J. : Automatic translation and the concept of sublanguage. In KITTREDGE R. & LEHRBERGER J. (eds.) : *Sublanguag,* de Gruyter, Berlin, pp. 81-106, 1982.

[MARTIN & TEN PAS 91]
MARTIN W. & TEN PAS E. : Subtaal en Lexicon, *Spektator,* tijdschrift voor nederlandistiek, nr 3/4, jaargang 20, 1991.

[McNAUGHT e.a. 91]
McNAUGHT J., MARTIN W., NKWENTI-AZEH & TEN PAS E. : *Feasibility of standards for terminological description of lexical items.* JR-Report Eurotra-7 Study, 1991.

[MICHAELIS e.a. 77]
MICHAELIS P.R., CHAPANIS A., WEEKS G.D., & KELLY M.J. : Word usage in interactive dialogue with restricted and unrestricted vocabularies. *IEEE Transactions on Professional Communication,* PC-20, 4, pp.214-221, 1977.

[MOSKOVICH 82]
MOSKOVICH W. : What is a sublanguage ? the notion of sublanguage in modern soviet linguistics. In KITTREDGE R. & LEHRBERGER J. (eds.) : *Sublanguag,* de Gruyter, Berlin, pp.191-205, 1982.

[PAULUSSEN & MARTIN 92]
PAULUSSEN H. & MARTIN W. : DILEMA 2 : A Lemmatizer-Tagger for Medical Abstracts, in *Proceedings of the Third Conference on Applied NLP,* ACL, Trento, pp. 141-146, 1992.

[SAGER 82]
SAGER N. : Syntactic formatting of science information, in KITTREDGE R. & LERBERGER J. eds.) : *Sublanguage : studies of language in restricted semantic domains,* de Gruyter, Berlin, 1982.

[SAGER e.a. 80]
SAGER J., DUNGWORTH D. & McDONALD P. : *English Special Languages; principles and practice in science and technology.* Oscar Brandstetter Verlag KG, Wiesbaden, 1980.