Institut Universitari de Lingüística Aplicada
Universitat Pompeu Fabra

Programa de doctorat: Lingüística Aplicada(lèxic i discurs)
Bienni 1994-1996

Tesi doctoral

Extracció de terminologia: elements per a la construcció d'un SEACUSE

(Sistema d'Extracció Automàtica de Candisats a Unitats de Significació Especialitzada)

Per optar al títol de doctora per la Universitat Pompeu Fabra

Rosa Estopà Bagot

Directora: Dra. M. Teresa Cabré Castellví

1999

Dipòsit legal: B.14059-2002
ISBN: 84-699-8512-4

Menció honorífica concedida pel Jurat dels premis INFOTERM Internacionals per a la recerca aplicada i el desenvolupament en el camp de la Terminologia, juntament amb el comitè executiu d'INFOTERM, per la tesi doctoral 'Extracción de terminología: elementos para la construcción de un SEACUSE"
Austria, agost de 2001

AGRAÏ MENTS

Una tesi doctoral és un procés llarg, resultat d'un conjunt de factors molt diversos: de la tradició i de la innovació, de l'opinió i de la creació, de la reflexió i de la imaginació, del dubtar i del preguntar, de la constància, de la sistematicitat, del fer i del refer... I en aquest camí de revolts i dreceres hi ha estones de solitud en què cerques raons que en justifiquin la continuació. És en aquests moments que paraules com les de Black Elk[1] em van ajudar a veure i valorar les persones que m'ajudaven a avançar "*sé que això que faré és una bona obra i com que* ***cap home sol no pot fer una bona obra****, primer faré una ofrena i pregaré a l'esperit del món que m'ajudi a ser veraç*".

Avui, a les acaballes del treball, vull recordar-me de tots els que me n'han facilitat, d'una manera o d'una altra, l'elaboració.

Crec sincerament que és un privilegi haver fet aquesta tesi com a membre de l'Institut Universitari de Lingüística Aplicada de la Universitat Pompeu Fabra i, en concret, col·laborant en el Projecte "La terminología científico-técnica: reconocimiento, análisis y extracción de información formal y semántica" (DGES - PB 96-0293), dirigit per la Dra. Cabré. Vull agrair també a l'IULA el fet de possibilitar-me l'any 1997 una estada de formació a França en el marc de l'Acció Integrada "Localización de secuencias nominales que permiten constituir términos en un corpus bilingüe (español-francés) especializado (informática y economía)" (HF1996-0215), en què vaig poder conèixer els projectes dels Drs. Bourigault, Habert i Jacquemin.

[1] "Als meus mestres" (Lakota Oglala, 1931). Bruchac, J. (1996) *La saviesa de l'indi americà*.

Agraeixo el suport rebut per professors, col·legues i amics de l'IULA. El meu agraï ment més sincer al professorat del programa de doctorat Lingüística Aplicada (Lèxic i discurs) (1994-1996) que han contribuï t i contribueixen en la meva formació acadèmica. Al Dr. Toni Badia pels seus comentaris i a la Dra. Mercè Lorente per la seva predisposició a ajudar-me i sobretot pels seus consells. A les professores Elisabet Solé, Montserrat Ribas, Cleci Bevilacqua, Meritxell Domènech, i molt especialment a Cristina Gelpí, que com amigues m'han escoltat, aconsellat i animat contínuament. A Judit Feliu que m'ha alleugerat les tasques professionals. I a Jordi Vivaldi per la seva predisposició i per les hores de lectures i converses que hem compartit.

També haig de donar les gràcies a diferents professionals que desinteressadament han col·laborat en aquesta tesi: als metges Lluís Barona, Meia Faixedas, Pere Horta, Àlvar Martínez i, sobretot, Toni Valero; als traductors Xevi Burjons, Janet DeCesaris, Xavier Mas; a les terminògrafes Àngels Egea, Cristina Gelpí, Sílvia Martorell; i als documentalistes Lluís Codina, Montserrat Culubret, Mireia Ribera, Ferran Vera. I També a Lluís Pujol, per als gràfics i a Xevi Burjons i Marta Duran, per les observacions prescriptives. A tots ells els agraeixo el seu temps i els seus comentaris.

Vull agrair a la Dra. M. Teresa Cabré la direcció d'aquesta tesi, el seu mestratge constant i la seva amistat. Com a directora li dono les gràcies per la disposició, la clarividència, la confiança, les propostes, les idees, els comentaris, les lectures i relectures del text, i també pels consells ("*no et precipitis, dorm-hi i reflexiona, sobretot reflexiona-hi*"); com a mestra, pels més de deu anys d'exemple i com a amiga per tot allò que no es diu però que es viu i es guarda per al demà.

Finalment, vull agrair a la meva família la paciència i l'estimació demostrades en tot moment; als meus pares l'encoratjament constant; a la Cristina i a la Montse, la comprensió i l'estimulació permanents.

A en Jordi li estic agraï da per tot, però sobretot pel somriure que em regala dia a dia.

A tots els qui m'han ensenyat a raonar. I, de primer, als meus pares que m'han ensenyat a valorar les petites coses.

Jo mateix no ho sé pas, però ens caldrà anar cap allà on el raonament, com si fos el vent, ens porti.

[Plató, La República, 394d]

ÍNDEX GENERAL

VOLUM 1

ABREVIACIONS

A	adjectiu
A no esp	adjectiu no especialitzat
Adv	adverbi
Aesp	adjectiu especialitzat
art	article determinat
N	nom[1]
$N_{comú}$	nom comú
$N_{llatí}$	nom llatí
$N_{no\ term}$	nom no terminològic
N_{propi}	nom propi
N_{term}	nom terminològic
prep	preposició
SAdj	sintagma adjectiu
SAdv	sintagma adverbial
SN	sintagma nominal
SPrep	sintagma preposicional
UD	unitat discursiva
UF	unitat fraseològica
UFE	unitat fraseològica especialitzada
UL	unitat lèxica
ULE	unitat lèxica especialitzada
UP	unitat polilèxica
USE	unitat de significació especialitzada
UT	unitat terminològica
UTM	unitat terminològica monolèxica
UTP	unitat terminològica polilèxica
V	verb

Sistemes informàtics

CERPAT	cercador de patrons
EXCAT1	explorador de candidats a terme
SEACAT	sistema d'extracció automàtica de candidats a terme
SEACUSE	sistema d'extracció automàtica de candidats a unitats de significació especialitzada

[1] Per defecte entenem nom comú.

Diccionaris

DEM	*Diccionari enciclopèdic de medicina* d'Enciclopèdia Catalana (1990)
DIEC	*Diccionari de la Llengua Catalana* de l'Institut d'Estudis Catalans (1995)
DLC	*Diccionari de la Llengua Catalana* d'Enciclopèdia Catalana (1993)

0. INTRODUCCIÓ

La terminología teórica y práctica se enfrenta, desde hace tiempo, a dos problemas básicos para el tratamiento de las unidades terminológicas: la detección de términos y la segmentación de unidades terminológicas de estructura compleja. Hasta ahora, la terminología sólo disponía de la consulta con el especialista para decidir sobre la identificación definitiva de un término, sobre su correcta segmentación y sobre la pertinencia de este término en un ámbito de especialidad. A estos dos problemas hay que añadir la complejidad que presenta la automatización de estos procesos.

Projecte DGES PB 96-0293[1]

Els primers intents per extreure automàticament unitats terminològiques (UT) dels textos especialitzats s'inicien a principis dels anys vuitanta amb la finalitat d'automatitzar o semiautomatitzar algunes tasques de certes aplicacions terminològiques; però no és fins a finals dels vuitanta i sobretot als noranta, amb la creació de grans corpus textuals informatitzats, que els primers programes d'extracció automàtica de terminologia comencen a donar resultats positius.

De fet, des de l'aparició del que es considera el primer sistema d'extracció automàtica de candidats a terme —TERMINO[2]— fins avui, s'han dut a terme nombrosos projectes per dissenyar extractors (semi)automàtics de terminologia de naturalesa diferent; però, malgrat la gran quantitat d'estudis que s'han realitzat en aquesta línia, el reconeixement i la

[1] Aquest estudi forma part dels projectes que s'estan desenvolupant en la línia de recerca en lèxic a l'Institut Universitari de Lingüística Aplicada (IULA) de la Universitat Pompeu Fabra. En concret, la tesi s'emmarca en el projecte "La terminología científico-técnica: reconocimiento, análisis y extracción de información formal y semántica" (PB 96-0293), finançat per la Dirección General de Enseñanza Superior (DGES).

[2] TERMINO va ser un dels primers sistemes d'extracció automàtica de terminologia de base lingüística. La versió 1.0 d'aquest sistema va ser creada l'any 1989 pel grup de *Recherche et développement en linguistique computationnelle* (RDLC) del Centre ATO (Analyse de textes par ordinateur) de la Université du Québec à Montréal. Per a més informació vegeu [Perron, 1991]. Prèviament, altres projectes en el camps de la indexació de documents havien aportat elements valuosos per a la detecció de terminologia [Pratt, A.; Pacak, M, 1969], [Pacak, M.; Dunham, G., 1973].

delimitació automàtics de les unitats terminològiques (UT) a partir de textos no són encara satisfactoris.

Les principals dificultats amb què es troben els extractors terminològics són bàsicament dues:

1. Reconèixer quan una unitat lèxica té caràcter especialitzat; és a dir, saber quan una unitat és usada amb valor especialitzat.

2. Delimitar les unitats lèxiques complexes; és a dir, saber on comença i on acaba un sintagma terminològic.

Els extractors no poden reconèixer fàcilment unitats lèxiques especialitzades perquè aquestes no presenten cap element que les faci formalment identificables. I tampoc no poden delimitar totes les unitats especialitzades dels textos especialitzats si utilitzen exclusivament patrons morfològics o morfosintàctics per reconèixer i delimitar les unitats terminològiques (UT) perquè està a bastament demostrat que els patrons estructurals són un filtre massa permissiu per identificar les unitats terminològiques d'un domini determinat. En efecte, si s'usen patrons que fan referència **només** a la forma de les unitats, sovint es proposen com a candidats a terme delimitacions errònies i segments sense interès especialitzat. Per tant, també cal dotar els extractors de coneixement semàntic per tal que puguin detectar i delimitar automàticament les unitats especialitzadament pertinents de manera més exhaustiva i més precisa.

D'altra banda se sap que els extractors de terminologia solen reduir-se a la detecció de la unitat terminològica polilèxica (UTP) que, si bé és la unitat de significació especialitzada més abundant dels textos especialitzats, no és l'única.

0.1 Del treball de recerca a la tesi: antecedents

Aquesta tesi forma part d'un procés iniciat amb el treball de recerca que vam presentar el desembre de 1996 amb el títol *Les unitats terminològiques polilexemàtiques en els lèxics especialitzats: dret i medicina*.

En aquell treball ens vam proposar, a partir d'un corpus lexicogràfic especialitzat:

- descriure de manera exhaustiva els patrons estructurals de les UTP en català de dues àrees del saber científic, el dret i la medicina, i

- caracteritzar el no-terme, és a dir, analitzar els tipus de categories gramaticals que no formen mai part d'una UTP.

Aquests objectius estaven motivats pel fet que no hi havia estudis lingüístics sistemàtics de la forma dels sintagmes terminològics del català (estructura morfosintàctica, categoria gramatical, freqüència d'ús de formes, tipografia, etc.), estudis que eren el punt de partida de la majoria de **S**istemes d'**E**xtracció **A**utomàtica de **C**andidats a **T**erme (SEACAT) que funcionaven en aquell moment per a altres llengües.

Una de les aplicacions immediates dels resultats del treball de recerca podia ser, doncs, el disseny d'un SEACAT que utilitzés patrons estructurals per detectar les unitats terminològiques polilèxiques dels textos especialitzats en català[3].

[3] I de fet, els resultats del treball de recerca han estat utilitzats per J. Vivaldi per elaborar un primer SEACAT per al català anomenat EXCAT1, de gran utilitat per a la tesi que presentem.

Però les estructures que vam proposar en aquell treball estaven basades en corpus lexicogràfics i no en corpus textuals. Per tant, un element d'enllaç entre el treball de doctorat i la tesi que presentem és la validació dels patrons formals de les UTP proposades per a les unitats que trobem en els textos; i un altre punt de connexió entre el treball de recerca i la tesi és l'anàlisi dels resultats que genera un SEACAT basat en aquests tipus de patrons.

0.2 Des de la perspectiva lingüística de la terminologia. Supòsits teòrics de base

Seguint la proposta de la teoria terminològica de base comunicativa de Cabré (1997a, 1998a i b), assumim una sèrie de supòsits teòrics pel que fa a la terminologia i al seu objecte d'anàlisi, principis que emmarquen aquest treball de recerca:

1. La terminologia és alhora una matèria interdisciplinària i transdisciplinària, que, consegüentment, es pot estudiar des de diferents punts de vista. En aquest treball, prioritzem el seu **vessant lingüístic**, sense oblidar, però, que la terminologia s'inspira també en supòsits teòrics de les ciències cognitives, de les ciències de la comunicació i de les especialitats.

2. Tot i que tradicionalment l'objecte principal de la terminologia és la **UT,** en els textos especialitzats, trobem **altres unitats de significació especialitzada** que no són termes.

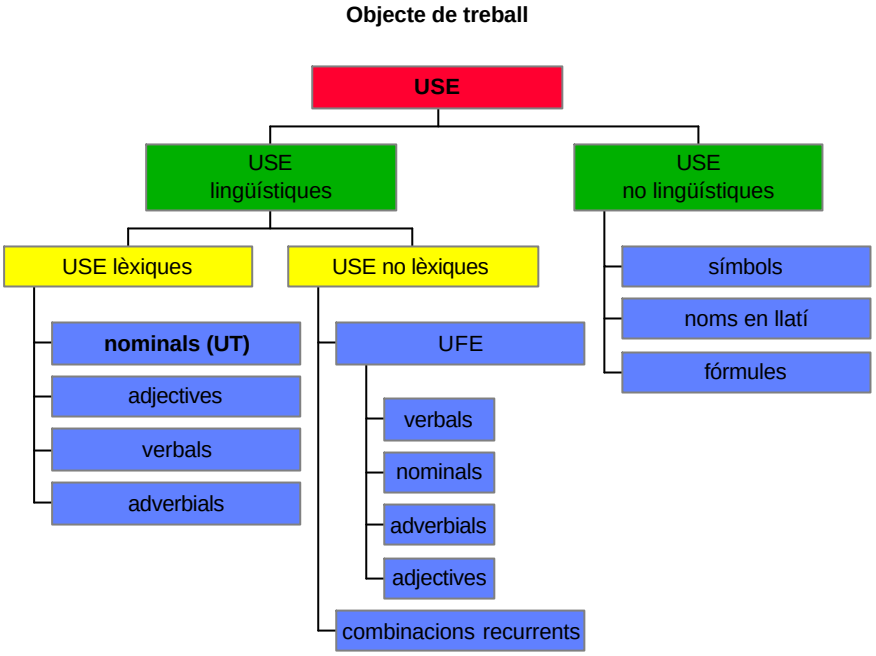
3. El context natural de les UT són els **textos,** però només **els especialitzats temàticament,** produïts bàsicament per especialistes. A partir d'aquest supòsit, deduïm que les UT només adquireixen valor terminològic en els textos especialitzats.

4. Tots els estudis aplicats estan condicionats per una finalitat professional concreta, i aquesta és una premissa que tindrem en compte en la proposta d'un extractor automàtic.

0.3 Les unitats de significació especialitzada. Objecte del treball

En un principi, l'objecte d'anàlisi de la tesi eren les **UT pertinents en un camp especialitzat concret,** enteses com un subconjunt de les **unitats lèxiques** del llenguatge natural que formen part d'una estructura conceptual especialitzada i que s'usen en discursos temàticament i intencionadament especialitzats. Però, a mesura que el treball ha anat avançant, hem constatat que la UT no és l'única unitat interessant dels textos especialitzats, sinó que hi ha altres unitats, que van més enllà del concepte clàssic del terme que també són especialitzadament pertinents per a una determinada finalitat professional. Per això, en el moment que hem constatat que els usuaris s'interessen per una unitat més genèrica que el terme, hem cregut oportú eixamplar el nostre objecte d'estudi. D'aquesta manera, hem passat de considerar només les UT a incloure-les en unitats més àmplies que anomenem **U**nitats de **S**ignificació **E**specialitzada (USE), unitats que vehiculen coneixement especialitzat i que formalment abracen diversos tipus d'unitats sígniques tant lingüístiques com no lingüístiques. El quadre següent mostra la diversitat d'USE[4]:

[4] Hi ha d'altres classes d'USE com, per exemple, les iconografies que no hem tractat perquè ens hem basat, com veurem més endavant, en el punt de vista de l'especialista.



0.4 Objectius del treball

Els objectius generals que ens proposem en aquest treball són els següents:

1. **Analitzar i avaluar el funcionament dels actuals sistemes d'extracció automàtica de candidats a terme** amb la finalitat de tenir elements per dissenyar un extractor per al català. Aquest objectiu implica:

    a) elaborar un estat de la qüestió dels sistemes d'extracció (semi)automàtica de terminologia per conèixer-ne els avantatges i inconvenients, i estudiar les causes de les seves limitacions

    b) comprovar que les estructures morfosintàctiques de les unitats terminològiques polilèxiques proposades en el treball de recerca *Les unitats terminològiques polilexemàtiques en els lèxics especialitzats* a partir d'un

corpus lexicogràfic, són les mateixes si s'apliquen a un corpus textual.

2. **Proposar elements de reconeixements de les Unitats de Significació Especialitzada (USE)** pertinents per poder dissenyar un sistema d'extracció automàtica d'aquestes unitats. Aquest objectiu es concreta en les accions següents:

    a) delimitar, des del punt de vista lingüístic, els diferents tipus d'USE que s'usen en l'àmbit de les ciències de la salut
    b) proposar elements i estratègies per reconèixer i detectar amb més precisió els diferents tipus d'unitats
    c) establir estratègies per eliminar al màxim les unitats no pertinents que solen generar els sistemes d'extracció automàtica.

3. **Dissenyar un model de Sistema d'Extracció Automàtica de Candidats a Unitats de Significació Especialitzada (SEACUSE) adequat a les necessitats professionals de diferents col·lectius d'usuaris.** Aquest propòsit recolza en diversos objectius que caldrà assolir prèviament:

    a) mostrar que cada activitat professional requereix unes necessitats diferents quant a les unitats especialitzades d'un text
    b) establir els criteris per determinar la pertinència d'una USE a l'interior de cada activitat professional
    c) proposar els perfils d'USE que requereixen algunes finalitats professionals específiques.

0.5 Idees prèvies

El treball parteix d'un conjunt de supòsits centrats bàsicament en dos eixos:

1. Els sistemes d'extracció.
2. Les unitats de significació especialitzada usades en l'àmbit de les ciències de la salut.

1. Pensem que els SEACAT actuals, per bé que són molt útils, no són del tot satisfactoris; i no ho són, d'una banda, perquè es limiten a detectar un sol tipus d'UT: les unitats terminològiques polilèxiques (UTP) deixant de banda les monolèxiques; i, de l'altra, perquè no totes les unitats que detecten són pertinents ni detecten totes les UTP pertinents dels textos.

   Per millorar un SEACAT, s'ha d'aconseguir reduir el silenci (les USE pertinents que no genera) i, per tant, detectar totes les USE sense limitacions; i, paral·lelament, s'ha de reduir considerablement el soroll (les unitats terminològicament no pertinents que genera) i, per tant, eliminar les seqüències que no siguin unitats de significació especialitzada.

El problema principal dels SEACAT actuals de base lingüística és que detecten les UT a partir de patrons morfològics i/o morfosintàctics que no són exclusius d'aquestes unitats. Per reconèixer i delimitar una unitat de significació especialitzada (USE) no es pot recórrer només als aspectes morfosintàctics, sinó que cal tenir en compte també els elements lèxics, morfològics, sintàctics, semàntics i pragmàtics.

2. Quant a les unitats de significació especialitzada en general i a les unitats terminològiques en particular, en primer lloc, partim del supòsit

que les UT no són les úniques unitats amb significat especialitzat, sinó que en els textos hi ha moltes altres unitats de significació especialitzada, lingüístiques i no lingüístiques, que interessen també els usuaris, que hauria de tenir en compte un extractor. Les unitats de significació especialitzada (USE) lingüístiques poden ser lèxiques i sintàctiques. Les USE lèxiques poden ser noms, adjectius, verbs o adverbis i les USE no lèxiques poden ser UFE o combinacions nominals molt freqüents. Les USE no lingüístiques inclouen unitats que formen part d'alguna nomenclatura científica (símbols, noms llatins) i fórmules[5].

En segon lloc, assumim que les UT —subconjunt de les USE— són totes nominals i tenen valor referencial, i pressuposem que les USE no nominals no corresponen a UT, tot i que el caràcter nominal no implica que la unitat sigui obligatòriament un terme.

En tercer lloc, creiem que totes les USE, simples o complexes, ofereixen alguns recursos morfològics, sintàctics, semàntics i pragmàtics —més sistemàtics o més estadístics— per reconèixer-les.

Finalment, pensem que la pertinència d'una USE depèn de l'activitat professional que es realitzi; de manera que en un text d'especialitat pot haver-hi unitats que des del punt de vista temàtic siguin pertinents, però que des del punt de vista de la finalitat professional no ho siguin.

0.6 Interès i aplicacions del treball

Aquest treball té al nostre entendre un triple interès. En primer lloc, ofereix un nou plantejament teòric de la unitat terminològica que té en

[5] En aquest treball hem decidit no tractar els dibuixos i iconografies que apareixen en els textos especialitzats.

compte els interessos professionals. En segon lloc, aporta una anàlisi descriptiva de les unitats de significació especialitzada usades en l'àmbit de la medicina. En tercer lloc, dóna criteris per valorar exhaustivament i críticament els SEACAT actuals i proposa un model de sistema d'extracció automàtica de candidats a unitats de significació especialitzada (SEACUSE) adequat a les necessitats de les activitats professionals.

L'aplicació més immediata dels resultats d'aquesta tesi és l'elaboració d'un **sistema d'extracció d'unitats de significació especialitzada (SEACUSE)** de base lingüística que tingui en compte els interessos dels col·lectius d'usuaris quan realitzen una activitat professional.

Els camps d'aplicació d'un SEACUSE són molts i diversos, i entre ells podem mencionar la documentació, la traducció, la docència, les especialitzacions, la terminografia, la lingüística, etc. Una eina que realitzi el buidatge de les unitats de significació especialitzada dels textos de manera automàtica pot ser un component de moltes aplicacions en lingüística aplicada, com la confecció de sistemes d'indexació automàtica, sistemes hipermèdia de relació de la informació, sistemes de categorització de documents especialitzats, sistemes d'interrogació, sistemes d'ajuda de creació de tesaurus, glossaris, diccionaris, bases de dades, bases de coneixement, sistemes de traducció automàtica o assistida, sistemes multimèdia d'autoaprenentage o de suport a la docència, sistemes experts per al processament del llenguatge natural, sistemes d'explotació de corpus textuals especialitzats, etc.

0.7 Organització del treball

En aquest capítol previ hem establert les bases del treball: el marc, els antecedents, l'àmbit d'anàlisi, les unitats d'anàlisi, els objectius, les hipòtesis i les aplicacions de la tesi.

A partir d'aquí el treball es divideix en tres parts: una primera part d'anàlisi i crítica del funcionament d'extractors existents —capítols 1, 2, 3 i 4—, una segona part d'experimentació i proposta d'un extractor multifuncional —capítols 5, 6 i 7—; i una última part que inclou les conclusions i la bibliografia de referència —capítols 8 i 9.

Més detalladament, el primer capítol està dedicat a la descripció dels sistemes d'extracció automàtica de candidats a terme (SEACAT) i a l'anàlisi i valoració dels principals SEACAT amb l'objectiu d'elaborar un estat de la qüestió en aquest camp que evidenciï les característiques i sobretot les limitacions d'aquests sistemes.

El segon capítol està consagrat primerament a validar les hipòtesis que vam postular en el treball de recerca sobre els patrons estructurals de les UTP i en segon lloc a comprovar les principals limitacions dels SEACAT que es basen en patrons morfosintàctics. Aquestes limitacions es manifesten en dos aspectes: el silenci (unitats pertinents no detectades per l'extractor) i el soroll (unitats no pertinents presentades com si ho fossin).

Les dades que es desprenen de l'anàlisi del silenci i el soroll s'estudien en el tercer i quart capítol, respectivament. Així, primer analitzem els tipus i les causes de silenci que produeixen els SEACAT, i tot seguit els tipus i les causes del soroll generat per aquests sistemes.

El cinquè capítol enceta la segona part dedicada a propostes amb proposicions de com un sistema d'extracció automàtica podria reduir el silenci i el soroll, és a dir de com es podria millorar el seu funcionament, de manera que els seus resultats fossin més acostats al reconeixement i delimitació manuals de les unitats de significació especialitzada.

El sisè capítol introdueix el punt de vista de l'usuari i planteja el fet que no totes les activitats professionals requereixen els mateixos tipus ni el mateix nombre d'unitats especialitzades d'un text. Aquesta hipòtesi és verificada a través d'una prova experimental basada en les necessitats de quatre activitats professionals diferents.

Tanca aquest segon bloc el capítol setè, en què s'exposa una proposta de model de SEACUSE que, a més de les estratègies plantejades en el capítol cinquè, té en compte les finalitats dels professionals a l'hora de presentar els resultats.

En les conclusions, se sintetitzen els supòsits de partida, es validen o refuten les hipòtesis que hem anat formulant al llarg del treball i es presenten els resultats més significatius. Tot seguit, es fa referència als suggeriments que han sorgit durant el treball i a les seves limitacions.

Al final s'inclou la bibliografia que hem tingut en compte, ordenada alfabèticament.

En un segon volum incloem en annex els materials de què hem partit. El primer annex inclou el text que configura el corpus utilitzat per al buidatge manual. El segon annex mostra els resultats dels buidatges manuals del text realitzats per diferents professionals. El tercer annex presenta la relació dels professionals que han col·laborat en aquest treball.

1. ELS SISTEMES D'EXTRACCIÓ AUTOMÀTICA DE TERMINOLOGIA

> *Automatic term recognition is much needed because a simple but coherently built terminology is the starting point of many applications such as human or machine translation, indexing, thesaurus construction, knowledge organisation, etc. and because manual efforts cannot keep up with the rapid growth of technical terms. Methodologically current ATR research is situated within the broad category or corpus-based aprroaches to computational linguistics which has become very popular in the last five or six years. Many ATR studies reveal interesting insights into various aspects of terms.*
>
> [Kageura i Umino, 1996:3]

1.1 Definició

Un sistema d'extracció automàtica de terminologia és un conjunt de programes informàtics que es proposa reconèixer les unitats terminològiques (UT) que apareixen en un corpus textual. Aquest objectiu, no obstant el que es proposa, s'ha de matisar per tal com és un sistema que no atribueix a les unitats detectades un valor terminològic definitiu, sinó que les proposa com a candidates a termes que han de ser validades per un expert humà. Per tant no es tracta d'un sistema automàtic sinó assistit.

Aquests sistemes no han estat concebuts per substituir el professional, sinó per alleugerir-lo d'unes tasques molt concretes del treball terminològic confiant a l'ordinador una part de les tasques del treball terminològic i reservant al terminòleg la part en la qual les seves competències són imprescindibles. Com remarca de Yzaguirre (1996: 71):

> *No se atisba ni remotamente la posibilidad de que la técnica nos lleve a prescindir del terminólogo; bien al contrario, le iremos descargando progresivamente de las partes más mecánicas y menos creativas del trabajo, le permitiremos abarcar más terreno y proporcionaremos nuevas salidas al producto de su esfuerzo.*

Aquesta mateixa idea era ja present l'any 1988 en el disseny de TERMINO:

> *Il ne s'agit pas de mécaniser la recherche terminologique mais de donner au terminologue les moyens de lui faciliter la tâche, de satisfaire à la demande et de produire plus rapidement un travail dont la qualité serait égale ou supérieure au travail effectué manuellement.*
>
> [Perron, 1988:30]

Si tenim en compte aquestes limitacions, podem dir que la utilització de l'etiqueta sistema d'extracció automàtica de terminologia no és suficientment precisa i per això a partir d'ara quan ens referim a aquests sistemes ho farem amb la denominació *sistema d'extracció automàtica de candidats a terme* (SEACAT).

1.2 Funcions

La funció principal dels SEACAT és, doncs, l'**automatització** de la **fase de buidatge** de qualsevol activitat terminològica. Aquesta fase consisteix a seleccionar totes les unitats d'un text amb significat especialitzat. La recerca de les UT pertinents per a un treball determinat és una tasca que requereix molt de temps i sistematicitat en l'aplicació de criteris; és, de fet, una de les fases més feixugues i llargues —sobretot quan es manipulen volums d'informació importants— que té el risc de convertir-se en poc sistemàtica i, consegüentment, ineficaç. A més, moltes vegades, i segons el tema i el grau d'especialització dels textos, la selecció de les UT es converteix en un treball d'equip en el sentit que requereix el treball coordinat de, com a mínim, un especialista en llenguatge i un expert en el domini que es tracti; característica que pot alentir i diversificar més la

tasca[1]. Amb l'ús dels SEACAT, doncs, el treball terminològic guanya **rapidesa** i **sistematicitat.**

1.3 Metodologies

S'han fet diferents classificacions dels SEACAT[2] a partir de la metodologia que aquests sistemes utilitzen per reconèixer les unitats terminològiques [Kageura i Umino, 1996], [Drouin, 1997], [Estopà, Vivaldi, Cabré, 1998]; tradicionalment s'ha distingit entre:

a. Sistemes que utilitzen només mètodes basats en coneixement estadístic.
b. Sistemes que utilitzen només mètodes basats en coneixement lingüístic.
c. Sistemes que utilitzen mètodes basats en coneixement estadístic i en coneixement lingüístic alhora.

[1]Aquests arguments eren el mateixos que van motivar el disseny de la primera eina informàtica d'ajuda al buidatge terminològic: *"Parmi toutes les tâches qui composent le travail terminologique, il en est une qui exige du terminologue beaucoup de son temps, de son attention et de ses efforts; il s'agit du dépouillement, cette étape au cours de laquelle le terminologue relève, dans le corpus parfois volumineux qu'il a constitué, les termes du domaine à l'étude et les données s'y rapportant dont il aura besoin lors de l'étape de l'analyse terminologique. Bien que, lorsqu'il dépouille les textes qu'il a rassemblés, le terminologue doive constamment user de ses connaissances et de son expérience, particulièrement lorsqu'il doit relever et découper les termes qu'il juge pertinents à son domaine de recherche, il n'en demeure pas moins que cette partie du travail terminologique, si essentielle soit-elle, présente un aspect fastidieux: que de textes il lui faut parcourir pour dépister les termes, trouver des contextes; que de fiches il doit remplir sur lesquelles doivent être consignées un nombre important de données."* [Perron, 1988: 24].

[2]La indexació de textos —que consisteix a identificar les paraules o expressions d'un document que serveixen per representar-lo de manera condensada— és una tasca que manté punts de contacte amb la detecció d'unitats terminològiques a partir de textos pel fet que aquestes unitats són també nusos de coneixement privilegiat. Per aquesta raó, alguns dels procediments o sistemes que mencionem procedeixen d'aquest camp. Spyns (1996) presenta un panorama molt complet de programes informàtics d'indexació de textos en el camp biomèdic.

1.3.1 Sistemes basats en coneixement estadístic

En general, els mètodes estadístics reconeixen les unitats terminològiques a partir de la seva freqüència d'aparició en un corpus marcat temàticament. Per conèixer el grau de concurrència entre els components d'un candidat a terme, es basen en càlculs estadístics que oscil·len des de simples freqüències fins a mesures molt més complexes.

Els mètodes que es basen exclusivament en coneixement estadístic [Salton i Buckley, 1988], [Evans i Lefferts, 1995], a diferència dels lingüístics, condicionen la llargada dels corpus que han d'utilitzar de forma que, si el corpus d'aplicació és petit, es genera molt de **silenci** ¾ nombre de termes no reconeguts del total de termes presents en un text; però tot i que el corpus estigui constituï t per milions d'ocurrències sempre hi ha un percentatge de paraules que per la seva baixa freqüència d'ús en els textos seleccionats no es poden recuperar. Els mètodes estadístics generen també molt de **soroll** ¾ nombre de candidats a terme sense valor terminològic¾, perquè en els textos especialitzats, a part de paraules gramaticals, trobem tant paraules especialitzades com paraules usades amb un significat no especialitzat que formen part de la llengua general. Moltes d'aquestes paraules sense valor especialitzat pareixen en els textos amb una freqüència d'ús alta (*cosa, causa, conseqüència, hipòtesi, tema, estudi, element, formació, raó, distinció, manera, fixació, diferència, procés*, etc.).

Els mètodes estadístics, a diferència dels lingüístics, no permeten arribar a generalitzacions que contribueixin a explicar fenòmens del llenguatge natural, les seves estratègies són independents del llenguatge. En canvi, entre els mètodes basats en coneixement lingüístic i les ciències del

llenguatge s'estableix un circuit de retroalimentació recursiva que permet avançar en el camp de l'explicació teòrica de fenòmens lingüístics[3].

1.3.2 Sistemes basats en coneixement lingüístic

Per reconèixer els termes, els sistemes basats en coneixement lingüístic utilitzen diferents recursos que contenen dades lingüístiques. Aquestes informacions són normalment:

lexicogràfiques:

- diccionaris de termes
- diccionaris de paraules auxiliars

morfològiques:

- patrons d'estructura interna del mot

morfosintàctiques:

- patrons morfosintàctics
- elements que marquen fronteres exteriors de la UT
- funcions sintàctiques

i més rarament,

semàntiques:

- classificacions semàntiques

pragmàtiques:

- representacions tipogràfiques
- informacions de disposició del terme en el text, etc.

Així, per exemple, hi ha extractors que utilitzen patrons morfològics de sintagmes terminològics, com el sistema TERMS [Justeson i Katz, 1995]; n'hi ha que funcionen a partir d'un diccionari de paraules auxiliars, d'un

[3] Hauríem de parlar d'una retroalimentació parcial perquè actualment els mètodes lingüístics, com ja hem avançat, parteixen només de la **part formal** d'una unitat lèxica i

diccionari terminològic i de regles que s'apliquen sobre termes ja reconeguts, com el FASTR [Jacquemin, 1994]; d'altres usen patrons morfosintàctics que assenyalen les fronteres externes del terme, és a dir, coneixement lingüístic del *no-terme*, com el LEXTER [Bourigault, 1994], aquest tipus de fronteres poden complementar-se amb informació tipogràfica sobre l'aparició dels mots, com fa DROUIN [Drouin, 1997]; un altre grup de sistemes es basen en patrons morfosintàctics del terme, com TERMINO [David i Plante, 1991] i d'altres es basen en una anàlisi sintàctica detallada del sintagma nominal, com el NODALIDA [Arppe, 1995]; etc.

El tipus de coneixement utilitzat fa que els sistemes d'aquest tipus siguin, majoritàriament, aplicables només a una llengua i, per tant, la utilització en textos d'una llengua diferent necessita d'un estudi lingüístic previ i, probablement, d'un nou disseny d'algun dels mòduls del sistema.

Com ja hem dit anteriorment, un dels problemes principals dels sistemes que treballen només amb dades formals (morfològiques, morfosintàctiques, sintàctiques i/o lèxiques) és la gran quantitat de soroll que generen (entre el 55% i el 75%). En efecte, no totes les paraules proposades com a UTP pels sistemes ho són, sinó que sovint els mateixos patrons corresponen també a unitats lèxiques i fraseològiques amb un ús no especialitzat; altres vegades, responen a unitats d'ús especialitzat, però no terminològiques, com les unitats fraseològiques especialitzades (UFE) o les combinacions molt recurrents; i, fins i tot, altres vegades, simplement són segments discursius, sense interès especialitzat.

Una idea que actualment és compartida per tothom és que l'única manera de reconèixer i delimitar exhaustivament les UT d'un text especialitzat és incorporant en els programes d'extracció algun tipus de **coneixement**

---

ignoren la complexitat de la unitat terminològica.

semàntic. En aquest sentit i simplificant les possibilitats reals, els extractors poden fer servir dues estratègies per adjuntar informació semàntica al lèxic.

La primera consisteix a utilitzar categories semàntiques d'una font lèxica externa al corpus textual de treball. Per exemple, WordNet[4] o AlethDic[5] són dos sistemes de classificació léxica, que organitzen el lèxic a partir del significat de les paraules i no a partir del seu significant, que podrien integrar-se en un extractor.

La segona via consisteix a extreure les categories semàntiques de les paraules del mateix corpus de treball a través d'elements cotextuals que fan referència a la combinació semanticosintàctica de les paraules.

L'extractor de terminologia dissenyat per Naulleau (1998) és un exemple del primer enfocament:

> *Contre nos convictions et en raison des problèmes de faisabilité, nous n'adoptons pas le point de vue contextualiste qui aurait pu s'appuyer sur une approche distributionnelle pour définir des catégories sémantiques. Nous catégories sémantiques proviendront donc d'une source extérieure au corpus. Nous avons récupéré les étiquettes sémantiques du lexique d'AlethDic, puis projeté celles-ci sur un nouveau jeu d'étiquettes, plus étroit.*
>
> [Naulleau, 1998: 70]

En el segon enfocament, se situa el model teòric d'interpretació del significat de les seqüències NA proposat per Fabre (1996):

> *Notre principale contribution concerne l'élaboration du modèle d'interprétation. Nous avons proposé une mise à plat des règles compositionnelles utilisables pour le calcul sémantique et défini une extension du principe d'attachement d'informations prédicatives au nom. Nous avons montré la nécessité de dépasser la limite de ce qui est linguistique pour s'engager dans la prise en compte de données pragmatiques. Nous avons de fait établi des principes d'interprétation pour*

[4][Millner, 1990].
[5][Naulleau, 1998].

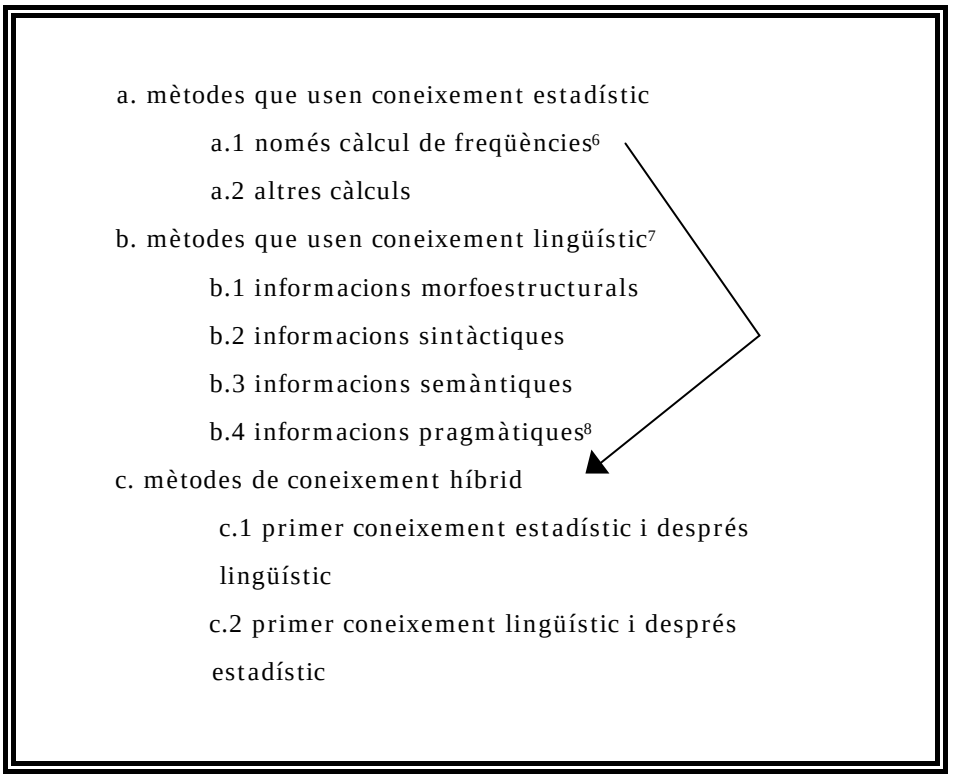
> *les séquences sans constituant prédicatifs, en nous basant sur certains éléments du lexique génératif de J. Pustejovsky.*
>
> [Fabre, 1996:154]

1.3.3 Sistemes basats en coneixement híbrid

Finalment, tenim els sistemes que apliquen alhora coneixement estadístic i coneixement lingüístic. En aquests sistemes —anomenats híbrids—, l'ordre d'aplicació dels tipus de coneixement és decisiu ja que condiciona els resultats. Els mètodes que apliquen, primer coneixement estadístic i després coneixement lingüístic, presenten els mateixos problemes de silenci que els sistemes basats exclusivament en coneixement estadístic, DROUIN [Drouin, 1997]. En canvi, si l'estadística es fa servir només com a mecanisme complementari de la lingüística (ACABIT [Daille, 1995] i CLARIT [Evans i Zhai, 1996]), els resultats finals poden ser més bons per tal com l'estadística pot ajudar en un moment del procés de detecció a reafirmar la condició de terme d'una unitat lingüística, o bé a rebutjar la condició de terme d'una unitat lingüística.

Avui dia, les tècniques estadístiques proporcionen dades en relació amb l'ús dels mots i, en certa manera, podem dir que supleixen aspectes de la competència pragmàtica que tot especialista té sobre els termes del seu domini.

Si tenim en compte no només el tipus de coneixement que els extractors utilitzen, sinó també la diversitat d'informació utilitzada i l'ordre d'utilització, obtindrem una proposta de classificació dels sistemes d'extracció automàtica de termes més afinada que la que proposàvem a l'inici de l'apartat:



1.4 Dominis d'aplicació

Una eina que realitzi el buidatge terminològic de manera automàtica, encara que sigui parcial, és un component central d'un sistema d'ajuda a qualsevol tipus d'aplicació terminològica. En efecte, moltes disciplines que tenen com a objectiu aplicat el tractament de les llengües des de diferents punts de vista —la terminologia, la traducció, la lingüística,

[6]Els termes tenen un vessant gramatical i un vessant pragmàtic i és justament en els aspectes pragmàtics on trobem la major part de les peculiaritats del lèxic especialitzat [Cabré, 1992]. La freqüència d'aparició de determinades estructures formals, categories gramaticals o esquemes semàntics en un text ajuden a conèixer l'ús que els parlants fan de les paraules. La lexicologia des dels seus inicis s'ha valgut de la freqüència d'ús —encara que aquesta s'hagi calculat de manera inexacta o intuï tiva— per explicar i descriure les unitats lèxiques. En aquest sentit, la freqüència està propera a la pragmàtica, a l'actuació i, per tant, a la lingüística entesa en una concepció àmplia.

[7] Els diferents nivells d'informació lingüística no són excloents, sinó que alguns sistemes en combinen més d'un, com veurem en el proper apartat.

[8]Entenem per informacions pragmàtiques totes les informacions tipogràfiques i de disposició del terme en el text.

l'ensenyament, la documentació, les especialitzacions— s'han adonat de la necessitat de comptar amb extractors de terminologia que funcionin amb coneixement lingüístic.

1.4.1 Traducció especialitzada

En traducció especialitzada, tant si es tracta d'un procés manual com automàtic, el buidatge terminològic d'un text és indispensable per realitzar la traducció o interpretació d'una comunicació especialitzada, amb un nivell òptim de qualitat[9]. En el cas, per exemple, de la traducció d'un volum important de documents feta per un equip de traductors, els SEACAT s'utilitzen per recollir els termes d'aquell domini i establir-ne glossaris; d'aquesta manera, l'equip es pot distribuir els documents de traducció amb la seguretat que la terminologia utilitzada per tots els membres de l'equip és homogènia. Segons Melby (1990), una bona traducció hauria d'anar acompanyada sempre d'un fitxer terminològic bilingüe.

El lloc de treball del traductor tecnicocientífic ha de comptar amb eines de tractament de la terminologia i sobretot amb bases de dades terminològiques plurilingües, en la constitució de les quals els SEACAT poden ser molt útils. Termight, [Dagan i Church, 1994], és un SEACAT concebut com una eina d'ajuda als traductors.

1.4.2 Terminografia

La creació de bases de dades terminològiques, de diccionaris, glossaris, vocabularis o qualsevol tipus de recull lèxic especialitzat requereix complir

una fase en què es localitzen i se segmenten els termes dels textos. En aquesta etapa, es realitza un buidatge sistemàtic d'un corpus textual del domini sobre el qual es vol fer el treball terminogràfic i, a continuació, es decideix quines són les unitats pertinents per al treball que es vol fer. En el treball terminogràfic, la finalitat de la fase de buidatge és recollir el material per confeccionar la nomenclatura d'una base de dades terminològica o d'un diccionari especialitzat. Els SEACAT, com ja hem vist, són una ajuda a aquesta tasca de buidatge i, actualment, s'estan convertint en eines auxiliars del treball terminogràfic. TERMINO [Perron, 1989] i HEID [Heid i al., 1996] són dos exemples d'aquest tipus d'eina creats amb un objectiu terminogràfic.

1.4.3 Documentació

Els SEACAT són també de gran utilitat en el disseny, elaboració i actualització de sistemes d'indexació automàtica, en la confecció automàtica de tesaurus, en l'elaboració de sistemes hipertextuals, en sistemes de categorització de documents especialitzats, en sistemes de recuperació d'informació basats en xarxes de descriptors, etc.

Cert és que les terminologies i els tesaurus són productes diferents: mentre que les terminologies són recopilacions dels termes d'un domini determinat, els tesaurus estan constituï ts de descriptors —una bona part d'ells termes— que serveixen per indexar els documents; tot i així, molts autors han remarcat la seves afinitats basant-se en el fet que la majoria de descriptors pertinents, per exemple per a la indexació de documents, són també UT[10]. Aquest fet justifica que molts SEACAT tinguin com a objectiu

[9] *"L'analyse du processus de la traduction technique a montré que le traducteur passe une grande partie de son temps (40% ou davantage) à résoudre des problèmes de terminologie."* [Felber, 1987: 31].

[10] Per a aquesta qüestió vegeu un article de Larivière (1989) on proposa el concepte de *tesaurus terminològic*.

primer la indexació de textos, com FASTR [Jacquemin, 1996] o CLARIT [Evans i Zhai, 1996], per citar dos exemples. En el camp de la biomedicina hi ha molts projectes que tenen com a objectiu la recuperació d'informació i, per tant, la indexació i la descodificació de documents són dues aplicacions que deriven d'aquest objectiu. Spyns (1996) presenta un panorama general dels projectes internacionals que en el domini de la medicina tenen aquest objectiu documental.

1.4.4 Gestió del coneixement

Actualment, els bancs de coneixement terminològic comencen a reemplaçar els bancs terminològics. Des del primer banc de dades DICAUTOM creat l'any 1963 (que va esdevenir l'any 1973 EURODICATOM), s'han acumulat una sèrie de crítiques que han mostrat la ineficàcia de les bases de dades terminològiques. Lerat (1995) distingeix cinc raons d'insatisfacció dels bancs de dades: la concepció, la finalitat, l'alimentació, l'actualització i el grau d'exactitud; Otman (1997) afegeix a aquestes insuficiències la manca d'una estructura multirelacional.

Els bancs de coneixement representen, en paraules de Badia (1996: 72), *"una estructura simbòlica d'una sèrie de fets sobre el món"*. Aquests nous bancs de saber es basen sempre en ontologies que relacionen els conceptes i els termes. Consegüentment, les bases de coneixement terminològic es poden entendre com *"une amélioration d'une Base de Données Terminologiques classique enrichie par des relations conceptuelles."* [Condamines, 1995: 37]. Per a Otman (1997: 244), una base de coneixement terminològic és a la vegada *"une banque de terminologie conceptuellement et sémantiquement structurée et une base de connaissances"*.

La base de coneixements terminològics modelada per l'equip de Condamines (1995), per posar un exemple, inclou dades conceptuals del domini i dades lingüístiques dels termes, i utilitza el sistema d'extracció automàtica de terminologia LEXTER en la primera fase del sistema per detectar els conceptes, a través del reconeixement de les UT que els representen, que configuraran una estructura.

1.4.5 Adquisició i processament del coneixement especialitzat

La majoria d'aplicacions informàtiques que tenen com a objectiu l'automatització de l'adquisició o el processament de coneixement especialitzat a partir de textos necessiten diferents tipus d'eines, que seguint Bourigault (1994) es poden classificar en tres grans categories segons la seva finalitat:

1. **Eines de transferència** que permeten transvasar, directament, els coneixements inscrits dels textos a un sistema informàtic, com ara una base de dades, una base de coneixements, un sistema d'autoaprenentatge de coneixements especialitzats o un sistema expert. Són exemples d'aquest tipus d'eines, entre d'altres, les creades per Schmidt i Wetter (1990) o per Copeck i al. (1992).

2. **Editors hipertextuals** que permeten editar un sistema de coneixements de manera molt flexible: és el cas, per exemple, de l'editor dissenyat per Motta i al. (1991).

3. **Eines d'anàlisi de textos**[11] que faciliten l'etapa inicial del disseny d'un sistema d'adquisició automàtica de coneixements a partir de textos. Aquest tipus d'eines no només són útils per al

[11] En paraules de Bourigault (1994) *outils de dépoulliment*.

disseny de sistemes d'adquisició de coneixement, sinó també per a totes les anàlisis realitzades en la **lingüística de corpus**. Seguint Bourigault, un SEACAT es pot concebre com una eina d'ajuda a l'anàlisi de textos: aquesta és la finalitat de LEXTER [Bourigault, 1994], un sistema d'extracció de terminologia concebut per Bourigault explícitament per a aquesta funció:

> *L'acquisition des connaissance est une activité de construction de modèles, au cours de laquelle le cogniticien effectue un travail d'interprétation "de haut niveau" hautement subjectif; les outils d'analyse de textes susceptibles de l'assister dans cette activité doivent se charger de tâches de "bas niveau", et exploiter les qualités de systématicité et de rapidité de l'ordinateur dans le traitement en masse de données textuelles. Nous pensons que, entre les deux extrémités de l'axe que nous venons de décrire , il y a de la place pour les outils que ne viennent pas se substituer au cogniticien pour l'interprétation des textes, mais ont pour rôle de lui apporter une aide conséquente en effectuant des traitement, en masse et de bas niveau, de dépouillement de textes dans le cas d'une documentation de taille important. Nous pensons que LEXTER constitue un outil de ce type.*
>
> [Bourigault, 1994: 228]

1.4.6 Lingüística computacional

La recerca en llenguatge natural es proposa implementar un model que doni compte de com els éssers humans adquireixen, estructuren i processen el coneixement. Per tant, tracta d'aconseguir que un ordinador actuï intel·ligentment, que observi, raoni i prengui decisions sobre fets concrets, que interactuï amb persones, etc. Però el llenguatge natural és molt complex i variat, i aquesta és la raó per la qual la lingüística computacional sovint ha limitat el terreny del tractament automàtic del llenguatge a un domini especialitzat temàticament.

Per exemple, reconèixer, delimitar i conèixer la pertinència de totes les UT d'un text són algunes de les tasques que l'ésser humà pot dur a terme, però no l'ordinador, tot i que el buidatge automàtic és un dels components

d'algunes aplicacions en lingüística computacional. Per posar un cas concret, per realitzar un sistema informàtic de comprensió o de generació de textos es necessita construir abans una base de coneixements terminològics amb informacions semàntiques, sintàctiques i pragmàtiques sobre els termes; i per poder elaborar una base de coneixements terminològics, com hem comentat anteriorment, és important disposar d'una eina d'extracció de terminologia, que és la que proporcionarà la matèria primera a partir d'un corpus textual informatitzat.

1.5 Estructura i funcionament: estat de la qüestió

> *Les termes, et plus particulièrement les termes complexes formés de plusieurs mots, sont des entrées lexicales ayant un comportement syntaxique et sémantique spécifique. Leur reconnaissance dans les textes ne se limite pas à la recherche de quelques patrons syntaxique spéciaux, ni à la comparaison de chaînes de mots avec des listes de termes contrôlés.*
>
> [Jacquemin, 1997:3]

En aquest apartat, i seguint un estudi més complet d'Estopà, Vivaldi i Cabré (1998), exposem una síntesi de les característiques internes que presenten els principals SEACAT amb la finalitat de dibuixar un panorama ampli de l'estat de la qüestió que permeti veure els seus avantatges i limitacions. D'aquesta manera, disposarem d'elements per millorar en algun aspecte aquests sistemes.

Hem partit de divuit extractors diferents elaborats des de 1989 (any del primer sistema semiautomàtic de buidatge terminològic) fins a 1997: ACABIT[12], ANA[13], ATELIER/FX[14], AUTOLEX[15], BLANK[16], CLARIT[17],

[12][Daille, 1994].
[13][Enguehard i Pantera, 1994].
[14]URL: http://www.ling.uqam.ca/Ato/FX/AtelierFX.html
[15][Planas, 1994].
[16][Blank, 1995].
[17][Evans i Zhai, 1996].

DROUIN[18], FASTR[19], HEID[20], LEXTER[21], NEURAL[22], NODALIDA-95[23], SBIC[24], TERMIGHT[25], TERMINO[26], TERMS[27], STELLA[28] i NAULLEAU[29].

L'estudi d'aquests sistemes se centra en els sis paràmetres següents:

- els nivells d'informació d'entrada
- les estratègies de reconeixement de candidats a terme
- les estratègies de filtratge de termes
- les estratègies d'alimentació de coneixement
- la interacció del sistema amb l'usuari
- els resultats obtinguts.

Aquests paràmetres permeten conèixer les principals estratègies de funcionament que actualment usen els SEACAT i saber quines són les seves limitacions.

1.5.1 Nivells d'informació d'entrada

El nivell d'informació d'entrada fa referència al tipus d'informació prèvia que requereix cada sistema per poder ser aplicat, és a dir, la matèria primera. La majoria d'extractors utilitza informació lingüística en algun moment del procés. Aquesta informació pot ser variada. Alguns sistemes parteixen d'una llista de paraules auxiliars, d'altres de filtres per a

[18][Drouin, 1996].
[19][Jacquemin, 1996].
[20][Heid i al., 1996].
[21]Bourigault, 1994].
[22][Frantzi i Ananiadou, 1995].
[23][Arppe, 1995].
[24][Bordoni, L.; Anzaldi, C., 1996].
[25][Dagan i Church, 1994].
[26][David i Plante, 1991].
[27][Justeson i Katz, 1995].
[28][Jacquin i Liscouet, 1996].
[29][Naulleau, 1998].

categories gramaticals, i gairebé tots utilitzen la combinació d'un analitzador morfològic amb un desambiguador[30].

La taula següent reflecteix la tria d'informació de cada sistema:

Sistema		Nivell d'informació d'entrada				
	nom	llistes de paraules	anàlisi morfològica	desambiguador	corpus d'aprenentatge	filtre de categories
1	ACABIT		X	X		
2	ANA	X				
3	ATELIER/FX		X	X		
4	AUTOLEX	X				
5	BLANK		X	X		
6	CLARIT		X			
7	DROUIN		X	X		
8	FASTR	X	X	X		
9	HEID		X	X		
10	LEXTER		X	X	X	
11	NAULLEAU	X	X	X		
12	NEURAL		X	X		
13	NODALIDA-95		X	X		
14	SBIC	X				
15	TERMIGHT		X	X		
16	TERMINO		X	X		
17	TERMS		X			X
18	STELLA		X			

1.5.2 Estratègies de reconeixement de candidats a terme

El reconeixement i la delimitació de les UT són dues de les fases més complexes d'aquest tipus d'aplicació; els programes analitzats es valen d'estratègies diferents per recuperar els termes, encara que cap de les estratègies és per si mateixa del tot satisfactòria:

- elements que actuen de frontera de mot
- patrons estructurals

[30]Els autors d'aquests sistemes coincideixen a considerar el desambiguador com una de les fonts d'error que fa augmentar l'índex de silenci.

- analitzadors sintàctics parcials
- elements de disposició de les paraules en el text
- elements tipogràfics
- llistes d'UT
- perfils d'aprenentatge[31]
- classificacions semàntiques del lèxic[32].

El quadre que es presenta a continuació resumeix les diferents opcions adoptades per cada sistema analitzat:

Sistema		Delimitació de termes				Desambiguació d'estructures	
	nom	fronteres	patrons	*parser*[33]	altres	aprenentatge	altres
1	ACABIT		X				-
2	ANA				X		-
3	ATELIER/FX				X		-
4	AUTOLEX	X					-
5	BLANK	X	X				-
6	CLARIT			X			estadística
7	DROUIN	X					-
8	FASTR		X	X	X		-
9	HEID		X				-
10	LEXTER	X				X	
11	NAULLEAU		X	X		X	
12	NEURAL		X				-
13	NODALIDA-95				X		-
14	SBIC	X					manual
15	TERMIGHT		X				-
16	TERMINO			X	X		-
17	TERMS		X				-
18	STELLA			X	X		

[31]Naulleau parteix d'uns perfils formats per UT pertinents per a un usuari determinat. De cada UT pertinent, el programa n'extreu informacions morfològica, sintàctica i semàntica.

[32]En el moment en què es va realitzar l'anàlisi [Estopà, Vivaldi, Cabré, 1998] no hi havia encara cap programa que usés informació semàntica. En l'actualitat, el projecte proposat com a tesi doctoral per Naulleau (1998) utilitza etiquetes semàntiques extretes d'una classificació semàntica del lèxic, AlethDic, tant en aquesta fase inicial com en les posteriors. Aquestes etiquetes permeten prioritzar certs esquemes d'interpretació i, per tant, afinar més en la selecció final dels candidats a terme.

[33]En aquest context s'ha d'entendre *parser* com un eina d'anàlisi parcial de les frases, mai com una eina que intenti donar una anàlisi completa i única de cada frase.

1.5.3 Estratègies de filtratge de termes

L'última fase —explícita o implícita— abans de presentar el conjunt final de candidats a terme dels extractors de terminologia és el filtratge de candidats a terme. El quadre següent concreta els tipus de filtre que fan servir els programes per intentar reduir el soroll inicial:

Sistema		Filtratge de termes					
	nom	manual	freqüència[34]	lingüístic	estadístic + lingüístic	lingüístic + estadístic	termes de referència
1	ACABIT					X	
2	ANA						X
3	ATELIER/FX			?			
4	AUTOLEX	X					
5	BLANK		X	X			
6	CLARIT				X	X	
7	DROUIN[35]				X		
8	FASTR						X
9	HEID			X			
10	LEXTER			X			
11	NAULLEAU			X			
12	NEURAL					X	
13	NODALIDA-95			X			
14	SBIC	X					
15	TERMIGHT		X	X			
16	TERMINO		X	X			
17	TERMS		X	X			
18	STELLA						X

1.5.4 Estratègies d'adquisició

Gairebé cap dels SEACAT analitzats aprofita els resultats obtinguts en l'aplicació del programa, és a dir, els sistemes no incorporen tècniques de *feedback* i, per tant, cada vegada que s'aplica de nou el programa es

[34]Hem considerat la tècnica de filtratge de termes mitjançant la freqüència com un cas particular, a mig camí entre els mètodes basats en coneixement lingüístic i els basats en coneixement extralingüístic.

parteix de zero. Només dos sistemes, FASTR i ANA, opten per una estratègia incremental: a partir d'un conjunt de termes ja reconeguts el sistema en reconeix de nous. En ambdós casos, però, malgrat que el mètode de reconeixement és recursiu, els termes identificats no es validen abans de fer-los servir en el cicle següent i, conseqüentment, sorgeix el problema que un segment considerat erròniament terminològic doni lloc a termes no vàlids en cicles posteriors.

1.5.5 Interacció del sistema amb l'usuari

Com ja hem comentat anteriorment, al final de l'aplicació d'un SEACAT s'arriba a una llista de segments que han de ser validats manualment per un usuari que posseeix una competència cognitiva i pragmàtica sobre el tema especialitzat de què tracta el text processat.

En alguns casos els resultats es presenten concisament —per exemple, SBIC—, mentre que en altres es presenten els candidats a terme en context, fet que facilita la tasca de revisió: a través de navegació hipertextual —LEXTER, ATELIER/FX—, de finestres amb múltiples contextos per a cada candidat —NODALIDA, HEID, TERMIGHT, TERMINO—, de xarxes semàntiques a partir dels termes detectats —ANA, STELLA—, relacionant totes les paraules d'un text —ATELIER/FX, FASTR—, o creant una xarxa terminològica mitjançant la descomposició dels termes en nucli i expansió —LEXTER.

1.5.6 Resultats obtinguts

Els resultats obtinguts se solen valorar en relació amb dos paràmetres:

[35]Aquest sistema incorpora també una etapa de postprocessament.

- **el silenci**
- **el soroll.**

En el primer cas, es valora el nombre de les paraules que en el text tenen valor terminològic i que el sistema no ha presentat com a candidates a terme. En el segon cas, es mesura, del total de candidats a terme presentats pel sistema, el percentatge d'unitats rebutjades per l'usuari perquè en el text no tenen un valor terminològic[36]. En general, els autors dels sistemes no faciliten de manera explícita i objectiva dades sobre l'èxit de l'aplicació i, en aquells casos en què no hem pogut experimentar amb el sistema, es fa difícil saber quin és el percentatge de soroll i de silenci de cada sistema perquè no hi ha massa dades públiques en aquest sentit.

Com es pot deduir de la taula següent, la majoria de sistemes s'apliquen sobre textos en una sola llengua —especialment en anglès o en francès— i sobre corpus d'un àmbit o d'un subàmbit especialitzat:

[36]Els conceptes de *silenci* i *recall,* i els de *soroll* i *precisió* donen la mateixa informació, respectivament, tot i que tenen en compte punts de vista diferents. Així, per exemple, un sistema amb un *silenci* del 25% li correspondria una xifra de *recall* del 75%.

	Sistema	Corpus de prova		
	nom	àrea	llengua	dimen. [K par.]
1	ACABIT[37]	Telecomunicacions	francès	200 800
2	ANA	Bricolatge	francès anglès	120 25
3	ATELIER/FX	Medicina	francès	?
4	AUTOLEX	?	?	?
5	BLANK	Jurídic (patents)	alemany	12.000
6	CLARIT[38]	Notícies	anglès	240 Mbytes
7	DROUIN	Geomètrica	francès	?
8	FASTR	Medicina	francès	1.560
9	HEID	Enginyeria	alemany	35
10	LEXTER	Enginyeria	francès	3.250
11	NAULLEAU			
12	NEURAL	Medicina (oftalmologia)	anglès	55
13	NODALIDA-95	Cosmologia Enginyeria Ind. de l'automòbil Premsa	anglès	20
14	SBIC	Medi ambient	italià	?
15	TERMIGHT	Informàtica	anglès	?
16	TERMINO	Medicina	francès	?
17	TERMS	Metal·lúrgia Eng. espacial Eng. nuclear estadística Semàntica Cromatografia	anglès	? ? ? 2,3 6,3 14,9
18	STELLA	Documents d'Internet	anglès francès	?

1.6 Sistemes de base lingüística

En l'apartat anterior hem analitzat diferents SEACAT a partir del seu funcionament. En aquesta secció ens centrarem exclusivament en alguns sistemes que utilitzen coneixement lingüístic. Concretament, descriurem

[37]Daille i al. (1996) indiquen que una aplicació experimental d'ACABIT sobre un manual de telecomunicacions de 200.000 mots (*The satellite communication handbook*) ha donat una taxa d'encert al voltant del 85%.

[38]El sistema s'ha provat molt intensivament pel que fa a l'eficiència de la indexació, però no en relació amb la qualitat del termes extrets.

de manera detallada les característiques de funcionament de sis SEACAT representatius de les metodologies de base lingüística:

1. TERMS. Es basa en esquemes morfosintàctics de les UT.
2. LEXTER. Es basa en coneixement morfosintàctic sobre el no-terme.
3. NODALIDA. Es basa en coneixement morfosintàctic del terme, però usa un detector sintàctic de frases nominals.
4. FASTR. Es basa en diccionaris lèxic i terminològic i regles morsintàctiques.
5. NAULLEAU. Es basa en coneixements morfològic, sintàctic i semàntic.
6. ACABIT. Combina coneixement lingüístic i coneixement estadístic.

Tots aquests sistemes es proposen només l'extracció dels segments complexos, candidats a **unitats terminològiques polilèxiques (UTP)**, i no d'unitats terminològiques monolèxiques.

1.6.1 TERMS

Autors:	Justeson, J.; Katz, S.
Data:	1995
Objectiu:	Eina d'extracció de candidats a terme
Llengua de treball:	Anglès

1.6.1.1 Sinopsi

Els autors de TERMS parteixen de les idees següents en relació amb els termes:

- Els termes tècnics estan formats gairebé sempre per sintagmes nominals (entre el 92,5 i el 99% d'una mostra aleatòria de 800 termes són sintagmes nominals).

- els sintagmes nominals terminològics estan formats bàsicament per noms i adjectius (97%) i algunes preposicions (3%) sempre entre dos sintagmes nominals.
- La longitud mitjana dels sintagmes nominals terminològics és de 1,91 paraules.
- Els sintagmes nominals dels textos tècnics són gairebé exclusivament terminològics.
- Els sintagmes nominals terminològics es diferencien dels no terminològics pel fet que els tipus de modificadors dels primers són molt més reduï ts que els que poden aparèixer en els segons.
- Un sintagma nominal terminològic, normalment, es repeteix en forma idèntica en un mateix document, ja que la sola omissió d'un modificador podria fer que l'entitat referenciada canviés i, en canvi, els sintagmes nominals no terminològiques presenten molta més variació perquè no estan fixats.

TERMS utilitza l'algorisme següent per buscar cadenes amb una freqüència igual o major a dos que responguin a un sintagma nominal terminològic[39]:

((A | N)+ | ((A | N)*(N P)?)(A | N)*N[40]

Els candidats a terme de longitud 2 (patrons: AN i NA) i longitud 3 (patrons: AAN, ANN, NAN, NNN i NPN) són, amb diferència, els més usuals. Alguns exemples de candidats a terme amb aquestes estructures diferents són:

AN: *linear function, lexical ambiguity*

[39] Aquesta expressió regular cobreix el 97% del termes presents en diccionaris terminològics (99% si s'accepten les preposicions).
[40]Els autors fan servir les abreviacions: A: adjectiu. N: nom. P: preposició.

NN: *regression coeficcients, word sense*

AAN: *Gaussian random variable, lexical conceptual paradigm*

ANN: *cumulative distribution frequency, lexical ambiguity resolution*

NAN: *mean squared error, domain independent set*

NNN: *class probability function, text analysis system*

NPN: *degree of freedom, energy of adsorption.*

Aquest algorisme es proposa d'oferir una bona cobertura[41] de la terminologia habitual dels manuals tècnics i alhora una alta qualitat[42] en l'extracció.

TERMS utilitza a més dels patrons estructurals restriccions de freqüència i de certes unitats gramaticals. En particular, TERMS només considera la preposició *de*, perquè, segons els autors, si en considerés d'altres la qualitat dels resultats baixaria molt, tot i que la cobertura augmentaria.

La implementació dels patrons gramaticals també afecta la relació qualitat/cobertura de l'extractor. TERMS primer analitza i lematitza cada paraula del text i, a continuació, identifica les seqüències que satisfan els patrons. Si una paraula no s'identifica com a nom, adjectiu o preposició, es descarta. De cada paraula es conserva només les lectures com a nom, adjectiu o preposició en aquest ordre de prioritat. La cadena es rebutja si hi ha més d'una paraula que s'identifica com a preposició, o bé no satisfà el patró.

Segons els autors, el filtratge utilitzat per TERMS té una cobertura com a mínim tan bona com la que s'obtindria amb un analitzador morfològic convencional, malgrat que la qualitat és inferior (per exemple *fixed*

[41] Proporció entre els termes vàlids que l'algorisme ha extret i el total de termes del text.
[42] Proporció entre els termes vàlids que l'algorisme ha extret i el total de termes proposats.

s'identifica només com a adjectiu *bug fixed*, però també pot ser verb *fixed disk drive*). En contrapartida, la velocitat de processament és notablement superior.

El sistema s'ha provat en diferents àrees (metal·lúrgia, enginyeria espacial i energia nuclear) i s'utilitza en centres de traducció d'IBM. Segons els autors, la cobertura de TERMS és del 71% i la qualitat dels seus resultats entre el 77% i el 96%.

1.6.1.2 Valoració

Aquest sistema és un exemple prototípic de l'aproximació lingüística clàssica basada en característiques estructurals de les UT. Malgrat que els autors han realitzat un estudi previ del comportament formal de les UTP —en alguns punts fins i tot amb afirmacions excessivament contundents—, l'algorisme proposat no sembla treure gaire profit d'aquestes anàlisis sobre els patrons formals dels termes.

Aquest sistema està dissenyat per aplicar-se sobre textos anglesos i, per tant, els patrons proposats com a filtre s'haurien d'adaptar si es volgués aplicar a una altra llengua.

1.6.2 LEXTER

Autor:	Bourigault, D.
Data:	1994
Objectiu:	Eina d'extracció de candidats a terme
Llengües de treball:	Una versió en francès i una en anglès

1.6.2.1 Sinopsi

L'origen d'aquest sistema s'ha de buscar en les necessitats de l'empresa EDF (*Electricité de France*) de millorar un sistema d'indexació de textos ja existent. La idea bàsica és la detecció de les fronteres entre les quals s'espera aï llar els sintagmes nominals susceptibles de ser considerats UT. L'anàlisi que realitza LEXTER (Logiciel d'EXtraction de TERminologie) és superficial i cada vegada que s'aplica elabora unes heurístiques dels termes *ad hoc* per tal d'obtenir sintagmes nominals de longitud màxima que considera candidats a terme.

El funcionament d'aquest sistema es basa exclusivament en tècniques lingüístiques i obté uns resultats considerablement bons. LEXTER es basa en el coneixement negatiu de la UT, és a dir, en la identificació d'aquells elements o patrons que no poden formar mai part d'un terme complex.

El programa està organitzat en mòduls principals:

- mòdul d'anàlisi morfològica i desambiguació
- mòdul de delimitació
- mòdul de descomposició
- mòdul d'estructuració
- mòdul de navegació.

1. El **mòdul d'anàlisi morfològica i desambiguació** atribueix a cada una de les paraules del text informació de la categoria gramatical i del lema.

2. En el **mòdul de delimitació** s'elabora una anàlisi sintàctica local per descompondre el text en grups nominals de longitud màxima. Per exemple:

alimentation en eau
pompe d'extraction
alimentation electrique de la pompe de refoulement.

En aquest moment el sistema s'aprofita del coneixement negatiu sobre la composició d'un terme complex. Amb aquest objectiu identifica els patrons que mai no formaran part d'un terme (com, per exemple, els verbs en forma personal, els pronoms, les conjuncions, etc.) i els considera com a "frontera d'un candidat a terme". Alguns d'aquests patrons són simples (pronoms, verbs en forma personal, etc.), mentre que d'altres són patrons complexos (seqüències de preposició + determinant). Un exemple en francès d'aquest últim cas és la seqüència: 'SUR(prep) + LE(article definit)'. L'anàlisi més habitual és considerar que aquesta seqüència marca una frontera entre grups nominals, com a:

> *on raccorde le câble d'alimentation du banc sur le coffret de décharge batterie.*

Però existeix un nombre de casos (al voltant del 10%) en què aquesta seqüència forma part del terme, com ara:

> *action sur le bouton poussair de réarmement*
> *action sur le système d'alimentation de secours.*

Per resoldre aquesta i altres situacions semblants[43], el sistema fa servir un mètode d'aprenentatge endogen dels patrons de subcategorització. Aquesta estratègia consisteix a buscar en el corpus les seqüències *(nom)+Sur+le* amb el context a la dreta. A continuació, s'eliminen els noms que no són productius (que no tenen contextos a la dreta diferents). En la segona passada, les seqüències *sur+le* són considerades fronteres de frase, excepte en els casos que van precedides pels noms productius detectats en la fase d'aprenentatge.

[43] L'autor es refereix a una classe de preposicions que anomena *Sur* i no a la preposició *sur* en concret; i en aquesta classe inclou les preposicions següents: *avec*, *contre*, *dans*, *par*, *pour*, *sans*, *sous*, *vers*, *en* i *sur*.

Per exemplificar aquest funcionament, suposem que en un primer preprocessament el sistema es troba amb seqüències com les següents:

*le **protection contre** le gel est assurée par*
***Protection contre** les grans froids*
*il s'agit de maintenir la teneur en oxygène de cette **eau dans** les limites fixées*
*on procède à l'injection d'**eau dans** les générateurs de vapeu*
*le système permet l'aiguillage desl'**automates sur** le prélèvement effectuè*

La productivitat de cada seqüència d'aquest tipus és:

protection contre	2
eau dans	2
automates sur	1

En el segon prepocessament les seqüències productives (freqüència més gran que 1) no són considerades pel sistema com una frontera de terme, mentre que les seqüències no productives sí que són considerades com un límit extern del candidat a terme. En l'exemple que hem presentat *protection contre* i *eau dans* no actuarà com a frontera mentre que *automates sur* sí que ho serà.

Així, amb aquesta estratègia es detecta positivament un nombre important de noms que d'altra manera s'haurien perdut. Però amb aquesta permeabilitat s'introdueix una quantitat de *brutícia* considerable, estimada per l'autor entre el 10% i el 50 %.

3. En el **mòdul de descomposició** s'analitza els grups nominals i se'n divideixen els constituents en nucli i expansió:

pompe d'extraction → *nucli:* *pompe*

expansió: extraction

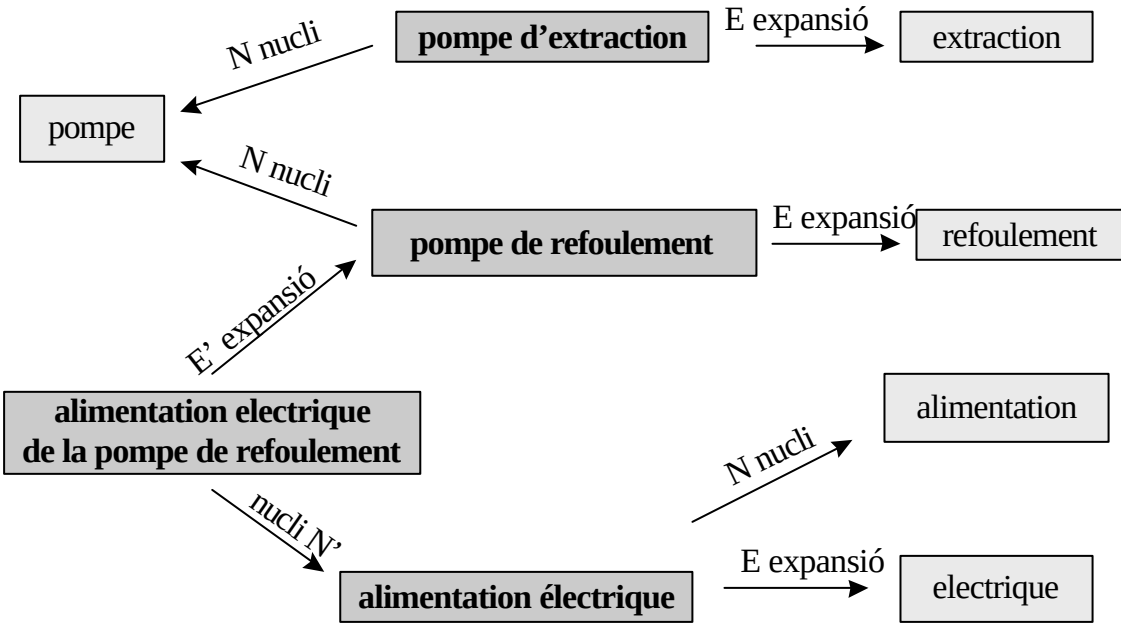
En aquest punt el sistema es troba amb situacions d'ambigüitat, com ara les seqüències:

N A A

N prep N A

que no sap com s'ha d'analitzar perquè treballa amb estructures planes i no fa cap anàlisi sintàctica. Per a la resolució d'aquests casos es recorre també a un sistema d'aprenentatge endogen similar al que hem presentat en el mòdul de delimitació.

4. En el **mòdul d'estructuració** s'organitza la llista de candidats a terme en una xarxa anomenada xarxa terminològica. Per obtenir-la és suficient recórrer a la llista de candidats a terme i reconèixer les distintes parts de cada un. Presentem a continuació un exemple d'aquestes xarxes:



Aprofundint en l'anàlisi dels candidats a terme, l'autor desglossa els tipus d'enllaços existents entre les diferents parts d'un candidat a terme. Així, en el corpus de prova LEXTER troba 16.526 enllaços (productivitat global p) que es reparteixen de la manera següent:

5.463 tipus N (pN)
5.463 tipus E (pE)
2.800 tipus N' (pN')
2.800 tipus E' (pE').

Aquesta classificació permet una agrupació dels components dels candidats a terme segons la posició que ocupen (N, E, N' i E'). Amb aquestes dades també es defineix la *taxa de productivitat normal* i la *taxa de productivitat ponderada*:

taxa de productivitat	
normal	ponderada
xN = pN/p	cN = xT log(pN)
xE = pE/p	cE = xE log(pE)
xN' = pN'/p	cN' = xT' log(pT')
xE' = pE'/p	cE' = xE' log(pE')

Aquests coeficients no s'utilitzen com a filtre, sinó que es presenten al terminòleg com una dada més per facilitar l'avaluació dels candidats a terme. A més, implícitament, serveixen per extreure possibles unitats monolèxiques amb significació especialitzada.

5. Finalment en el **mòdul de navegació** es construeix una interfície de consulta, *hipertext terminològic,* a partir del corpus inicial, de la xarxa de candidats a termes i dels coeficients i llistes ja mencionades.

LEXTER s'utilitza actualment en l'explotació de diferents corpus de la societat EDF, bàsicament, en la indexació automàtica de textos, en els sistemes de consulta hipertextual de documentació tècnica, en l'adquisició de coneixement i en la construcció de bases de dades terminològiques. LEXTER també s'utilitza com a extractor de terminologia en la base de coneixements terminològics dissenyada pel grup de terminologia de TIA [Condamines, 1995] i a SYCLADE [Habert i al., 1996], eina de classificació de paraules[44].

1.6.2.2 Valoració

Actualment, LEXTER és el SEACAT més citat i el més utilitzat per a diferents aplicacions. L'eficiència d'aquest sistema es ressent, però, d'errors en el marcatge i en la desambiguació prèvia. Aquest sistema (com tots els que utilitzen tècniques simbòliques) genera una quantitat notable de soroll: d'un corpus de 200.000 paraules s'obtenen 20.000 candidats a terme

[44] Syclade simplifica l'arbre d'anàlisi complex en arbres elementals. Així, de l'arbre d'anàlisi del candidat a terme *stenose serre de le tronc commun gauche,* s'obtenen els arbres elementals de les seqüències següents:

- *stenose serre*
- *stenose de tronc*
- *tronc commun*
- *tronc gauche.*

Aquests arbres elementals permeten construir classes de contextos sintàctics. Del primer arbre elemental (*stenose serre*) deriven dos contextos possibles: ~ *serre* i *stenose* ~, on "~" representa el nucli del sintagma. Les paraules nucli es convertiran en el nusos d'un graf mentre que el context servirà d'etiqueta de les extensions. Per exemple, a partir dels candidats a terme *stenose sevère* i *lesion sevère* es pot establir la relació següent:

estenose —*severe*— *lesion*

La idea subjacent és que les extensions relacionen paraules semànticament pròximes. Si cada paraula de l'extrem té associada la freqüència d'aparició de cada arbre elemental, la profunditat del graf variarà segons el sostre escollit. Això faria surar les paraules més pròximes que pertanyen a classes conceptuals com: *malalties, actes mèdics, parts del cos, graus d'afecció, relacions "part de".*

que després de la validació es redueixen a 10.000. L'autor[45] remarca també el problema del silenci (termes no reconeguts) que valora com el 5% del total dels termes vàlids.

Com la majoria de sistemes, es limita a la detecció de sintagmes nominals, ja que els verbs són considerats frontera de terme i mai no s'incorporen. LEXTER es va dissenyar inicialment per ser aplicat sobre textos anglesos, més tard s'ha adaptat a la llengua francesa.

Un dels punts més valorats del sistema és el mecanisme d'aprenentatge endogen que permet treballar autònomament sense necessitat de tenir accés a un diccionari complex i voluminós.

1.6.3 NODALIDA

Autor:	Arppe, A.
Data:	1995
Objectiu:	Eina d'extracció de candidats a terme
Llengua de treball:	Anglès

1.6.3.1 Sinopsi

NODALIDA és un producte dissenyat per l'empresa Lingsoft a partir de l'eina NPtool[46] desenvolupada a la Universitat de Hèlsinki, Departament de Lingüística General. L'objectiu de NPtool és, bàsicament, generar llistes dels sintagmes nominals de les frases d'un text i proporcionar una

[45] Bourigault, D. (desembre, 1996): *Detecció de Termes : estat de la qüestió*. Jornada de treball. Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra, Barcelona

avaluació de la certesa d'aquests sintagmes com a candidats a termes (OK/?). D'aquestes llistes s'extreuen totes les subcadenes permissibles i es multiplica per tres la llista inicial. Així, per exemple, per a la frase:

exact form of the correct theory of quantum gravity

NPtool proposa la llista següent de NP addicionals:

exact form of the correct theory
exact form
form of the correct theory of quantum gravity
form
form of the correct theory
correct theory
theory
quantum gravity
gravity.

D'aquesta manera i encara que no és un objectiu explícit, NODALIDA pot reconèixer els termes monolèxics que formen part dels sintagmes nominals complexos.

Paral·lelament, operen una sèrie de premisses, com la següent, que serveixen de primer filtre:

"Els NP que comencin amb determinant, adjectiu o una frase prefixada (kind of, some, one,...) són eliminats."

[46] Aquesta eina s'utilitza també en el marc del projecte TRANSTERM, finançat per la Unió Europea. Per a més informació podeu consultar [Ahmad i al., 1996].

Per a la resta de sintagmes nominals, es calcula la freqüència d'aparició, s'ordenen i agrupen segons el nucli gramatical i la freqüència, i es presenten al terminòleg acompanyats del context.

NODALIDA utilitza finestres per presentar els resultats; d'aquesta manera, el terminòleg pot determinar fàcilment els sintagmes nominals proposats susceptibles de ser considerats termes.

L'eina NPtool [Voutilainen, 1993], el cor del sistema, és un detector de sintagmes nominals i de resolució d'ambigüitats estructurals basat en el formalisme de la *gramàtica de restriccions* (*constraint grammar*) [Karlsson, 1990]. Les característiques bàsiques d'aquesta eina són que:

- L'anàlisi morfològica es basa en una descripció molt acurada que inclou la categoria morfològica i la funció sintàctica.
- La descripció morfològica es basa en regles lingüístiques. S'eviten, però, distincions que depenguin del coneixement a un nivell superior del sintàctic.
- L'anàlisi, tant en la gramàtica com en el lexicó, s'aplica a un corpus amb text no controlat.
- La desambiguació s'efectua amb criteris estrictament lingüístics. Aquest procés deixa ambigües entre un 3% i 6% del total de paraules.

El text se sotmet a un procés previ per determinar les fronteres de frase, locucions, compostos, signes tipogràfics, etc. A continuació, s'analitza morfològicament i s'obté un resultat com el següent[47]:

*("<*the>" ("the" DET CENTRAL ART SG/PL (@>N)))*
("<inlet>" ("inlet" N NOM SG))

[47]El significat dels símbols utilitzats per l'autor és el següent: @>N: premodificadors; @<N: postmodificadors; @: conjuncions de coordinació i subordinació; V: verb i marcador

("<and>" *("and" CC (@CC)))*
("<exhaust>" *("exhaust" <SVO> V SUBJUNCTIVE VFIN (@V))*
("exhaust" <SVO> V IMP VFIN (@V))
("exhaust" <SVO> V INF)
("exhaust" <SVO> V PRES -SG3 VFIN (@V))
("exhaust" N NOM SG))
("<manifold>" *("manifold" N NOM PL)).*

En aquest moment, es produeix la desambiguació. Per exemple per a la frase:

> *The inlet and exhaust manifolds are mounted on opposite sides of the cylinder head.*

S'obtenen les dues anàlisis següents:

> *a) ... on/@AH opposite/@N sides/@NH of/@N< the/@>N* ***cylinder/@NH head/@V***
> *b) ... on/@AH opposite/@N sides/@NH of/@N< the/@>N* ***cylinder/@>N head/@NH.***

El fet de considerar o no la seqüència final (*cylinder head*) com un sintagma nominal és l'única diferència entre les dues anàlisis proposades. El procés només dóna dues anàlisis possibles per a cada frase: una en què es dóna preferència als NP de longitud màxima (*NP-friendly*) i una altra en què es dóna preferència als NP de longitud mínima (*NP-hostile*). A continuació, el sistema compara les dues estratègies i assigna una valoració d'aquest tipus:

---

d'infinitiu; NH: nucli nominal; "<" i ">" indiquen la direcció en què es troba el nucli,

OK: la mateixa anàlisi és proposada per les dues estratègies

?: l'anàlisi és proposada només per a una de les estratègies

Així, per a l'última frase, el sistema dóna les anàlisis següents:

OK: inlet and exhaust manifolds
OK: exhaust manifolds
?: opposite sides of the cylinder.
?: opposite sides of the cylinder head.

D'aquesta manera, el terminòleg rep una llista de candidats a terme per validar, amb aquesta informació addicional. Segons les dades fetes públiques pels autors [Arppe, 1995], els resultats obtinguts per l'eina auxiliar NPtool són força bons: precisió = 95-98% i recall = 98,5-100% respecte d'un text d'unes 20.000 paraules.

1.6.3.2 Valoració

El sistema NODALIDA es basa en l'ús de coneixement lingüístic mitjançant una aproximació estructural (detecció d'estructures sintagmàtiques i corresponent desambiguació estructural).

Els resultats presentats per l'autor mostren un alt grau de qualitat amb corpus força reduï ts, però no se sap si es mantindrien amb corpus de dimensions més grans. Tampoc no queda clar com l'autor calcula les xifres de precisió i *recall,* ni quins són els termes considerats correctes: si només els que tenen el signe OK o tots. S'ha de fer notar també que aquests resultats es refereixen a l'eina auxiliar NPtool i no al sistema NODALIDA.

---

esquerra o dreta.

NODALIDA aconsegueix una desambiguació molt alta, encara que amb un elevat nombre de regles, que comporten un problema considerable de gestió i de control.

La llista que el sistema proposa al terminòleg per validar són tots els candidats valorats amb els signes **OK** o **?**. La manera d'obtenir els possibles sintagmes nominals fa pensar que en la llista a validar hi ha molts candidats que el terminòleg ha de rebutjar, encara que segons els autors sigui un mètode amb un alt grau de precisió.

També cal remarcar que és un dels únics sistemes que, implícitament, reconeix les unitats terminològiques monolèxiques, tot i que només les que formen part d'un sintagma nominal complex.

1.6.4 FASTR

Autor:	Jacquemin, C.
Data:	1996
Objectiu:	Eina d'extracció de candidats a terme
Llengües de treball:	Una versió en francès i una versió en anglès.

1.6.4.1 Sinopsi

El punt de partida d'aquest treball realitzat per Jacquemin a la Universitat de Nantes és la idea de construir una eina per detectar terminologia que s'aprofiti dels termes ja coneguts i acceptats. Aquests termes poden provenir d'una base de dades existent o ser recollits prèviament per la mateixa eina i validats pel terminòleg. La idea de base és no començar cada vegada de zero.

El primer pas és, doncs, obtenir i analitzar un conjunt de termes ja existents per després extreure'n les variants possibles. Per assolir aquest primer objectiu, FASTR utilitza un analitzador parcial que a partir de cada terme obté una regla que després expandeix en les seves variants. Per exemple, el terme *serum albumin* que correspon a una seqüència **Nom-Nom** respon a una regla d'aquest tipus:

regla 1: $N_1 \rightarrow N_2\ N_3$
<N2 lema>= serum
<N3 lema>= albumin.

El pas següent consisteix a aplicar al corpus textual les regles generades amb el corpus d'unitats inicial i, a través de les metaregles, l'extractor genera possibles variants de cada terme que es trobin a la llista referència. Si tornem a la regla anterior, per exemple, es podria aplicar la metaregla següent:

$\text{Coor}(X_1 \rightarrow X_2\ X_3)\quad X_1 \rightarrow X_2\ C_4\ X_5\ X_3$

i s'obtindria una regla nova:

$N_1 \rightarrow N_2\ C_4\ X_5\ N_3$

Mitjançant aquesta nova regla s'accepten noves construccions que substitueixen C_4 per una conjunció i X_5 per una paraula aï llada, com ara *serum and egg albumin*. El candidat a terme no és la nova construcció sencera, sinó el terme *coordinat*[48], en aquest cas: *egg albumin*. Les paraules que han donat lloc a la nova regla (*egg* i *albumin*) mantenen la seva funció d'equacions de restricció de la regla original i, a més a més, serveixen d'ancoratge per aplicar la metaregla.

Una metaregla pot tenir associades restriccions específiques per limitar-ne l'aplicació. Per exemple:

<C_4 lema> ! but
<X_5 cat> ! Dd
<X_5 cat> ! Di

D'aquesta manera, es rebutgen les seqüències sense cap relació lèxica com és el cas de *serum and the albumin*[49]. Aquesta regla pertany a la classe de les regles de coordinació, però el sistema en fa servir també d'inserció i de permutació.

Un exemple de regla d'inserció és la següent:

X_2 X_3→ X_2 **X_4** X_3
donat *medullary carcinoma*, si troba *medullary thyroid carcinoma* el terme incorporat a la llista és *thyroid carcinoma*.

I una regla de permutació:

X_2 X_3→ X_3 **P_4** **X_5** X_3

donat *control center*, si troba *center for desease control* el nou terme proposat és *disease control*.

La metagramàtica de FASTR per a l'anglès inclou 73 metaregles :

- 25 de coordinació

[48] L'autor utilitza aquesta denominació per referir-se a una UTP formada per juxtaposició.
[49] La restricció que s'aplica en aquest cas és que l'element X_5 no pot ser un determinant.

- 17 d'inserció
- 31 de permutació.

Cada metaregla està lligada a un extractor de patrons, eina que permet adquirir informació molt ràpidament. El formalisme gramatical utilitzat per FASTR és una extensió de PATR-II [Shieber, 1986], llenguatge que permet escriure gramàtiques utilitzant estructures de trets. Les regles que descriuen els termes estan formades per una part lliure de context ($N_1 \rightarrow N_2\ N_3$) i un conjunt d'equacions que indiquen les restriccions (ex. <N2 lema>=serum i <N3 lema>=albumin). El sistema, en primer lloc, filtra les regles que ha d'aplicar segons el text d'entrada i, a continuació, analitza el text (aquest mecanisme, però, no sempre té èxit, ja que l'èxit depèn de la paraula afegida).

Pot donar-se el cas que la metaregla permeti redescobrir un terme ja reconegut prèviament i d'aquesta manera s'estableixi un lligam entre aquestes dues unitats. Els casos d'elisions (com ara *Kerr magnetooptical effect* i *Kerr effect*) no es tracten, ja que l'aproximació escollida no és apropiada per a aquest tipus de referències.

És important notar que el procés és incremental: a partir d'un conjunt de termes ja coneguts el sistema en detecta de nous, la qual cosa permet reiniciar una altra vegada el cicle i detectar més candidats. El procés continua fins que ja no es detecten nous termes.

L'autor va realitzar un experiment a partir d'un corpus de medicina d'1,5 milions de paraules i una llista de referència de 70.000 termes de diferents àrees temàtiques. Després de quinze cicles es van detectar 17.000 termes dels quals 5.000 eren nous. La velocitat de processament va ser de 2.562 paraules/minut. Aquest sistema, però, es degrada quan la llista de referència disminueix.

Es planteja l'existència d'una relació conceptual entre els nous termes i el terme que n'ha permès el reconeixement. Aquesta relació varia segons el tipus de regla aplicada: inserció o coordinació. La permutació no permet d'establir cap relació donada la naturalesa sintagmàtica de la relació.

Es parteix del supòsit que quan dos termes apareixen coordinats comparteixen sempre un mateix esquema semàntic. Per exemple, *dorsal spine* and *cervical spine* poden relacionar-se per coordinació i, en canvi, cap del dos apareixerà coordinat amb *fish spine* (eriçó) perquè pertanyen a classes semàntiques diferents. Així, el sistema defineix una classe conceptual de la manera següent:

> "*Dos termes t i t' estan en la mateixa classe si i només si existeix una cadena de coordinació entre t i t'.*"

Amb aquest tipus de relació es presenta la possibilitat de construir grafs que agrupin les diferents *classes* de paraules indicant quin és el terme origen i quin/s el/s derivat/s. Per exemple la relació:

> *normal control* → *uraemic control*

ens indica que el segon terme s'ha obtingut a partir d'una coordinació amb el primer (ha trobat la seqüència *normal and uraemic control* després de trobar o tenir com a referència *normal control*).

Quan es troba un nou terme per inserció, també existeix una relació conceptual entre els dos termes, a més, en aquest cas l'autor observa una gran semblança amb la taxonomia corresponent. Per posar un exemple:

→ *superficial tumor*

malig/benign tumor → *mixed tumor*
→ *mediastinal tumor*
→ <*part-of-body*> *tumor.*

1.6.4.2 Valoració

FASTR és un sistema retroalimentatiu, amb l'inconvenient, però, que els termes que s'afegeixen a la llista de termes *vàlids* de manera automàtica no reben cap tipus de validació. Aquest tractament permet afegir a la llista inicial termes no vàlids que poden donar lloc, en els cicles successius, a la incorporació de més termes no vàlids. Tot i aquesta limitació, l'autor opina[50] que la validació dels termes no és una font d'error important ja que *normalment* candidats incorrectes no produeixen nous possibles candidats a termes.

És veritat que utilitzar els termes ja reconeguts i acceptats és un mecanisme molt útil per detectar unitats terminològiques, però, en contrapartida, limita, en certa manera, la possibilitat de reconèixer termes que no estan relacionats amb els de l'origen. Aquesta tècnica s'hauria de poder complementar amb altres estratègies d'extracció per poder reconèixer els termes que no estan relacionats amb cap unitat terminològica polilèxica (ja que els termes inicials sempre són unitats terminològiques polilèxiques i una metaregla no pot donar com a resultat una unitat monolèxica).

També cal destacar que la noció de variació que s'aplica en aquest sistema no té fonaments lingüístics, sinó gràfics. L'autor utilitza el terme *variació terminològica* només en el sentit formal, és a dir, FASTR no distingeix si dos segments són el mateix concepte o són dos conceptes diferents;

[50]Comunicació personal.

d'aquesta manera, dos termes cohipònims són considerats variants, i un hiperònim respecte del seu hipònim també.

1.6.5 NAULLEAU

Autor:	Naulleau, E.
Data:	1998
Objectiu:	Eina d'extracció de sintagmes nominals pertinents
Llengua de treball:	Francès

1.6.5.1 Sinopsi

El sistema d'extracció de terminologia dissenyat per Naulleau és un programa d'indexació lliure de grups nominals complexos, ja que l'autor explícitament rebutja tractar les UT formades d'un sol mot perquè les considera massa polisèmiques. L'objectiu del programa és la indexació lliure de documents restringida per informacions morfològiques, sintàctiques i semàntiques. Per això, utilitza una gramàtica del sintagma nominal creada *ad hoc* basada en les generalitzacions de les propietats lingüístiques que presenten les llistes de sintagmes nominals que es proporcionen al sistema cada vegada que es vol aplicar. De tal manera que, a partir d'aquestes llistes de sintagmes nominals pertinents i de sintagmes nominals no pertinents, es configura un perfil negatiu i un de positiu sobre els quals es construeix la gramàtica.

En tots els SEACAT dissenyats fins a 1997, l'usuari ha d'intervenir al final del procés; en el sistema de Naulleau també, però en aquest sistema l'usuari es converteix en essencial perquè també ha d'intervenir abans de començar a aplicar-se el programa; com afirma el mateix autor *"nous souhaitons donner la possibilité à l'opérateur humain d'intervenir directement sur le jugement d'intérêt qu'il porte sur la forme des marques textuelles qui constituent ses indices de recherche dans un document"* [Naulleau, 1998:13]. Abans de l'aplicació de l'extractor, l'usuari li facilita dues llistes: una amb els sintagmes nominals que l'usuari consideri pertinents i una altra amb els sintagmes que considera no-pertinents per tal que el sistema pugui construir una gramàtica sobre la qual es fonamentarà l'extracció.

Una vegada proporcionades les llistes, el sistema les analitza utilitzant coneixement morfològic (categoria gramatical), sintàctic (aplicació d'un *parser* sintàctic) i semàntic (aplicació d'una reducció de les classes semàntiques d'AlethDic[51]) i crea dos tipus de filtres: un conjunt de filtres negatius i un conjunt de filtres positius. Aplicant els filtres es detecten les unitats considerades terminològicament pertinents per a l'usuari inicial que són analitzats també amb coneixement morfològic, sintàctic i semàntic

Segons Naulleau, és difícil analitzar el grau d'èxit del seu programa perquè aquest varia segons el tipus de segments que l'usuari proporciona, per bé que l'autor ha comprovat que els millors resultats s'obtenen quan el sistema utilitza informació semàntica per al reconeixement dels termes.

1.6.5.2 Valoració

[51]És el diccionari electrònic, amb informació morfològica, sintàctica i semàntica, de la Societé ERLI.

Aquest sistema, presentat com a tesi doctoral per Naulleau a la Universitat Paris XIII [Naulleau, 1998], es pot considerar el primer SEACAT que incorpora informació semàntica a partir d'un tesaurus.

És interessant notar que per primera vegada d'una manera *explícita* l'usuari i la noció de pertinència d'una UT són a la base de la concepció d'un sistema d'extracció automàtica de terminologia. Com ja hem remarcat en l'apartat anterior, aquests dos paràmetres són presents *implícitament* en la filosofia de diversos sistemes com TERMINO o LEXTER, quan els autors afirmen que el soroll generat no ha de preocupar excessivament perquè d'aquesta manera l'usuari pot escollir, sense restriccions, tots els segments que realment l'interessen.

Tot i la seva innovació, l'estratègia proposada per Naulleau suposa un procés massa llarg, ja que cada vegada que es vol fer servir el sistema un usuari ha de proporcionar a l'ordinador una llista de sintagmes terminològicament pertinents i una altra de sintagmes no-pertinents. A més, el fet que l'usuari intervingui al principi del procés no pressuposa que no hagi d'intervenir també al final del procés, ja que el sistema genera silenci i soroll.

Un altra limitació del sistema és el fet que només filtra sintagmes perquè es basa en dependències sintàctiques. Per filtrar unitats simples hauria de posseir, com a mínim, un analitzador de dependències elementals a nivell de frase o fer servir altres tipus de recursos.

1.6.6 ACABIT

Autor:	Daille, B.
Data:	1994
Objectiu:	Eina d'extracció de candidats a terme
Llengua de treball:	Francès

1.6.1.1 Sinopsi

La idea bàsica d'aquest treball és combinar el coneixement lingüístic amb mesures estadístiques. Per això, el corpus ha d'estar marcat morfològicament (les proves s'han realitzat sobre dos corpus de l'àmbit de les telecomunicacions de 500.000 paraules cadascun). A continuació, es crea una llista de candidats a terme d'acord amb les seqüències de text que responen a patrons sintàctics de formació de termes. Amb aquesta informació, s'utilitzen mètodes estadístics per filtrar aquesta primera selecció.

Partint del supòsit que tot banc terminològic està format bàsicament per noms compostos i que la majoria de compostos de longitud igual o major a tres constituents poden descompondre's en forma binària, el programa se centra en la detecció dels **noms compostos binaris**.

Els patrons que Daille ha considerat rellevants per al francès són: N_1 PREP (DET) N_2, i N A, juntament amb algunes variants tractades específicament com ara N_1 PREP de(DET) N_2 i N_1 PREP à(DET) N_2, i estructures com ara la coordinació a la dreta o a l'esquerra. Sobre aquests patrons s'apliquen els algorismes estadístics.

La tècnica utilitzada per reconèixer els patrons és la dels autòmats finits. Els autòmats són representats per un subconjunt d'etiquetes gramaticals a les quals s'afegeixen alguns lemes, formes flexionades i algun signe de puntuació. Així, podem considerar els autòmats com a filtres lingüístics que seleccionen els patrons definits i en determinen la freqüència d'aparició, la distància i la variància. Cadascun dels patrons morfosintàctics té associat un autòmat finit específic.

El tractament estadístic que s'aplica al corpus es basa en una sèrie de mesures estadístiques de diferents tipus: mesures de freqüència, criteris d'associació, criteris de diversitat i mesures de distància. ACABIT considera els dos lemes que formen una parella dintre d'un patró com a dues variables sobre les quals es vol mesurar el grau de dependència. Les dades es representen en una taula de contingència que té aquest aspecte:

	L_2	L_n
L_1	a	b
L_m	c	d

on:

a = ocurrències de la seqüència L_1L_2

b = ocurrències de $L_1 + L_n$ (n != 2)

c = ocurrències de $L_m + L_2$ (m != 1)

d = ocurrències de $L_m + L_n$ (m != 1 i n != 2).

En total s'apliquen divuit mesures diferents amb l'objectiu d'establir el grau d'independència de les variables de la taula de contingència. D'una primera anàlisi dels resultats, l'autora dedueix que només quatre d'aquestes mesures són pertinents per al propòsit fixat:

- la freqüència
- el criteri d'associació amb el numerador al cub[52] (IM^3)
- el criteri de versemblança [53]

[52] Fórmula obtinguda experimentalment a partir de la xifra d'associació descrita per Brown i al. (1988) amb l'objectiu d'afavorir les parelles més freqüents:

$IM^3 = \log_2 (a^3/(a+b)(a-b))$

En la seqüència de paraules que satisfan el filtre lingüístic N1 (Prep (Det)) N2, els valors més alts depenen del nombre d'ocurrències independentment del nombre de seqüències trobades; i els valors més petits de la xifra d'associació i de la xifra d'associació amb numerador al cub corresponen a paraules que poques vegades apareixen juntes, sinó separades (*système terre, code signalisation* ...).

[53]Aquest coeficient introduï t per Dunning (1993) és la prova de relació de versemblança aplicada a una llei binomial:

- el criteri de Fager/MacGowan[54].

1.6.6.2 Valoració

En aquest sistema, a diferència del que succeeix en d'altres[55], la freqüència ha estat una de les mesures més importants per detectar els termes d'una àrea, però la classificació que resulta d'aplicar aquesta freqüència selecciona en un nombre important seqüències freqüents que no són termes i, en canvi, no proposa termes que són poc freqüents. L'autora és conscient que aplicar mesures estadístiques implica una certa taxa de silenci perquè els termes de freqüència molt baixa no poden ser reconeguts.

Així, la mesura que es pren com a òptima és el criteri de versemblança ja que:

- és un veritable test estadístic
- proposa una classificació que té en compte la freqüència
- té un comportament correcte amb corpus mitjans i grans
- no està definit en els casos que no interessa recollir.

De totes maneres, utilitzar aquesta mesura comporta un cert soroll perquè:

- hi ha errors en el marcatge morfològic
- es presenten com a candidats combinacions que no són mai compostes (*ko bits*, *à titre d'exemple*, etc.)

Loglike = a log a + b log b + c log c + d log d - (a + b) log (a + b)
(a+c) log (a+c) - (b+d) log (b+d) - (c+d)log (c+d) + N log N

Aquest coeficient selecciona les mateixes parelles que la xifra d'associació amb numerador al cub per als valors més grans i no està definit quan els seus components només apareixen dintre de la d'una parella de noms compostos.

[54] Criteri de biologia que dóna resultats semblants al criteri d'associació. Dóna molta rellevància a parelles que apareixen sovint juntes i poques vegades separades, sense rebutjar sistemàticament els termes, els constituents dels quals apareixen sovint aï lladament.

- es presenten combinacions de longitud major o igual a tres que no són totalment pertinents (*bande latérale -unique-*, *service fixe -par satellite-*, etc.).

1.7 Conclusions

Els aquest capítol hem analitzat els trets substancials que defineixen els extractors de terminologia i hem posat de relleu les seves limitacions tenint en compte que la valoració dels resultats està condicionada pels corpus d'aplicació, que són molt **petits** i altament **especialitzats** —tant pel que fa al tema com al nivell d'especialització; i també per la mancança de dades públiques. Tenint en compte aquestes limitacions, podem dir a mode de síntesi que:

a) Cap dels SEACAT és totalment **satisfactori.** Aquesta afirmació se sustenta bàsicament en dos fets: d'una banda, tots els sistemes produeixen una quantitat massa gran de **silenci,** sobretot els de base estadística[56]; de l'altra, tots generen una quantitat molt elevada de **soroll,** sobretot els de base lingüística que utilitzen una sèrie de **patrons** morfosintàctics per identificar els termes complexos i, per tant, només es basen en l'aspecte formal de la UT.

b) Atès el soroll que es genera, tots els sistemes d'extracció proposen **llistes de candidats a terme** que, al final del procés, s'han d'acceptar o rebutjar manualment. Consegüentment, podem afirmar que tots els SEACAT són només parcialment eficaços perquè el procés de buidatge terminològic no és del tot automàtic.

---

[55][Church i Hanks, 1989].

[56]Hi ha un consens bastant generalitzat entre els diferents autors a considerar la **freqüència** com un bon indicador que un candidat a terme sigui finalment una unitat

c) Tots aquests sistemes avaluats se centren exclusivament en el **substantiu**, cap sistema d'extracció fa referència a altres categories gramaticals. Aquest fet està motivat per l'alt percentatge d'unitats terminològiques que s'usen en els textos especialitzats. El que també és un fet, però, és que en tots els llenguatges d'especialitat hi ha unitats lèxiques verbals, adjectivals i adverbials amb significat especialitzat, i combinacions específiques de base verbal, encara que el seu percentatge sigui inferior.

d) Tots els sistemes se centren en el reconeixement de la unitat terminològica polilèxica (UTP), tot i que alguns extractors reconeixen de manera indirecta també les unitats de significació especialitzada monolèxiques que integren una UTP.

e) Cap sistema d'extracció de terminologia fa referència a la delimitació entre col·locacions nominals i UTP; ni tampoc a l'extracció de fraseologia verbal[57].

f) Només un dels sistemes analitzats utilitza informació semàntica per reconèixer i delimitar les UT.

g) Cap sistema fa servir a fons les característiques combinatòries pròpies dels termes dels llenguatges d'especialitat lligats a una temàtica. Seria important disposar de més estudis sobre el tipus de restriccions que presenten les UT en relació amb:

- el camp conceptual

terminològica, encara que per si sola **no** sigui un paràmetre **suficient**, perquè si s'usa només la freqüència, es genera força silenci.

[57]Segons l'aplicació de l'extractor de terminologia, per exemple com a suport a la traducció automàtica, s'haurien de tenir en compte els verbs terminològics i la fraseologia.

- el tipus de text.

Algunes de les estratègies usades per diferents sistemes que ens semblen particularment interessants són:

- l'ús de regles heurístiques, en relació tant amb la unitat terminològica com amb allò que no pot ser mai una unitat terminològica
- la construcció de xarxes de complements i de nuclis dels termes complexos
- la reutilització de termes ja reconeguts
- l'anàlisi parcial de les frases per obtenir sintagmes nominals potencialment terminològics
- l'extracció de relacions semàntiques entre els termes —i entre els seus components
- la importància de les característiques de disposició de les unitats terminològiques en els textos, etc.
- la combinació de més d'una estratègia
- l'ús d'un diccionari amb informació semàntica sobre el lèxic.

Per millorar aquests sistemes d'extracció de terminologia i aconseguir que es redueixi tant el silenci com el soroll que generen, caldria aprofundir principalment en dos tipus d'estudis. D'una banda, caldrien **més estudis lingüístics** sobre:

- les categories gramaticals susceptibles d'ésser termes en els diferents àmbits d'especialitat
- la influència de la funció sintàctica dels sintagmes terminològics en els textos

- les relacions semàntiques entre els diferents constituents d'una unitat terminològica
- les relacions semàntiques dels termes
- la interpretació semàntica del lèxic a partir de corpus textuals
- la representació semanticolèxica
- la seva disposició dels termes en els textos
- les relacions en llengües diferents dels termes d'una mateixa temàtica.

I, d'altra banda, caldria treballar en la via de **sistemes informàtics** que:

- alternessin de manera més activa els mètodes estadístics amb els lingüístics
- milloressin les mesures estadístiques
- combinessin més d'una estratègia
- milloressin les interfícies per afavorir la interacció màquina/usuari.

En definitiva, si es vol avançar en el camp de l'extracció automàtica de terminologia, s'han d'interaccionar activament els mètodes lingüístics i algunes tècniques estadístiques. Els objectius d'aquestes millores aconseguirien que els extractors, d'una banda, reduï ssin al màxim el silenci i el soroll, i, de l'altra, reconeguessin tots els termes d'un text, tant els monolèxics com els polilèxics.

1.8 Recapitulació

En aquest capítol hem explicitat què són els SEACAT, quina és la seva funció principal i les aplicacions en diverses matèries aplicades. Hem analitzat diversos SEACAT i hem arribat a la conclusió que tots ells presenten un funcionament correcte, encara que no satisfactori i, en conseqüència, millorable. Aquesta millora creiem que ha de venir fonamentalment per una caracterització lingüística més específica i pertinent de les unitats terminològiques reals.

En el treball de recerca *Les unitats terminològiques polilexemàtiques en els lèxics d'especialitat* [Estopà, 1996b] vam descriure les UTP a partir de corpus lexicogràfics; en el proper capítol compararem els resultats de l'anàlisi que vam fer amb el funcionament de les unitats especialitzades en els textos, amb la finalitat de validar les hipòtesis postulades al treball de recerca i poder detectar per què no són satisfactoris els resultats que generen els sistemes d'extracció semiautomàtica de terminologia basats en algun tipus de patró formal, sigui morfosintàctic o lèxic.

2. LES UNITATS TERMINOLÒGIQUES EN ELS TEXTOS

> *On est aujourd'hui sur un tout autre paradigme: le travail scientifique est considéré comme en grande partie constitué par du langage, plus spécialement par des textes et la connaissance scientifique est elle-même considérée comme une information conceptuelle obtenue à partir de textes.*
>
> [Slodzian, 1995: 14]

2.1 Punt de partida: resultats del treball de recerca

En el treball de recerca *Les unitats terminològiques polilexemàtiques en els lèxics especialitzats: dret i medicina*, presentat el 1996, ens vam proposar, a partir d'un corpus lexicogràfic especialitzat[1]:

- descriure de manera exhaustiva els patrons formals de les UTP en català de dues àrees del saber científic, el dret i la medicina, i

- caracteritzar el no-terme, és a dir, analitzar el tipus de categories gramaticals que no formen part d'una UTP, almenys en aquests dos àmbits.

Per dur a terme aquests objectius vam analitzar quatre aspectes de les unitats terminològiques polilèxiques:

- la categoria gramatical
- les estructures formals
- la freqüència de les estructures formals
- la relació semanticoformal dels constituents.

Les conclusions a què vam arribar amb aquesta anàlisi van ser les següents:

1. Vam constatar que la categoria gramatical única de les UTP presentades en els diccionaris és la **nominal**.

2. Vam comprovar que la gran majoria d'UTP correspon a un **nombre molt baix de patrons estructurals morfològics**:

dret: 90% de les UTP	5 estructures
medicina: 85% de les UTP	6 estructures

i que, contràriament, hi ha una gran dispersió d'estructures morfològiques que equivalen a molt poques UTP; de fet, la majoria d'aquests patrons només presenten una sola ocurrència dins del nostre corpus.

dret: 10% de les UTP	70 estructures
medicina: 15% de les UTP	50 estructures

3. D'aquesta manera, vam poden afirmar que les estructures morfosintàctiques **més productives** en els dos dominis es concreten en les dues següents[2]:

	dret	medicina
N SAdj[3]	49,85% de les UTP	64,05% de les UTP
N SPrep	47,39% de les UTP	33,3% de les UTP
	97,24% de les UTP	**97,35%** de les UTP

[1]Vegeu el capítol primer *Introducció: marc de treball, tema d'anàlisi, unitat d'anàlisi, àmbits, corpus, idees prèvies, objectius, pla de treball* de [Estopà, 1996b: 8-28].
[2]Vegeu [Estopà, 1996b: 134, 154].
[3] En aquest capítol hem preferit utilitzar les estructures planes que havíem fet servir en el treball de recerca de doctorat per facilitar la comparació.

4. També vam constatar que la gran majoria de les UTP estan formades per un nucli i **un sol complement**:

	dret	medicina
1 complement	93% de les UTP	94% de les UTP
2 complements	5% de les UTP	4,7% de les UTP
3 complements	0,4% de les UTP	0,7% de les UTP
4 complements		0,4% de les UTP

5. Així mateix i pel que fa a les característiques del no-terme, vam remarcar, com ho havien fet altres autors per a l'anglès i per al francès [Bourigault, 1994], [Drouin, 1996], que hi ha unes categories gramaticals[4] que, en català, no formen part mai d'una UTP:

- els verbs en forma personal[5]
- els adverbis[6]

[4]Els signes de puntuació també marquen frontera d'UTP ja que no poden formar part d'una unitat nominal.

[5]S'ha de tenir en compte, però, que a vegades un participi pot funcionar com un adjectiu i un infinitiu com un substantiu, encara que formant part de les UTP no és gaire freqüent:

	dret	medicina
adjectiu-participi	4.3% de les UTP	1.5% de les UTP
adjectiu-gerundi	1.3% de les UTP	0.5% de les UTP
substantiu-infinitiu	0.4% de les UTP	

[6]Això és cert per a tots els adverbis menys per al *no* adverbial que precedeix a adjectius. Encara que aquests casos són molt poc productius, vam trobar una ocurrència en dret i una en medicina (*edema pulmonar no cardiogènic*) en què hi havia un adjectiu precedit per aquest adverbi. De totes maneres, aquest adverbi *no* davant d'un nom o d'un adjectiu és considerat per alguns autors un prefix negatiu, [Fabra, 1932], [Badia i Margarit, 1995],

- els determinants demostratius
- els determinants possessius
- els determinants quantitatius
- els determinants indefinits
- els determinants interrogatius
- els determinants numerals
- els pronoms
- els superlatius
- la majoria de preposicions
- la majoria de conjuncions.

En contrapartida, vam comprovar que en català són constituents representatius d'una UTP:

- els substantius
- els adjectius[7]
- la preposició *de*[8]
- la conjunció *i*[9]

[Gaatone, 1987], [Marchand, 1960], [Iacobini, 1992] entre d'altres; per tant, si considerem que aquest *no* és un prefix de negació, l'afirmació continuaria sent vàlida.

[7]En el 99% del casos l'adjectiu es posposa al nom. Els adjectius que es poden col·locar davant del nom són molt restrictius i aquesta restricció varia segons el tema. Així, per exemple, en medicina només vam trobar posposats al nom els determinants numerals ordinals (*segon adductor, tercer trocànter, quart ventricle,* etc.).

[8]Vam remarcar que la preposició *de* és la més usada en els discursos especialitzats, molt més que en altres llengües romàniques com el francès en què el camp semàntic de la preposició *de* es reparteix formalment entre la preposició *de* i la preposició *à* [L'Homme, 1996b]. En aquest sentit, els resultats del corpus lexicogràfic analitzat són els següents:

	dret	medicina
a, en, contra, davant, per	5.5% de les UTP	
de	43% de les UTP	100% de les UTP

[9]La conjunció *i* apareix en molt poques estructures com a constituent d'una unitat lèxica i, a més, aquestes estructures són molt poc productives:

	dret	medicina
estructures	9	1
ocurrències	15	1

- els articles definits *el, la, els, les.*

6. Vam observar també que el N del sintagma preposicional de l'estructura N de N, sobretot en medicina, és, moltes vegades, un nom propi (6,37% de les UTP en medicina). Aquest N sol correspondre al nom de l'investigador que ha descobert un nou concepte (*amígdala de Luschka, ampul·la de Vater, banda de Broca, banya d'Amon, gangli de Scarpa, gangli de Gasser, hiat de Fal·lopi, hiat de Scarpa,* etc).

7. Per acabar, vam observar que en les unitats lèxiques de dret hi ha una forta presència de noms abstractes que es refereixen a figures jurídiques (*oï dor de comptes, senyor útil, secretari d'estat, usurpació d'atribucions, ocultació de fill,* etc.) i, en canvi, en medicina hi ha un predomini de noms concrets —sobretot en subàmbits com l'anatomia, la patologia o l'instrumental clínic— (*nervi cortical, papil·la renal, orella interna, úlcera circular, icterícia dissociada, bisturí elèctric, bisturí de valvulotomia,* etc.).

8. Pel que fa a l'estructura morfològica NA —que és la més productiva—, vam advertir que, de manera molt general, en els dos corpus l'adjectiu de relació és el predominant i que trobem també adjectius qualificatius que adquireixen valor relacional (*paladar dur/paladar tou, taca blanca/taca blava, òrgan genital femení/òrgan genital masculí, intestí gros/intestí prim,* etc.).

9. Vam observar que, en medicina, l'adjectiu relacional indica bàsicament localització (*laberint cortical, orella externa, galea del cap,* etc.) o funció (*òrgan respiratori, òrgan genital, dent molar,* etc.).

10. Partíem de la idea que el **significat** dels sintagmes complexos especialitzats, a causa del seu caràcter volgudament descriptiu, és, en la majoria de casos, **componencial.** Tot i així, els aspectes semàntics de les UTP van ser un dels punts menys tractats del treball per tal com ens vam limitar a fer algunes observacions com les que acabem de comentar.

Fins aquí hem fet una síntesi de les principals conclusions del treball de recerca, a les quals vam arribar després d'analitzar les UTP d'un **corpus lexicogràfic**. En aquell estudi, ja vam assenyalar que el fet de partir d'un corpus lexicogràfic, en què les UTP estan fixades, podia condicionar els resultats[10], i no correspondre a l'anàlisi de les UT en context.

Per analitzar aquest aspecte, en aquest capítol, hem cregut necessari comprovar si els resultats a què vam arribar amb l'anàlisi de les UTP dels diccionaris són també vàlids quan s'apliquen a les UTP dels textos. Per això, hem configurat un corpus textual en català de l'àmbit de la medicina que servirà, en primer lloc, per verificar els resultats del treball de recerca i, en segon lloc, per constituir els primers materials de treball d'aquesta tesi doctoral.

---

[10]"*Hem evidenciat diverses vegades al llarg del treball que per estudiar més a fons les UTP i sobretot per poder-les contrastar amb unitats que traspassen els seus límits necessitaríem realitzar un estudi que es recolzés en la base d'un corpus textual. Aquesta necessitat, doncs, estableix un primer pont cap a la tesi doctoral. Així, l'obligatorietat d'establir un corpus de treball de textos especialitzats ens permetrà, d'una banda, anar més enllà en l'estudi de les UTP en el seu context natural i, de l'altra, poder contrastar les UTP amb altres unitats lingüístiques que generalment no s'acostumen a recollir en les obres lexicogràfiques. Els resultats del present estudi són interessants per poder comparar les dades extretes de les anàlisis de corpus textuals, ja que permetran de comprovar fins a quin punt les obres lexicogràfiques es desvien de l'ús real de les unitats terminològiques.*"
Estopà, R. [1996b: 200]

L'objectiu central d'aquest capítol és, doncs, comprovar que els patrons de les UTP dels diccionaris corresponen també als de les UTP que trobem en els textos. Una vegada feta aquesta verificació, el segon objectiu d'aquest capítol és analitzar com funcionen aquestes estructures quan es fan servir com a punt de partida d'un extractor de terminologia.

Per tal de poder aprofundir en l'estudi de les unitats complexes hem cregut oportú centrar aquest treball des del punt de vista de la temàtica en un sol tipus de text: textos de medicina.

2.2 Corpus de comprovació

Amb la finalitat de dur a terme aquesta comprovació hem seleccionat un text especialitzat de medicina per fer-ne el buidatge terminològic. Aquest text que constitueix el **corpus de comprovació**[11] de les hipòtesis formulades en el treball de recerca. En concret, hem seleccionat alguns dels capítols del llibre: Farreras, P.; Rozman, C. (1997) (13 ed.) *Medicina interna*. Madrid, Harcout Brace, que es caracteritza bàsicament per tres elements:

- els usuaris
- el grau d'especialització
- el nombre d'ocurrències.

Medicina Interna és un document escrit, produï t per especialistes i adreçat tant a especialistes com a estudiants universitaris. L'obra ha estat elaborada per més de 350 especialistes de les diferents branques de la medicina interna i, com s'explicita en la mateixa introducció, està

[11] Els textos que constitueixen el corpus de comprovació formen part del *Corpus textual especialitzat i plurilingüe* de l'Institut Universitari de Lingüística Aplicada. L'annex 1 recull els textos que configuren aquest corpus.

destinada *"a l'estudiant de medicina i al metge en exercici, ja es tracti del professional que exerceix dins la comunitat com a metge d'assistència primària o bé de qualsevol tipus de metge internista, tant l'especialista en medicina interna general com l'internista especialitzat en qualsevol de les branques de la medicina interna."*

Segons els mateixos experts, aquest llibre és un manual clàssic de medicina interna considerat per la qualitat dels seus autors, l'abast temàtic i l'actualitat de les informacions, una obra de referència de consulta obligatòria. És, per tant, un text representatiu i actual de l'àmbit temàtic de la medicina interna. Una mostra de la vigència del document és el fet que s'actualitza contínuament —treballarem sobre la tretzena edició— i que es pot trobar paral·lelament en castellà[12]:

> *L'any 1929 apareix la primera edició que va traduir el Dr. Pere Farreras Sampere a partir de la 3a edició alemanya del llibre A. von Domarus "Grundriss der inneren Medizin". L'any 1940 es va publicar la segona edició en les mateixes condicions. Entre 1949 i 1967 van veure la llum les edicions 3a-7a amb la intervenció creixent del professor Pere Farreras Valentí. Des de la seva prematura desaparició, ocorreguda l'any 1968, vaig assumir la direcció de l'obra, en la qual col·laborava des de 1955, així el llibre es va convertir en Farreras-Rozman en les edicions 8a-13a. Aquest breu recull històric és prou demostratiu que l'obra que presentem, malgrat ser una traducció del castellà, és originària de Catalunya, perquè en aquest país va néixer, es va consolidar i es va convertir en un clàssic. D'altra banda, els experts en Història de la Medicina m'asseguren que no té precedents en la bibliografia catalana i que, per tant, el llibre pot ser considerat el primer text de medicina interna realitzat en català.*
>
> [Farreras-Rozman, 1997: XXI]

Per tot això, podem considerar que el grau d'especialització d'aquest text és alt.

El llibre s'estructura en 20 seccions[13] distribuï des en dos volums. En aquest estudi, utilitzarem la secció sobre "Malalties infeccioses" (pàg. 2209-

[12]En castellà també s'ha editat una versió de l'obra en CD-ROM.
[13]Les seccions que integren l'obra *Medicina interna* són les següents:

2586) dividida en nou parts[14], una de les quals està dedicada a les "Malalties produï des per Rickettsia" (pàg. 2393-2403), que farem servir per realitzar el buidatge manual. El nombre d'ocurrències del capítol sobre les malalties infeccioses és de 60.948 i l'apartat corresponent a les malalties infeccioses produï des per rickèttsia està format per 10.045 ocurrències.

El quadre següent recull, de manera abreujada, les principals característiques del corpus de comprovació:

I volum: Principis de la pràctica mèdica, Malalties de l'aparell digestiu, Gastroenterologia, Hepatologia, Cardiologia, Angiologia i hipertensió arterial, Pneumologia, Nefrologia, Reumatologia i malalties sistèmiques, Oncologia mèdica, Genètica mèdica, Geriatria, Dermatologia en medicina interna; II volum: Neurologia, Psiquiatria, Hematologia, Metabolisme i nutrició, Endocrinologia, **Malalties infeccioses**, Toxicologia, Malalties per agents físics i Immunologia.

[14]Les nou parts que integren la secció 17 *Malalties infeccioses* són les següents:
Generalitats, Malalties produï des per bacteris, **Malalties produïdes per *Rickettsia***, Malalties produï des per *Mycoplasma* i *Chlamydia*, Micosis i malalties produï des per fongs, Malalties produï des per paràsits, Malalties produï des per helmints, Malalties produï des per virus, Problemes especials en les malalties infeccioses.

CORPUS DE COMPROVACIÓ

TEMA: medicina, medicina interna, malalties infeccioses
CORPUS1: "Malalties infeccioses" (pàg. 2209-2586)
SUBCORPUS1.1: "Malalties produï des per *Rickettsia*" pàg. 2393-2403)
EMISSORS: especialistes
RECEPTORS: especialistes i estudiants universitaris
CANAL: escrit
FUNCIÓ: adquisició de coneixements, resolució de dubtes
NIVELL D'ESPECIALITZACIÓ: alt
NOMBRE D'OCURRÈNCIES TEXT1: 60.948
NOMBRE D'OCURRÈNCIES SUBTEXT1.1: 10.045

2.3 Buidatge terminològic del corpus de comprovació

Hem realitzat el buidatge terminològic del corpus textual emprant dos mètodes diferents:

- un buidatge manual
- un buidatge automàtic.

A través del buidatge manual volem verificar que totes les unitats del text considerades terminològiques responen, efectivament, a una de les estructures que vam recollir en el treball de recerca. En canvi, l'objectiu del buidatge automàtic és, d'una banda, establir la freqüència d'ús de les estructures de les UTP proposades en el treball de recerca com a pertinents i, de l'altra, disposar d'uns llistats "bruts" de sintagmes complexos que ens permetin estudiar quins tipus de paraules i combinacions, malgrat que no

siguin termes, tenen les mateixes estructures que els termes. L'anàlisi comparada d'aquests llistats amb els resultats del buidatge manual permetran identificar els problemes de reconeixement i de delimitació que sorgeixen quan s'aplica un SEACAT que funciona amb patrons formals de tipus morfosintàctic, així com també establir quins tipus d'unitats no reconeixen la majoria de SEACAT i per quins motius no les poden reconèixer.

El buidatge manual l'hem fet sobre el fragment del text "Malalties produït des per *Rickettsia*"; per al buidatge automàtic s'ha fet servir el text sencer de les "Malalties infeccioses".

2.3.1 Buidatge manual d'unitats terminològiques

2.3.1.1 Procés de buidatge manual

Com acabem de dir, l'única manera de comprovar si les estructures formals de les UTP proposades en el treball de recerca són completes és verificant-ho manualment, és a dir marcant d'un text especialitzat aquelles unitats que un usuari competent reconeix com a UT pròpies d'un camp.

Es tracta, doncs, del procediment clàssic que se segueix per extreure la terminologia d'un text:

> *La méthodologie de travail, en terminologie, se fonde essentiellement sur les corpus et sur leur analyse; c'est cette étape que nous nommons dépouillement. L'étape principale du dépouillement est le repérage de termes qui implique, lors de la lecture systématique du corpus, une identification manuelle des termes.*
>
> [Drouin, 1996: 45]

> *El buidatge terminològic és una operació que consisteix a extreure dels corpus de buidatge aquells segments que es consideren termes propis d'un camp d'especialitat del qual s'elabora la terminologia. (...) La primera acció*

> *d'aquesta fase de buidatge consisteix a reconèixer en els textos dels corpus de buidatge els segments lingüístics que corresponen a un concepte de l'àrea especialitzada i a delimitar-los. Un especialista en una determinada matèria reconeix amb més facilitat que no pas un llec en la matèria aquests segments terminològics que representen conceptes de la seva disciplina.*
>
> [Cabré, 1992: 273-316]

Com es desprèn d'aquesta darrera cita, una de les premisses per realitzar un buidatge terminològic és conèixer el contingut del tema sobre el qual es treballa, perquè les UT són les representacions lingüístiques dels conceptes especialitzats. Tradicionalment, s'ha considerat que el professional ideal per reconèixer les UT d'un text és l'especialista en el tema, que és qui posseeix la competència cognitiva i pragmàtica més alta sobre l'especialitat. En el cas del nostre corpus, els metges especialistes en medicina interna són, en principi, les persones més adients per realitzar el buidatge terminològic del text "Malalties infeccioses per Rickettsia". Per aquesta raó, vam encarregar inicialment a un especialista en medicina interna —el Dr. Pere Horta— que marqués totes les unitats del text que considerava pertinents en tant que unitats de significat especialitzat[15].

2.3.1.2 Resultats del buidatge manual

Del text sobre les malalties infeccioses produï des per rickèttsia, que està format de 10.045 ocurrències, l'especialista ha marcat 730 unitats diferents, de les quals:

[15] És significatiu remarcar que aquest especialista no té nocions específiques de lingüística i, en aquest sentit, el buidatge del text està guiat per l'ús professional i el sentit comú.

316 (43,28%) són unitats lingüístiques monolèxiques
375 (51,36%) són unitats lingüístiques polilèxiques
41 (5,61%) són unitats no lingüístiques

Així, ha considerat com a especialitzadament pertinents tant unitats lingüístiques com unitats no lingüístiques i, dins de les unitats lingüístiques, tant unitats monolèxiques com unitats polilèxiques.

De l'anàlisi de les unitats marcades podem extreure dos tipus de conclusions: unes sobre les UTP que l'especialista ha marcat, en comparació amb les UTP analitzades en el treball de recerca; i unes altres sobre els altres tipus d'unitats.

a. En relació amb les UTP

Comparats els resultats de les anàlisis de les UTP del text assenyalades per l'especialista amb els resultats de les anàlisis de les UTP del corpus lexicogràfic, observem que:

1. Totes les UTP del corpus textual seleccionades manualment per l'especialista responen a un dels esquemes morfosintàctics proposats en el treball de recerca:

NSAdj: 81,60%
NSPrep: 17,33%
NN: 1,06%

2. Les dues estructures més productives coincideixen en el corpus lexicogràfic i en el corpus textual (NSAdj i NSPrep suposen el

98,93% de les ocurrències del corpus lexicogràfic i el 98,32% de les ocurrències dels corpus textual).

3. Si despleguem analíticament aquestes estructures morfosintàctiques, ens adonem que en el corpus textual no hi ha tanta variació d'estructures morfològiques com en el lexicogràfic, fins i tot, en el corpus lexicogràfic, hi ha 16 seqüències morfològiques que no es donen en el corpus textual. L'explicació d'aquesta diferència la trobem en el fet que el corpus lexicogràfic de medicina general contenia nombroses estructures morfològiques que només es donaven en una o en dues ocurrències i que, per tant, és lògic que en un corpus textual de només 10.000 ocurrències sobre un tema específic de medicina no es donin.

Les estructures que hem recollit en el corpus textual són les següents:

Corpus textual: NSAdj	306 ocurrències	81,6 % de les UTP
NA[16]	258 ocurrències	68,8% de les UTP
NAA	30 ocurrències	8% de les UTP
NAAA	1 ocurrències	0,26% de les UTP
NA + símbol	1 ocurrències	0,26% de les UTP
N + sigla	1 ocurrències	0,26% de les UTP
N + símbol	6 ocurrències	1,6% de les UTP
sigla + A	2 ocurrències	0,53% de les UTP

Corpus textual: NSPrep sense article	45 ocurrències	12 % de les UTP
N de N	26 ocurrències	6,93% de les UTP
N de N_{propi}	12 ocurrències	3,2% de les UTP
N de NA	2 ocurrències	0,53% de les UTP
N de + símbol	2 ocurrències	0,53% de les UTP
N A de N_{propi}	1 ocurrències	0,26% de les UTP
N de N de N	2 ocurrències	0,53% de les UTP

Corpus textual: NSPrep amb article	17 ocurrències	4,53 % de les UTP
N de art N	11 ocurrències	2,93% de les UTP
N de art N A	2 ocurrències	0,53% de les UTP
N de art + sigla	2 ocurrències	0,53% de les UTP
N A de art $N_{propi}N_{propi}$	1 ocurrències	0,26% de les UTP
N A de art N N_{propi}	1 ocurrències	0,26% de les UTP
N de art N de art N_{propi}	1 ocurrències	0,26% de les UTP

[16] En aquest capítol hem preferit utilitzar les mateixes representacions que vam fer servir per al treball de recerca, és a dir estructures planes.

Corpus textual: NSPrep mixtos	3 ocurrències	0,8 % de les UTP
N en N de art N	1 ocurrències	0,26% de les UTP
N de art N de N_{propi}	2 ocurrències	0,53% de les UTP

Corpus textual: N N	4 ocurrències	1,06 % de les UTP
N N	3 ocurrències	0,8% de les UTP
N N_{propi}	1 ocurrències	0,26% de les UTP

En la taula següent es pot veure el nombre total de subestructures morfològiques per a cada un dels dos tipus de corpus:

	corpus lexicogràfic	**corpus textual**
NSAdj	9 subestructures	6 subestructures
NSPrep sense article	15 subestructures	6 subestructures
NSPrep amb article	10 subestructures	6 subestructures
NSPrep mixtos	2 subestructures	2 subestructura
NN	2 subestructures	2 subestructures
	38 subestructures diferents	22 subestructures diferents

Tot i aquesta dispersió, podem afirmar que les estructures efectivament productives són les mateixes en els dos tipus de corpus i que, a més, les subestructures del corpus textual estan incloses en les proposades a partir del lexicogràfic. Això és cert en tots els casos menys en alguns sintagmes en què algun dels constituents és una sigla o un símbol. En el corpus lexicogràfic ja havíem detectat la seqüència N símbol (*soca E, febre Q, linfòcits T, proteus OX-19,*

proteus OX-2, proteus OXK), però en el corpus textual, a més d'aquesta combinació, hem detectat les quatre següents[17]:

(1)

sigla + A: *DNA rickecttsià, pH àcid, pH baix*

N de art + sigla: *tècnica del PCR, afecció de l'SNC*

(2)

N de + símbol: *títol d'IgA, títol d'IgM, títol d'IgG*[18]

N A + símbol: *anticossos específics IgM*

4. La productivitat de les estructures també coincideix en els dos corpus, perquè, malgrat que en el corpus textual estan representades 23 estructures morfològiques diferents, només les cinc següents sobrepassen les deu ocurrències:

NA: 258 ocurrències (68,8% del total d'UTP)
N AA: 30 ocurrències (8% del total d'UTP)
N de N: 26 ocurrències (6,9% del total d'UTP)
N de N_{propi}: 12 ocurrències (3,2% del total d'UTP)
N de art N: 11 ocurrències (2,9% del total d'UTP)
representen el 89,8% de les UTP

5. Totes les UTP seleccionades per l'especialista són noms, dada que ratifica els resultats de l'anàlisi del corpus lexicogràfic. Això no

[17] Encara que aquestes estructures són molt poc productives —entre totes representen l'1,86% de les UTP del text— s'haurien de tenir en compte a l'hora de dissenyar un extractor automàtic de terminologia, perquè la sigla i el símbol són dos elements molt presents en el lèxic especialitzat en la mesura que proporcionen fluï desa al discurs científic. D'aquesta manera, el símbol, que és una unitat no lingüística, entra a formar part del sistema lingüístic a través de l'ús d'aquestes unitats lèxiques mixtes i, consegüentment, els llindars de la terminologia es comencen a eixamplar.

[18] *Títol d'immunoglobulina A, Títol d'immunoglobulina M, Títol d'immunoglobulina G.* En el tractat clàssic sobre fisiologia de Guyton (1987) s'explica el significat d'aquestes unitats lèxiques: "*Hay b clases generales d'anticuerpos llamados internacionalmente IgA, IgG, IgA, IgD i IgE. Ig significa inmunoglobulina, las otras cinco letras simplemente designan las diferentes clases de inmunoglobulinas.*"

obstant, hi ha altres unitats de significació especialitzada polilèxiques que no són nominals, tot i que l'especialista no les ha marcades.

6. El nombre de complements d'una UTP és normalment **un**, tant en el corpus lexicogràfic com en el textual:

	corpus lexicogràfic	corpus textual
1 complement	**94% de les UTP**	**89,05% de les UTP**
2 complements	4,7% de les UTP	10,93% de les UTP
3 complements	0,7% de les UTP	
4 complements	0,4% de les UTP	

Les unitats formades per més de dos complements, encara que n'existeixen, no són gens freqüents. La diferència entre els 4,7% sintagmes amb un nucli i dos complements i els 10,93% sintagmes amb un nucli i dos complements que té el corpus textual segurament està causada per les 30 ocurrències amb estructura morfològica NAA que ha assenyalat l'autor. Aquesta productivitat (la segona estructura més productiva de les UTP després de NA) està motivada per un objectiu de classificació exhaustiva. Generalment, el segon adjectiu indica o bé la localització —com a (3)— o bé la gravetat —com a (4)— o bé la temporalitat —com a (5):

(3)

infiltrat radiològic bilateral
infiltrat cel·lular perivascular
insuficiència vascular perifèrica
extravasació hemorràgica intersticial
col·lapse vascular perifèric

(4)

insuficiència cardíaca greu

hemorràgia digestiva alta

concentració inhibidora mínima

(5)

exantema maculós transitori

tractament antibiòtic precoç

insuficiència cardíaca progressiva[19]

7. Com vam comprovar en el corpus lexicogràfic, en català només poden formar part d'una UTP[20]:

    - els substantius[21]
    - els adjectius
    - la preposició *de* i en menor freqüència *per*[22]
    - els articles definits *el, la, els, les*[23].

8. Hem pogut confirmar que més del 99% dels adjectius de les UTP de medicina es posposen al nom; també hem confirmat que els únics adjectius que es poden col·locar davant del nom són els determinants numerals ordinals.

[19] Cap dels termes de (3), (4) i (5), marcats per l'especialista, estan documentats en el *Diccionari Enciclopèdic de Medicina* (1990).

[20] Hem de remarcar que, encara que en el corpus lexicogràfic vam documentar una ocurrència d'una unitat lèxica en què un dels seus constituents és la conjunció *i* (*vaccí de Calmette i Guérin*), en el corpus textual que hem treballat no n'hem trobat cap, cosa que no vol dir que en altres documents mèdics no puguem trobar aquesta estructura (que deu ser, però, molt poc productiva).

[21]Com ja havíem notat en el treball de doctorat, en medicina el segon nom de l'estructura N de N molt sovint és un nom propi. Moltes malalties o síndromes es denominen amb el nom d'una persona que, generalment, és l'individu que va identificar primer el trastorn. Aquest recurs lèxic es coneix amb el nom d'*epònim*. Per això, el nucli d'aquests sintagmes sol ser un mot genèric com *malaltia, trastorn, síndrome* o una malaltia concreta (*malaltia d'Alzheimer, malaltia d'Addison, malaltia de Brill-Zinsser, síndrome de Barret, cèl·lula de Kupffer, paràlisi de Bell*). Els *epònims* solen conviure amb un terme mèdic més descriptiu.

[22] La preposició *per* apareix en el nom d'alguns mètodes o tècniques: *control per estímuls, control per retroacció, biòpsia per agulla, biopsia per incisió, biòpsia per aspiració*, etc.

9. Com en el corpus lexicogràfic, la preposició per excel·lència que forma part d'una UTP en els textos de medicina és *de*. En aquest corpus hi ha només una seqüència que trenca aquesta homogeneï tzació dels resultats: *reacció en cadena de la polimerasa* (PCR). Aquesta unitat introdueix un sintagma adverbial, *en cadena*, a l'interior d'una UTP nominal. L'estructura d'aquesta seqüència amb nucli deverbal (N + SAdv + Sprep) no l'havíem documentat en el corpus lexicogràfic.

10. I, finalment, voldríem comentar que, encara que continua sent productiva la seqüència N de N_{propi}, ha disminuï t la seva freqüència. Aquesta disminució pot ser causada per dos motius: pel fet que els epònims solen tenir el seu corresponent nom mèdic descriptiu i les nomenclatures internacionals recomanen no utilitzar epònims en textos especialitzats. I també podria estar causat pel tema; en aquest cas, caldria diversificar i augmentar el corpus temàticament per comprovar la productivitat real d'aquesta seqüència[24].

b. En relació amb les unitats que no són UTP

Com ja havíem comentat a l'inici de l'apartat, si tenim en compte totes les unitats que ha marcat l'especialista com a pertinents, observem que hi ha unitats que no són UTP. Aquestes unitats rellevants des del punt de vista de l'especialista són les següents:

[23]A més, hem de tenir també present que els símbols, encara que amb una freqüència molt baixa, també són elements —encara que no lingüístics— que poden ser constituents d'una unitat terminològica.

[24] Avancem que, en el corpus textual de més de 60.000 ocurrències sobre les malalties infeccioses, aquesta seqüència és tan productiva com en el corpus lexicogràfic.

1. UT que estan formades per una sola paraula entesa com una seqüència entre dos blancs: les unitats terminològiques monolèxiques (UTM):

N: 272 ocurrències (86,07% de les UTM i 35,47% de les USE)

2. Altres unitats amb significat especialitzat (USE) que no són noms, sinó adjectius, verbs o adverbis. Són unitats que no són termes perquè no són referencials, però sí que tenen un significat especialitzat:

V: 23 ocurrències (7,27% de les USE monolèxiques i 3,15% de les USE)
A: 16 ocurrències (5,06% de les USE monolèxiques i 2,19% de les USE)
Adv: 4 ocurrències (1,26% de les USE monolèxiques i 1,08% de les USE)

   Aquesta diversificació de categories gramaticals és una de les raons per la qual l'objecte d'interès d'un extractor no pot reduir-se només a les UT.

3. Les sigles, que representen el 4,11% de les USE monolèxiques i l'1,78% del total d'USE del text remarcades per l'especialista:

(6)

ADA, DNA, IFD, IFI, PCR, TNF, RNA, etc.

És un fet que, explícitament, els extractors no tenen gairebé mai en compte les unitats monolèxiques i, en canvi, són unitats essencials en el discurs cientificotècnic. La raó per la qual les ignoren és, bàsicament, pel fet que, des del punt de vista de la forma, no presenten elements específics que permetin reconèixer-les. Aquesta raó té contraarguments almenys en dos casos:

1. La freqüència d'alguns sufixos específics, com *-itis* en medicina o *-ina* en bioquímica, permet identificar les UT d'un àmbit específic. En el text, entre els termes assenyalats per l'especialista, trobem una gran quantitat de termes amb aquests sufixos:

(7)

-itis:

artritis, cerebil·litis, conjuntivitis, endocarditis, glomerulomafritis, hepatitis, mastitis, meningitis, meningoencefalitis, miocarditis, monoartritis, nefritis, orquitis, osteomielitis, pancreatitis, pericarditis, pleuropericarditis, pneumonitis, poliartritis, polineuritis, poliradiculeneuritis, tromboflebitis i *vasculitis*.

(8)

-ina:

alcalina, ciprofloxecina, creatinina, doxiciclina, eritromicina, fibrina, josamicina, neoptorina, ofloxacina, perfloxacina, permetrina, rifampicina, roxitromicina i *tetraciclina*.

2. La presència d'alguns formants cultes que són propis d'un àmbit especialitzat determinat permet saber si una unitat és especialitzada:

(9)

hemo- (sang):

hemocultiu, hemòlisi, hemograma, hemorràgia, hemoglutinació[25].

(10)

hepato- (fetge):

hepatitis, hepatòlisi, hepatomegàlia, hepatosplenomegàlia.

(11)

-àlgia (dolor):

artàlgia, artromiàlgia, miàlgia.

(12)

-lògic (que pertany a la ciència de):

biològic, etiològic, epidemiològic, histològic, histopatològic, immunològic, neurològic, patològic, radiològic i *serològic.*

En els capítols següents veurem quina millora comportaria usar aquest tipus d'elements (sufixos, prefixos i formants cultes) per disminuir el silenci i el soroll que generen els SEACAT basats en patrons lingüístics.

A més de les unitats monolèxiques, l'especialista ha marcat seqüències que no formen part del llenguatge natural, però que tenen un valor especialitzat i que, segurament, seria interessant de tenir en compte per a molts tipus de treballs terminològics. Aquests elements són bàsicament dos:

1. Els noms científics en llatí de plantes o d'animals pertinents per a l'estudi de les malalties infeccioses, que representen el 4,93% del total d'unitats del text marcades com a especialitzadament pertinents per l'especialista.

[25]Molts dels exemples extrets del mateix text estan formats per més d'un formant culte; per citar només alguns casos: *hemòlisi (hemo: sang + lisi: dissolució), hemograma (hemo: sang + grama: dibuix), hemorràgia (hemo: sang + ragia: vessament).*

2. Els símbols que representen el 0,68% de les unitats assenyalades[26].

Per bé que havíem hipotetitzat que, en un principi, l'especialista no hauria de dubtar a l'hora de reconèixer les unitats que representen conceptes del seu àmbit professional, això no ha estat cert del tot, perquè encara que les detecti, s'adona que no totes les unitats tenen el mateix valor[27]. Els comentaris de l'especialista, en casos de dubte, plantegen problemes de fons de la terminologia, com ara: si existeix un **lèxic de** o és més pertinent parlar d'un **lèxic que s'usa en** i si existeixen **mots generalitzats** i **mots restrictius.** Les dues últimes afirmacions introdueixen el tema de la **fraseologia** i dels paratermes que acompanyen els termes en els usos especialitzats restringits.

En aquesta línia, l'especialista ha marcat amb un senyal especial 24 unitats que considerava que eren frases que es deien i s'escrivien i que, encara que "*es podien separar en diverses unitats, tenien un sentit juntes i dins del text*"[28].

[26] Ara bé, l'interessant pel que fa al símbols és quan aquests es combinen amb les paraules per formar unitats lèxiques complexes. Aquest fenomen és prou significatiu dins del conjunt de termes remarcats, ja que representa el 2,66% del total d'UTP.

[27] En paraules de l'especialista: "he seleccionat mots exclusivament de medicina, com *hepatitis* i *anèmia nemocrònica*; mots que s'usen en medicina, però que coneix tothom fins i tot un nen, com *mà, cara, cor, pulmó, malaltia, malalt, sang*; mots que són d'altres disciplines, però que també les utilitzem en medicina, com *cèl·lula* i *àcar*; mots que no formen part del lèxic tècnic de la medicina, però que els usem molt, com *cas, tolerar* i *contaminació*; he marcat també frases, perquè els seus mots per separat no volen dir res d'especial, però totes juntes tenen un significat mèdic".

[28] *àcar del ratolí, aglutinació en làtex, alteracions en el complement, arrels dels membres, augment al quàdruple dels títols, augment de la permeabilitat vascular, curs de l'afecció hepàtica, defenses de l'hoste, dèficit de glucosa-6-fosfatdeshidrogenasa, diagnòstic precoç de la malaltia, febre per picada de la paparra sud-africana, formes benignes de la malaltia, fragmentació de la paret vascular, hemòlisi associada a dèficit congènit de glucosa-6-fosfatdeshidrogenasa, índex de l'activitat de les paparres, lloc de la inoculació, mecanismes de defensa de l'hoste, mediadors de la inflamació, pneumonitis amb infiltrat alveolar, picada de paparra, predisposició de l'hoste, punt d'inoculació, vasculitis a l'SNC, vector principal de la malaltia.*

D'aquestes observacions, se'n desprenen dues idees que discutirem en el capítol cinquè:

- el lèxic especialitzat dels textos de medicina no està preestablert
- el lèxic especialitzat dels textos de medicina no és uniforme; l'ús sembla constatar que presenta una gran variació, variació que no mostren els diccionaris especialitzats.

2.3.2 Buidatge automàtic d'unitats terminològiques

2.3.2.1 Procés de buidatge automàtic

Per complementar les proves que demostren que els resultats del treball de recerca són vàlids, hem cregut oportú buidar terminològicament el text sencer sobre les malalties infeccioses —60.948 ocurrències— mitjançant mètodes automàtics d'explotació de corpus textuals.

Hem utilitzat dos programes d'explotació de textos de l'Institut de Lingüística Aplicada dissenyats en el projecte *Corpus textual especialitzat i plurilingüe*:

- un programa d'extracció de segments amb estructures determinades: CERPAT (CERcador de PATrons)

- un programa de delimitació de fronteres terminològiques: EXCAT1 (EXplorador de Candidats A Terme).

CERPAT és un programa que permet —entre d'altres possibilitats— extreure, a partir d'un text etiquetat, lematitzat i desambiguat morfològicament, ocurrències amb una determinada estructura morfològica. Aquest sistema ens ha permès extreure totes les UTP que

responien a una de les estructures morfològiques proposades en el treball de recerca amb les seva freqüència d'aparició.[29] L'objectiu primer de l'aplicació d'aquest programa al corpus és comprovar **la freqüència d'ús de les seqüències estructurals** proposades en el treball de recerca[30].

La majoria d'extractors de terminologia parteixen de patrons morfosintàctics predefinits per delimitar les UTP. Els sistemes creats per Blank (1995), Daille (1994), Jacquemin (1996), Heid i al. (1996), Frantzi i Ananiadou (1995), Dagan i Church (1994), Justeson i Katz (1995) són sistemes amb aquest tipus de plantejament, que es troben, però, amb una sèrie de limitacions ja assenyalades a Estopà, Vivaldi i Cabré (1998) i que descriurem més amb detall en aquest i en els propers capítols:

> *Els sistemes que utilitzen una sèrie de patrons morfosintàctics per identificar els termes complexos que, encara que responguin a la majoria d'unitats terminològiques, solen ser molt reduï ts i alhora massa poc restrictius; per a l'anglès AN i NN, per al francès NA i N prep N, de tal manera que hi ha alguns termes amb altres estructures que no es detecten mai. Els sistemes d'extracció basats en aquests tipus de tècniques lingüístiques generen massa soroll.*
>
> [Estopà, Vivaldi i Cabré, 1998:]

Amb els materials que genera CERPAT, es pot començar a estudiar el soroll que produeixen determinats patrons formals i, implícitament, el soroll que produeixen *grosso modo* els extractors que es basen en aquest tipus de patrons morfològics.

El segon programa, EXCAT1, és un explorador de terminologia que funciona amb coneixements lingüístics sobre allò que no pot ser mai un

[29] Aquesta eina, a més, permet extreure les seqüències amb el seu context i, d'aquesta manera, saber si una unitat terminològica està ben delimitada.

[30] Ara bé, ja que aquesta eina funciona a base de categories morfològiques i no sintàctiques, no ens permet d'extreure, directament, estructures sintàctiques del tipus NSAdj o NSPrep. Per detectar aquest tipus de sintagmes, es necessitaria un analitzador sintàctic (*parser*), però, en el moment de realitzar el buidatge, per a la llengua catalana encara no n'hi havia cap; tot i que, actualment, a l'Institut Universitari de Lingüística Aplicada de la Universitat Pompeu Fabra s'està treballant en l'elaboració d'un analitzador sintàctic per a la llengua catalana.

constituent d'una UTP. Així, aquest programa delimita les UT d'un text establint fronteres a partir d'elements que no poden formar part d'un terme (com ara relatius, demostratius, possessius, etc.). Els resultats de l'aplicació d'aquest programa són tots els segments que queden delimitats per dues fronteres de terme. El principal problema dels sistemes que funcionen amb coneixement negatiu del terme sol ser la delimitació de les UT, perquè es generen segments massa llargs —d'un nucli i tres i quatre complements— que no són en conjunt una UT, tot i que algunes de les seves "parts" poden ser-ho. Així doncs, els resultats obtinguts constitueixen també un material molt vàlid per explorar la segmentació i la delimitació de terminologia[31].

Consegüentment, els sistemes d'extracció semiautomàtica de terminologia basats en aquest tipus d'informació negativa també generen molt de soroll:

> *Si partim només d'elements que marquen frontera de terme, el soroll que genera el sistema és força alt. LEXTER és un dels mètodes que utilitza aquesta estratègia per detectar termes i, encara que té una cobertura del 95% (per tant, genera molt poc silenci), el soroll se situa entre el 40% i el 70%.*
>
> [Estopà, Vivaldi, Cabré, 1998: 57]

2.3.2.2 Resultats del buidatge automàtic

L'aplicació dels programes automàtics que acabem de presentar al text de buidatge, ens permet constatar que, de les 60.948 ocurrències que presenta el corpus de comprovació, 5.766 (9,33%) corresponen a possibles UTP. D'aquestes unitats complexes:

[31]EXCAT1 permet tornar en qualsevol moment al text original per observar com s'ha segmentat el text, extreure freqüències de les estructures utilitzades i extreure automàticament els patrons morfosintàctics de les seqüències generades.

1. Les estructures **més productives** coincideixen amb les estructures més freqüents del corpus lexicogràfic i amb les del buidatge manual del corpus textual 1.1:

	corpus lexicogràfic	buidatge manual corpus textual	buidatge automàtic corpus textual
N SAdj	64,05%	81,60%	68,26%
N SPrep	33,3%	17,33%	28,61%
	97,35%	**98,93%**	**96,87%**

Aquest 96,87% d'unitats es materialitzen principalment en les seqüències següents:

	buidatge automàtic corpus textual
NA	48,90%
N de N	11,71%
N de art N	8,65%
NAA	4,92%
N de N A	4,2%
N de art NA	3,02%
total	**81,4%**

El percentatge total del conjunt de les sis estructures morfològiques (81,4%) augmenta, lògicament, una vegada l'usuari ha fet la tria manual dels candidats a terme, ja que molts dels segments amb un nucli i dos complements extrets automàticament provoquen soroll i acaben reconvertint-se en una de les seqüències anteriors, sobretot en una de les tres primeres. Vegem com es desglossen terminològicament els sintagmes-candidats següents:

(13)

fase crònica de la malaltia: *fase crònica* i *malaltia*;
paràlisi espinal de les extremitats: *paràlisi espinal* i *extremitats*;
sistema nerviós dels primats: *sistema nerviós*;
parasitisme de les cèl·lules del sistema reticuloendotelial de l'hoste: *parasitisme, cèl·lules, sistema reticuloendotelial* i *hoste*;
etc.

És sabut que els extractors automàtics de terminologia basats en patrons formals proposen com a candidats a terme segments molt llargs formats per un nucli i més d'un complement, i quan se'n fa la revisió manual, la majoria es redueixen a UT monolèxiques, a UTP formades per un nucli i un sol complement o es rebutgen totalment. De fet, esporgar aquestes unitats suposa un treball manual important.

2. Encara que els percentatges entre el buidatge lexicogràfic, el buidatge textual manual i el textual automàtic, en general, coincideixen, si analitzem les ocurrències generades pels extractors, ens adonem que aquesta mateixa estructura no només correspon a UT, sinó també a seqüències discursives, unitats fraseològiques especialitzades i unitats fraseològiques o unitats lèxiques no especialitzades. I observem també que sovint només una part de l'estructura és terminològica i l'altra part constitueix una unitat discursiva. Això vol dir, com ja havíem intuït, que les estructures morfosintàctiques no són exclusives de les UTP.

Certament, de les 697 ocurrències del text que responen a l'estructura morfològica N de N —una de les estructures més

freqüent del textos especialitzats—, només 122 (17,50%) són terminològiques:

(14)

catèter de polietilè, catèter de Foley, cèl·lules de Mott, mal de coll, malaltia de Brill-Zinsser, medi de cultiu, síndrome de malabsorció, tinció de Giemsa, etc.

En algunes ocasions succeeix que el segment N de N no inclou cap UT:

(15)

condicions de treball, majoria dels casos, errors de judici, granets de sal, nombre de dies, punt de vista, unanimitat de criteris, etc.

El cas més normal, però, és que el segon nom —el nucli del sintagma preposicional— sigui una UT i el primer nom no ho sigui; en aquestes circumstàncies, a vegades la seqüència N de N pot esdevenir una unitat fraseològica especialitzada (UFE):

(16)

aparició de ***granulomes****, aparició de* ***petèquies****, casos de* ***rubèola,*** *concepte de* ***paludisme****, creixement de* ***papil·lomes,*** *episodi de* ***febre****, elevació de* ***transsaminases****, formes de* ***paludisme****, grau de* ***trombocitopènia****, mesures de* ***profilaxi****, percentatge de* ***seqüeles****, rigidesa de* ***nuca****, risc de* ***metàstasi****, tipus de* ***vacuna,*** etc.

Els exemples de (16) ens permeten observar que el nucli dels sintagmes subratllats és un nom format a partir d'un verb, que

conserva les característiques argumentals del verb del qual prové. La presència d'un nom deverbal en un sintagma nominal format per un nucli nominal i un complement preposicional és un indici fefaent de l'existència d'una unitat fraseològica especialitzada (UFE). Aquesta combinació es dóna encara amb molta més freqüència en les estructures en què la UT —l'argument— va precedida d'un article definit, el qual reforça aquesta idea d'unitat fraseològica, perquè la presència d'un article reforça la no lexicalització del segment:

(17)

absorció del ***sèrum***, *administració del* ***melarsoprol***, *administració de la* ***vacuna*** **,** *afectació de la* ***microcirculació***, *afectació de les* ***neurones*** ***motores***, *alentiment de la* ***circulació capil·lar,*** *alteracions de l'****aparell genital***, *alteracions del* ***llenguatge***, *complicacions de la* ***pneumòmia***, *complicació dels* ***picornavirus***, *prevenció de les* ***amebiasis***, *tractament de l'****aspergil·losi***, *tractament del* ***paludisme***, etc.

Finalment, podem trobar-nos també amb segments N de N que inclouen dues UT i que, malgrat això, el resultat global de la seqüència no és una altra UT, sinó una combinació recurrent en un domini temàtic concret:

(18)

biòpsia *de* ***pell***, ***cèl·lules*** *de* ***ronyó***, ***diagnòstic*** *de* ***toxoplasmosi, radiografia*** *de* ***tòrax***, ***tractament*** *de* ***medul·la***, ***injecció*** *de* ***sèrum***, ***lesió*** *de* ***cardiospasme***, etc.

3. Hem de confirmar la presència important d'epònims en el text (N de propi: 6,38%), encara que alguns no són UT. Així, al costat de les UT de (19) l'extractor també genera les unitats discursives de (20):

(19)

malaltia de Brill-Zinsser, malaltia de Chagas, malaltia de Sabouraud, antigen de Giardia, catèter de Foley, cèl·lula de Kuppfer, cèl·lula de Mott, cèl·lula de Türk, limfoma de Burkitt, limfoma de Hodgkin, miocardopatia de Chagas, etc.

(20)

illa de Nantucket, sud de Mèxic, nord de Moçambic, república de Myammar, escorxador de Brisbane, etc.

Els exemples de (20), òbviament, generen soroll ja que responen a l'estructura N de N_{propi}, però, en canvi, en cap cas es tracten d'UT. En aquest corpus, aproximadament el 15% dels candidats amb aquesta estructura s'haurien de rebutjar manualment.

4. Algunes seqüències estructurals amb tres i quatre complements que apareixen en el corpus lexicogràfic no surten en el corpus textual. De totes maneres, s'ha de remarcar que es tracta d'estructures hàpax del corpus lexicogràfic que corresponen, generalment, a estructures molt llargues, com les següents:

(21)

N A de art N de art N A A de art N (*beina sinoidal del tendó del múscul extensor cubital del carp*)

(22)

N A de art N de art N de art N (*beina sinoidal dels tendons dels dits de la mà*)

(23)

N de art N de art N A A de art N de art N (*beina dels tendons dels múscul flexor llarg dels dits del peu*)

I, a la inversa, amb els programes d'extracció automàtica de candidats a terme que funcionen amb coneixement negatiu ens trobem amb estructures candidates a terme que no havíem detectat en el buidatge lexicogràfic:

(24)

N A A de art N de art N A (*manifestacions clíniques procedents de l'afectació del parèmquima nerviós*)

(25)

N A de N de N A A (*formació consegüent d'immunocomplexos de pes molecular elevat*)

(26)

N de art N de N A de N (*causa de la persistència de títols significatius d'anticossos*)

Cap USE del text, però, presenta una estructura tan llarga com aquestes; perquè de fet no són una sola unitat, cosa que no significa que no incloguin alguna UT. En aquests casos, diem que el segment ha estat terminològicament mal delimitat i ha provocat soroll.

5. Tot i que sovint els segments més utilitzats en un text especialitzat són UT, hem comprovat que la freqüència no és un indici segur per saber si un segment és terminològic. Vegem una

mostra de les unitats que apareixen amb més freqüència en el corpus de comprovació en què podem observar aquesta barreja d'unitats:

(27)

NA

quadre ***clínic***	38 ocurrències
mononucleosi infecciosa	25 ocurrències
anatomia patològica	24 ocurrències
via oral	22 ocurrències
dolor abdominal	13 ocurrències
ésser humà	11 ocurrències
fase aguda	11 ocurrències
zona ***endèmica***	10 ocurrències

(28)

N de N

*període d'****incubació***	16 ocurrències
fixació de complement	16 ocurrències
mm de diàmetre	13 ocurrències
tractament d'elecció	6 ocurrències
pèrdua de pes	5 ocurrències
mm de llargada	5 ocurrències
esgarrapada de gat	5 ocurrències

(29)

N de art N

majoria dels casos	11 ocurrències
començament de la ***malaltia***	5 ocurrències
majoria dels ***pacients***	5 ocurrències
transmissió de la ***malaltia***	5 ocurrències

i el mateix passa amb les USE monolèxiques:

(30)

N

dia	158 ocurrències
pacient	149 ocurrències
cas	143 ocurrències
malaltia	136 ocurrències
infecció	134 ocurrències
vegada	98 ocurrències
setmana	90 ocurrències
any	80 ocurrències
tractament	73 ocurrències
febre	60 ocurrències
manera	54 ocurrències
nen	50 ocurrències
cosa	49 ocurrències
virus	47 ocurrències
home	45 ocurrències
espècie	42 ocurrències
dosi	41 ocurrències
diagnòstic	41 ocurrències
persona	33 ocurrències
etiologia	31 ocurrències
exantema	30 ocurrències
sang	29 ocurrències
fase	28 ocurrències
freqüència	28 ocurrències
anticossos	28 ocurrències
microorganisme	26 ocurrències

epidemiologia	25 ocurrències
lesió	25 ocurrències
patogènia	25 ocurrències
adult	24 ocurrències
cop	24 ocurrències
quist	24 ocurrències
excrement	23 ocurrències
fetge	21 ocurrències
mes	20 ocurrències
ou	20 ocurrències
pronòstic	20 ocurrències

(31)

A

freqüent	82 ocurrències
greu	36 ocurrències
elevat	26 ocurrències
eficaç	23 ocurrències
possible	20 ocurrències
capaç	20 ocurrències
responsable	19 ocurrències
útil	18 ocurrències
superior	18 ocurrències
rar	15 ocurrències
crònic	14 ocurrències
diferencial	13 ocurrències
inferior	13 ocurrències
intens	13 ocurrències
màxim	12 ocurrències
següent	12 ocurrències

Cap d'aquests adjectius aï lladament són especialitzats, per bé que alguns com *greu, superior, crònic, inferior, intens, màxim* poden integrar certes UT.

Així doncs, la freqüència tampoc no pot ser una prova única per demostrar el caràcter especialitzat d'una seqüència, sinó que només s'ha de tenir en compte com un element complementari.

2.4 Corpus de confirmació

El fet que el text de comprovació sigui un discurs altament especialitzat, escrit per especialistes i adreçat a especialistes no creiem que esbiaixi els resultats. Malgrat tot, per confirmar que els resultats són vàlids per a qualsevol tipus de text especialitzat, independentment de les seves característiques contextuals, hem repetit les mateixes proves de buidatge que havíem aplicat sobre el text de les malalties infeccioses a un altre text que presentés característiques oposades. Aquest text té la funció, doncs, d'actuar de corpus de control.

Així, sense moure'ns del marc de la medicina interna, hem seleccionat un altre tipus de malalties, les malalties respiratòries i, dins d'aquest àmbit, hem escollit un document sobre l'asma:

del Hoyo, J. (1985): *L'asma*. Barcelona, Ed. Proa.

L'asma és un llibre de caràcter divulgatiu, escrit per un metge —el Dr. Josep del Hoyo Calduch—, però adreçat al públic en general:

> *Els especialistes de l'aparell respiratori no podem deixar de sentir-nos satisfets per l'aparició d'un llibre que col·labora en la divulgació del coneixement de la malaltia asmàtica. (...) Un llibre de divulgació ha de ser clar i entenedor per a les persones a qui es dirigeix: lectors que no tenen un coneixement profund sobre el tema.*

[del Hoyo, 1985:7]

La funció bàsica d'aquest text és actuar de pont en la comunicació entre l'especialista i el gran públic (sobretot pacients d'asma i familiars) amb la finalitat que les persones asmàtiques puguin informar-se sobre la malaltia:

> *La lectura atenta d'aquest llibre que presentem aclarirà els dubtes que una persona asmàtica podria tenir sobre aquesta afecció i, sense dubte, l'ajudarà a orientar millor la convivència amb la seva malaltia.*
>
> [del Hoyo, 1985:9]

Per això, el grau d'especialització d'aquesta obra es pot considerar baix. El text de control és més curt que el corpus de comprovació —està format de 17.052 ocurrències. El quadre següent sintetitza les característiques contextuals bàsiques d'aquest document:

CORPUS DE CONTROL

TEMA: medicina, medicina interna, malalties respiratòries
CORPUS2: *L'asma*
EMISSORS: especialistes
RECEPTORS: públic general
CANAL: escrit
FUNCIÓ: superar obstacles cognitius
NIVELL D'ESPECIALITZACIÓ: baix
NOMBRE D'OCURRÈNCIES TEXT2: 17.052

Hem buidat el text sobre l'asma primer manualment i després automàticament amb la mateixa metodologia de buidatge que havíem utilitzat per al corpus textual de comprovació.

Per a l'extracció manual de les USE, al·legant que el text és divulgatiu, hem demanat a una metgessa de medicina general —la Dra. Faixedas— que realitzés el buidatge del text[32].

Per a l'extracció automàtica de les UTP, hem utilitzat, com en el corpus de comprovació, els programes EXCAT1 i CERPAT[33].

2.4.1 Resultats dels buidatges

Globalment, podem dir que els resultats que hem obtingut del buidatge del text sobre l'asma, tant del manual com de l'assistit, coincideixen amb els resultats dels buidatges realitzats sobre les malalties infeccioses, que alhora coincideixen amb els resultats del buidatge lexicogràfic. Les coincidències fan referència a:

1. el nombre d'unitats
2. les estructures formals
3. la categoria gramatical
4. el nombre de complements
5. la freqüència d'ús
6. altres característiques rellevants.

1. El document de *L'asma* està compost de 17.052 ocurrències, de les quals l'especialista només ha assenyalat **130** unitats diferents. Si comparem aquesta xifra amb la del corpus de comprovació (en què de 10.000 ocurrències, 730 de diferents havien estat assenyalades per l'especialista),

[32] En aquest cas, doncs, ens ha semblat que no era necessari recorre a un especialista en medicina interna o en malalties respiratòries per realitzar el buidatge del corpus ja que els conceptes que vehicula el text són fonamentals, per tant, des del punt de vista cognitiu, cap metge ha de tenir problemes per identificar les unitats especialitzades d'aquest tipus de textos.

[33] Per a més informació sobre aquests programes vegeu l'apartat anterior 2.3.

podem dir que la diferència indica el baix nivell d'especialització del text de control. En efecte, el text sobre l'asma és divulgatiu i això s'evidencia, entre d'altres coses, en el nombre tan baix d'UT que conté.

Els resultats del buidatge automàtic constaten que la seqüència NA equival a 323 ocurrències diferents; la seqüència N de N, a 129 ocurrències diferents, i la seqüència N de art N, a 267 ocurrències diferents. Aquesta gran diferència entre el nombre d'unitats marcades per l'especialista (recordem que només n'ha marcat 130 en total) i el nombre de candidats a terme generats pels extractors revela un percentatge significatiu de candidats rebutjables.

Si agafem les 323 ocurrències que presenten l'estructura NA i les comparem amb els resultats proposats per l'especialista, veiem que només 40 ocurrències coincideixen, cosa que significa que només el **12,38%** dels candidats a terme amb NA generats per EXCAT1 són pertinents; així, alguns segments no són pertinents, en d'altres només és terminològicament pertinent "una part" del segment, i d'altres unitats, encara que no siguin termes, són unitats especialitzades.

L'estructura no, és doncs, una prova definitiva per detectar les UTP i és finalment l'usuari qui davant una llista de candidats a terme, com la de (32), que responen a l'estructura NA, ha de discernir si algun element d'aquests segments (nucli i/o complement) és especialitzat, i si els segments sencers són UT, UFE o unitats discursives:

(32)

asma bronquial
aparell respiratori
aire lliure
alteració ***mental***

causa freqüent
***crisi** greu*
***crisi** intermitent*
element indispensable
factor extern
finalitat bàsica
*persones **asmàtiques***
població humana
***tractament** adequat*, etc.

El mateix passa si observem les unitats que responen a la seqüència morfosintàctica N de N, en què només 7 (5,42%) dels 129 sintagmes proposats com a candidats a terme van ser assenyalats per l'especialista. Per tant, la resta del 122 sintagmes que responen a l'estructura N de N —que han estat generats per l'extractor automàtic— presenten el mateix problema de reconeixement i sobretot de delimitació:

(33)

*administració de **medicaments***
*conjunt de **cèl·lules***
*dotzena de **substàncies***
emissió de pol·len
*grup de **malalties***
*mica de **tos***
període de temps
oli de colza
*tipus de **tractament***
tubs de goma
etc.

En seqüències menys productives, la possibilitat que es tracti d'una UT és molt inferior. Observem que per l'especialista cap o gairebé cap dels candidats a terme proposats pel programa informàtic és realment una UT. Així, dels candidats a terme amb un patró N de NA (29 ocurrències diferents) o N de art NA (82 ocurrències diferents), no hi ha cap segment complet que es pugui considerar una UT; tot i que aquesta seqüència pot incloure alguna UT monolèxica i/o polilèxica. És en aquests casos quan sorgeix el problema de delimitació del segment, com podem veure en els casos següents:

(34)

asma *de manera ràpida*
mena de ***reacció*** *defensiva*
*nom d'****asma intrínseca***
quantitat de ***gasos irritants***
realització de proves ***cutànies***

(35)

augment de la pol·lució atmosfèrica
*cas de l'****asma bronquial***
diagnosi *de l'****asma bronquial***
*empitjorament de l'****obstrucció bronquia****l*
famílies de les persones ***asmàtiques***
inflamació de les ***vies respiratòries***
part dels ***vasos sanguinis***
obstrucció *de les* ***vies respiratòries***
símptomes *de l'****asma crònica***
vida de la persona afectada, etc.

Totes aquestes observacions ens fan arribar a una primera conclusió: els sistemes d'extracció automàtica de candidats a terme basats en patrons

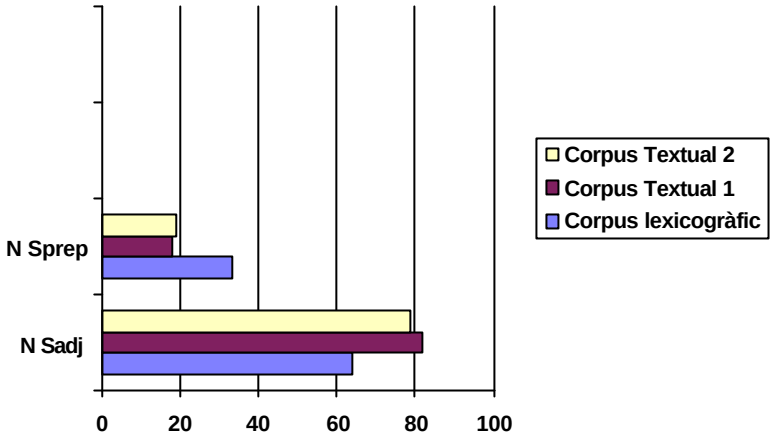
formals generen menys soroll com més alt és el nivell d'especialització d'un text; per bé que el soroll és present en tots dos tipus de text ja que les estructures morfosintàctiques que presenten les unitats no són exclusives de la terminologia.

2. Pel que a les estructures, de les 130 UT marcades per la metgessa com a pertinents, 62 són monolèxiques i 68 són polilèxiques. Per tant, tenint en compte que els SEACAT se centren només en el reconeixement de les UTP, deixarien sense reconèixer les 62 unitats monolèxiques.

Els textos divulgatius es caracteritzen per la seva claredat discursiva[34] i aquesta és la causa per la qual en aquest text no trobem ni sigles ni unitats d'altres codis no lingüístics, com ara els símbols; aquest tipus d'unitats, si bé serveixen per agilitar el discurs científic, són totalment opaques per a un llec en la matèria. També és significativa l'absència d'epònims.

Si ens centrem en les UTP, veiem que, com en el corpus textual de comprovació i en el corpus lexicogràfic, aquestes unitats es concentren prioritàriament en dues estructures:

[34] En el pròleg de *L'asma* [del Hoyo J., 1985: 7] es fa referència de manera molt clara a aquest tret que caracteritza el llibre "*Un llibre de divulgació ha de ser clar i entenedor per a les persones a qui es dirigeix —lectors que no tenen un coneixement profund sobre el tema— i alhora ha de tenir un bon nivell d'informació i rigor científic per tal que la persona que el llegeix pugui adquirir una idea sòlida del tema tractat.*"



Dintre de cada un d'aquests dos grans blocs d'estructures com menys especialitzat és un text, menys variació estructural presenta:

	corpus lexicogràfic	**corpus textual 1**	**corpus textual 2**
NSAdj	9 subestructures	6 subestructures	3 subestructures
NSPrep sense article	15 subestructures	6 subestructures	2 subestructures
NSPrep amb article	10 subestructures	6 subestructures	1 subestructura
NSPrep mixt	2 subestructures	2 subestructura	2 subestructures
NN	2 subestructures	2 subestructures	1 subestructura
total	**38**	**22**	**9**

Si busquem —a través dels programes informàtics utilitzats— totes les estructures que vam recollir en el treball de recerca, algunes no responen a cap ocurrència, com ara les seqüències N de N i N; N amb NA, VN, N de art N de art NA de art NA; d'altres sí, però després no corresponen a cap UTP, com és el cas de la seqüència N de art N de art NA malgrat que aparegui en 6 segments del text:

(36)

*causa de l'empitjorament de l'**asma crònica***

disminució del diàmetre de les ***vies respiratòries***
gravetat dels ***símptomes*** *de l'****asma crònica***
interior de les parets de les ***vies respiratòries***
majoria de les famílies de les persones ***asmàtiques***
part de l'arxiu de les ***substàncies innòcues.***

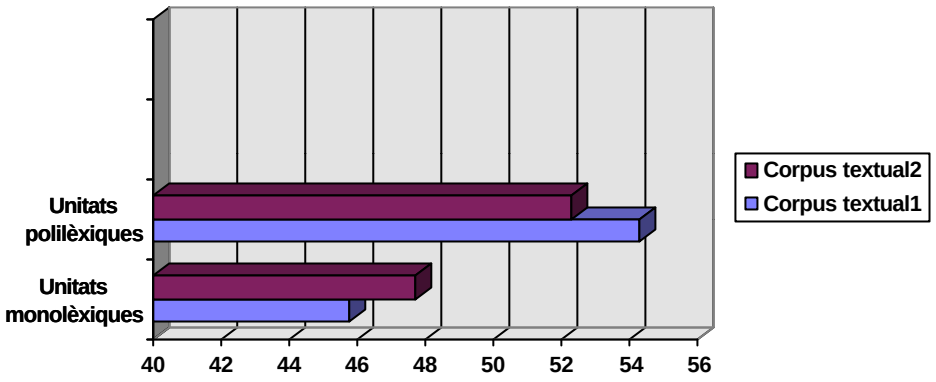
o de la seqüència N de NA de art NA amb 8 ocurrències, de les quals no n'hi ha cap que respongui a una UTP:

(37)

*factors desencadenants de l'****asma bronquial***
broncoconstricció *típica de l'****asma bronquia****l*
*factors desencadenants de l'****asma al·lèrgica***
fenòmens propis de la reacció ***immunitària***
origen ***al·lèrgic*** *de la* ***malaltia asmàtica***
*part fonamental de l'****aparell respiratori***
resposta inadequada del ***sistema immunitari***
símptomes *característics de l'****asma bronquial****.*

D'aquestes observacions podem extreure la segona conclusió: tot i que les estructures formals productives coincideixen en tots tres tipus de corpus (lexicogràfic, textual altament especialitzat, textual divulgatiu), com menys especialitzat sigui el corpus, les seves UT presentaran menys variació en l'estructura morfològica, cosa que no significa que es generi menys soroll, com hem vist a l'apartat anterior.

3. Funcionalment, la distribució en categories gramaticals coincideix en els dos corpus textuals:



Les unitats polilèxiques assenyalades són totes nominals, en canvi, les monolèxiques presenten categories gramaticals diverses:

USE monolèxiques: 62: 47,69%

noms: 58: 93,54%

verbs: 3: 4,83%

adjectius: 1: 1.61%

USE polilèxiques: 68: 52,30%

noms: 68: 100%

En aquest sentit, la tercera conclusió a què arribem és que el nombre d'UT del total d'USE és també molt elevat si tenim en compte que només són UT les USE nominals, i encara només les que tenen caràcter específic, tot i que en els dos corpus textuals hi ha USE d'altres categories gramaticals.

4. Quant al nombre de complements, hem constatat que totes les UTP marcades per l'especialista estan formades per un nucli i un sol complement (96,4%), excepte dues (3,5%). En canvi, quan apliquem EXCAT1 trobem que, encara que la majoria d'ocurrències continuen presentant un nucli i un sol complement, hi ha 60 estructures (que equivalen a 75 ocurrències diferents) amb un nucli i més de dos complements. Aquesta diversitat d'estructures morfològiques, però, no responen a cap UTP.

La quarta conclusió, doncs, a què ens permet arribar aquesta constatació és que, generalment, les UTP estan compostes d'un nucli i un sol complement, ja sigui un sintagma preposicional ja sigui, sobretot, un sintagma adjectiu.

5. La freqüència de les unitats en el corpus de control, com havíem també comprovat en el document sobre malalties infeccioses, no sempre és una garantia fiable per decidir que una unitat és terminològica. Per posar un exemple, hem comprovat que no totes les ocurrències que apareixen més de deu vegades en el corpus de l'asma són UT[35]. Aquesta constatació, una vegada més, demostra que les estructures morfològiques productives són sempre les mateixes —NA, N de N, N de art N— i que aquestes no són exclusives de les UT.

I el mateix problema —o més accentuat— ocorre amb el mots monolèxics[36]: de les 28 unitats monolèxiques que en el text de control s'usen

[35] (38) NA: ***asma bronquial*** (56 ocurrències), **v*ies respiratòries*** (42 ocurrències), ***obstrucció bronquial*** (37 ocurrències), *persones* ***asmàtiques*** (32 ocurrències), ***resposta immunitària*** (22 ocurrències), ***reacció al·lèrgica*** (12 ocurrències), ***aparell respiratori*** (11 ocurrències), ***asma al·lèrgica*** (11 ocurrències), *proves* ***cutànies*** (11 ocurrències), ***vasos sanguinis*** (11 ocurrències), *esforç* ***físic*** (10 ocurrències), *estat* ***asmàtic*** (10 ocurrències).
(39) N de N: ***diòxid de carboni*** (14 ocurrències).
(40) N de art N: *pas de l'aire* (12 ocurrències).
[36] Per donar algunes mostres, veiem les unitats amb més de 20 ocurrències:

aï lladament vint o més de vint vegades, només 13 (46,42%) són UT de medicina. Aquesta verificació dóna lloc a la cinquena conclusió: la freqüència no és un criteri totalment fiable per determinar el caràcter especialitzat d'una unitat, i menys en els textos de caràcter divulgatiu.

6. Finalment, com en el corpus textual de comprovació, hem verificat algunes constants que s'han repetit en tots tres tipus de corpus —lexicogràfic, textual molt especialitzat, textual divulgatiu— i que confirmen algunes prioritats morfològiques de les UTP:

- la preposició de totes les UTP amb sintagma preposicional és la preposició *de*
- l'únic determinant que pot formar part d'una UTP és l'article definit
- cap UTP presenta la conjunció *i*, i, per tant, si aquesta conjunció s'incorpora en els extractors que funcionen per patrons negatius es genera molt de soroll
- no se solen utilitzar en els textos divulgatius ni sigles, ni epònims, ni símbols, ni mots en llatí, unitats que, en canvi, s'utilitzen molt en els textos dirigits a especialistes.

(43) N: *persona* (109 ocurrències), ***substància*** (87 ocurrències), ***cas*** (84 ocurrències), *manera* (61 ocurrències), ***pulmó***(58 ocurrències), ***asma*** (58 ocurrències), *vegada* (48 ocurrències), ***organisme*** (48 ocurrències), ***via*** (47 ocurrències), ***reacció*** (45 ocurrències), ***obstrucció*** (45 ocurrències), ***medicament*** (40 ocurrències), *causa* (35 ocurrències), *aire* (33 ocurrències), ***al·lèrgia*** (32 ocurrències), ***metge*** (31 ocurrències), ***crisi*** (31 ocurrències), ***oxigen*** (27 ocurrències), *alteració* (27 ocurrències), ***especialista*** (23 ocurrències), ***símptoma*** (23 ocurrències), *mena* (23 ocurrències) *exemple* (23 ocurrències), ***sang*** (22 ocurrències), ***complicació*** (22 ocurrències), ***bronqui*** (21 ocurrències), *mecanisme* (21 ocurrències), *quantitat* (20 ocurrències).

2.5 Conclusions globals

De l'anàlisi de les dades de tots els buidatges dels dos corpus podem afirmar que, tal com havíem hipotetitzat, el fet de canviar l'objecte temàtic, els destinataris, les funcions i la llargada d'un corpus textual, no invalida els resultats del treball de recerca. Per tant, dins d'un mateix domini temàtic —les ciències de la salut—, els patrons morfosintàctics productius són útils independentment de les característiques del corpus d'aplicació. Aquesta afirmació significa que moltes estructures productives de les UTP dels textos de medicina són controlables. Malgrat això, en els textos divulgatius, les estructures morfosintàctiques es materialitzen en menys tipus de seqüències morfològiques i, alhora, aquestes provoquen més soroll que en els textos més especialitzats.

A mode de síntesi es poden formular cinc conclusions significatives:

1. Els patrons morfosintàctics productius de les UTP (més del 99%) són efectivament N SAdj i N SPrep, amb predomini del primer per sobre del segon, en tot tipus de text especialitzat.

2. Els sistemes d'extracció automàtica de candidats a terme basats en patrons formals funcionen pitjor com més baix és el nivell d'especialització d'un text, és a dir, aquests sistemes són més fiables si s'apliquen a corpus altament especialitzats.

3. Tot i que les estructures formals coincideixen en tots tres tipus de corpus (lexicogràfic, textual altament especialitzat, textual divulgatiu), com menys especialitzat és el corpus, les seves unitats terminològiques presenten menys variació morfològica, fet que no implica menys soroll.

4. La freqüència no és un factor determinant del caràcter especialitzat d'una unitat, i encara és menys fiable en els textos de caràcter divulgatiu.

5. En el buidatge manual dels dos corpus, divulgatiu i especialitzat, l'especialista ha considerat com a pertinents altres unitats que no són les UTP, unitats amb significat especialitzat que no tenen en compte els extractors automàtics.

2.6 Causes de les limitacions dels SEACAT basats en patrons estructurals

Si comparem les unitats que l'especialista ha assenyalat com a unitats pertinents d'un text de medicina amb les generades pels sistemes informàtics d'extracció de terminologia a partir del mateix text, constatem que, paradoxalment, no coincideixen. Els resultats divergeixen en dos sentits: d'una banda, l'especialista considera pertinents unitats lèxiques que els SEACAT no tenen en compte i, de l'altra, els SEACAT proposen com a candidats a terme segments que l'especialista no reconeix com a especialitzats.

Les principals limitacions del buidatge mitjançant un SEACAT respecte del buidatge manual de l'especialista són dues:

1. El silenci que comporten. Tot allò que un SEACAT hauria de reconèixer i que no reconeix.
2. El soroll que generen. Tot allò que el SEACAT ha reconegut i que no hauria d'haver generat.

I aquestes dues insuficiències —silenci i soroll— són causades pels elements següents:

1. l'exclusió d'USE pertinents
2. la inclusió d'unitats que no tenen valor especialitzat
3. la mala delimitació d'unitats
4. la manca d'explicitació de les relacions semàntiques que aquestes unitats mantenen dins del text.

2.6.1 Exclusió d'USE pertinents (silenci)

Tots els sistemes que funcionen a partir d'**esquemes morfosintàctics preestablerts,** ja sigui per patrons com per no-patrons, es caracteritzen perquè, en un principi, no generen gaire silenci. Això és lògic si tenim en compte que, com abans hem vist, totes les UTP proposades pels professionals responen a una de les estructures establertes i, per tant, teòricament, el nombre d'esquemes possibles és limitat. Malgrat aquesta important cobertura, els sistemes d'extracció basats en estructures no detecten totes les USE del text.

D'entrada, aquests programes se solen centrar exclusivament en les UTP. Consegüentment, podem considerar com un tipus especial de silenci totes les USE (i per tant també totes les UT) que apareixen en el text, però que no pertanyen al grup de les UTP, com són:

a) les USE monolèxiques[37]:

(42)

N: *arítmia, antibiòtic, artràlgia, cap, cefalea, malaltia, miàlgia, miocarditis, pacient, palmell,* etc.

(43)

A: *al·lèrgic, antimicrobià, cel·lular, clínic, endèmic, exantemàtic, histològic, infectat, papulós, serològic subclínic,* etc.

(44)

V: *administrar, descrostar, desinfectar, expirar, infectar, injectar, inocular, lesionar, prevenir, respirar, tolerar,* etc.

(45)

Adv: *biològicament, genèticament, histològicament, immunològicament, serològicament,* etc.

b) les sigles:

(46)

ADA	*CIM*	*CPK*
DDT	*DNA*	*IFD*
IFI	*LDH*	*PCR*
RNA	*SNC*	*VSG*

c) la fraseologia verbal i adverbial:

(47)

desenvolupar una (malaltia), patir (una malaltia), ser sensible a; en cadena, per via oral, per via intravenosa, etc.

I també podem considerar com un tipus especial de silenci aquelles unitats que, tot i que no pertanyen al llenguatge natural, tenen un interès especial en el textos temàticament especialitzats i que els SEACAT tampoc tenen en compte:

[37]Tots els exemples són extrets del text sobre les malalties infeccioses.

a) símbols:

(48)

IgG, IgM, IgA, Oc-19, Oc-12, Pg_1, Pg_2, soca E, etc.

b) mots i expressions llatins que formen part de nomenclatures:

(49)

Coixiella, Hyalomma, Ixodes, pediculus humanus corporis, Rickettsia, Rickettsia conorii, Rickettsia rickettsii, Rickettsia typhi, in vitro, in vivo, etc.

Però, fins i tot dins de les UTP —que són l'objecte d'extracció de la majoria de SEACAT—, trobem termes no detectats per una de les causes següents:

- errors en la desambiguació morfològica
- estructures que contenen més d'un terme (*superposició de termes*)
- termes amagats per anàfora.[38]

Els errors de desambiguació estan en relació amb la robustesa del desambiguador morfològic i/o estadístic que faci servir el sistema.

En canvi, les estructures que contenen més d'una USE autònoma suposen un problema menys fàcil de solucionar:

(50)

pneumonitis intersticial és un unitat terminològica, però també ho és *pneumonitis*

[38] Aquest tipus de silenci evidencia les relacions d'hiperonímia i meronímia que s'estableixen entre els termes que normalment són mots que comparteixen el mateix nucli [Bourigault, 1994].

(51)

febre tacada d'Israel és una unitat terminològica, però també ho són *febre tacada* i *febre.*

Però el problema de més difícil solució és, sens dubte, detectar les UTP, de les quals, per raons discursives, s'ha suprimit una part (el nucli o el complement). Alguns autors han anomenat aquest tipus de fenomen discursiu, *anàfora* [Kister, 1993]. Com assenyala Sager (1990, 1993), la variació contextual és molt freqüent en les formes lèxiques complexes:

> *Los nombres múltiples compuestos muchas veces aparecen en forma truncada en el texto y pueden ser idénticos en forma a sus hiperónimos que representan otros conceptos superordinados.*
>
> [Sager, 1993: 94]

Tot i que dóna lloc a estructures diferents; vegem-ne alguns exemples extrets del text sobre malalties infeccioses:

(52)

Des del punt de vista clínic, cal fer el diagnòstic diferencial amb ***malalties víriques i bacterianes.***

malalties víriques

malalties bacterianes

(53)

La malaltia per esgarrapada de gat és una ***malaltia ganglionar inflamatòria benigna, subaguda o crònica****, que gaureix espontàniament, amb fusió de les adenopaties o sense.*

malaltia ganglionar inflamatòria benigna

malaltia ganglionar inflamatòria subaguda

malaltia ganglionar inflamatòria crònica

(54)

*Amb certa freqüènica se'n presenta formes greus que inclouen diverses combinacions d'**insuficiència orgànica greu (neurològica, respiratòria, renal, cardíaca, hepàtica).***

insuficiència neurològica greu

insuficiència respiratòria greu

insuficiència renal greu

insuficiència cardíaca greu

insuficiència hepàtica greu

(55)

*En el 9% de casos sol haver-hi **conjuntivitis bilateral. Quan és unilateral** (...)*

conjuntivitis bilateral

conjuntivitis unilateral

2.6.2 Inclusió d'unitats no especialitzades (soroll)

El soroll produï t per les estructures proposades és una de les limitacions principals de tots els extractors de terminologia, tant dels basats en coneixement estadístic com dels basats en coneixement lingüístic. Entre les ocurrències proposades com a candidates a terme que responen, doncs, a una estructura preestablerta observem que:

1. Hi ha segments del discurs que presenten la mateixa estructura d'un terme, però que no ho són i, això, per diferents raons:

a. són segments que no són termes perquè són **UFE** —que per a alguns col·lectius són tan importants com

els termes i, per tant, també són pertinents de recollir[39]—,

b. són segments que no són termes perquè són **unitats discursives**

c. són segments que no són termes perquè són **UL o UF de la llengua comuna.**

Així, al costat de termes com

(57)

trastorn digestiu (NA), *vasos sanguinis* (NA), *sistema nerviós central* (NAA), *músculs del tòrax* (N de art N), *glàndules de la pell* (N de art N), *atac d'asma* (N de N), *úlcera d'estómac* (N de N), *proves de provocació bronquial* (N de NA), *síndrome de l'oli de colza* (N de art N de N)

hem trobat

(58)

tema complex (NA), *persona afectada* (NA), *tècniques noves* (NA), *nen asmàtic* (NA), *tos persistent* (NA), *exercici físic intens* (NAA), *augment de la temperatura* (N de art N), *causa de l'asma* (N de art N), *inici de la crisi* (N de art N), *profunditat de la respiració* (N de art N), *períodes de temps* (N de N), *episodis de durada curta* (N de NA), *importància dels canvis de clima* (N de art N de N), *aportació extraordinària d'oxigen* (NA de N), etc.

[39] El problema principal no és la no detecció d'aquestes unitats, sinó que els SEACAT

2. Hi ha segments del discurs proposats per un sistema com a candidats a terme en què només un o alguns dels seus constituents tenen valor especialitzat. Així, en el segment ***reactivitat*** *especial*[40], que respon a l'estructura NA, només el nucli és una USE; en canvi, en el segment *problemes d'****al·lèrgia*** (N de N) només el substantiu de l'extensió és terminològic; en el segment ***infeccions*** *de les* ***vies terminològiques*** (N de art NA), que podria ser una UT des del punt de vista de la seva estructura morfològica, només ho és el nucli i el sintagma nominal del complement, però no el segment sencer. Altres segments que exemplifiquen aquest fenomen, extrets dels corpus textuals, són els següents:

(59)

crisi *greu* (**N**A), *tipus* ***cardíac*** (N**A**), *augment de les* ***secrecions*** (N de art **N**), ***bronquis*** *de les persones* (**N** de art N), *causa d'****asma*** (N de **N**), *varietat d'****asma*** (N de **N**), *alteració de l'****aparell respiratori*** (N de art **NA**), ***símptomes*** *de l'****asma bronquial*** (**N** de art **NA**), ***dilatació*** *de la* ***paret del tòrax*** (**N** de art **N de art N**), *factors desencadenants de l'****asma al·lèrgica*** (N A de art **NA**), *acumulació notable de* ***cèl·lules*** (NA de **N**), etc.

Aquest problema està estretament relacionat amb la delimitació correcta de les UT que tractarem en l'apartat següent, perquè una delimitació deficient produeix, gairebé sempre, soroll.

---

clàssics presenten les UTP i les UFE sense cap mena de distinció.

En relació amb aquest mateix fenomen, cal que fem esment al fet que els extractors que funcionen a partir de fronteres de no-termes, com el LEXTER o l'EXCAT1, generen una gran quantitat d'estructures hàpax que no corresponen exactament a UT perquè són estructures massa llargues formades per un nucli i tres o més complements.

2.6.3 Imprecisió en la delimitació de les UT (soroll)

El problema de la delimitació de les unitats és alhora una de les limitacions dels SEACAT i una de les causes del soroll que generen aquests sistemes. Així, un candidat a terme pot ser que estigui mal delimitat perquè li falti un element (un o més d'un dels constituents) de la unitat o perquè li'n sobri una part.

En els mètodes basats en coneixement lingüístic, que generen realment molt poc silenci, el 99% de les causes d'una mala delimitació defectuosa d'una UT és la presència d'un element (o de més d'un) sobrer. Els exemples són múltiples i de fet, implícitament, ja hi hem fet referència en l'apartat anterior:

(60)

característiques ***biològiques***
període ***febril***
insuficiència renal
absència de ***leucòcits*** *de la* ***tripanosomiasi***
pacients *afectats d'****anèmia*** *de* ***cèl·lules falciformes***
efecte ***citopàtic*** *de les* ***cèl·lules falciformes***

[40]Hem remarcat en negreta la part del segment que creiem que té valor terminològic.

El problema es concentra sobretot en les estructures amb dos o tres complements, en què un dels constituents de la seqüència no forma part de la unitat. En aquests casos, sol ser el nucli de la seqüència el que no forma part de la UT, per bé que sovint aquest nucli, que sovint es tracta d'un nom deverbal (com veurem en el proper capítol), és un indici del començament d'una UFE nominals:

(61)

afectació de les ***cèl·lules hematoètiques***
*alteració de l'****aparell genital***
*complicació de l'****ambiasi intestinal***
secreció de les ***glàndules salivals***
augment de la xifra de ***proteïnes***
afectació de les ***neurones motores*** *supervivents*
multiplicació de ***virus poliomielítics salvatges***
observació microscòpica directa de ***plasmodi***
aplicació de la ***prova d'immunofluorescència indirecta***
causa de la infecció dels ***pacient immunodeficients***
proliferació exuberant de ***teixit de granulació***

Les estructures molt llargues —problema que presenten sobretot els sistemes que funcionen per informació negativa del no-terme— solen ser segments mal delimitats perquè els nuclis de les UTP en molt poques ocasions presenten més de dos complements. Així, estructures d'unitats generades per l'EXCAT1 com:

N de art N de N de art N de art N
N de N de art N A de N
N de N de N de art N de N
N A de N A A A

estan mal delimitades perquè globalment són unitats que no corresponen a cap UTP.

2.6.4 Ignorància de les relacions semàntiques entre les USE d'un text (silenci)

Les USE dins del text estan totes relacionades entre si. Aquestes relacions, que poden ser de molts tipus, no sempre són possibles d'identificar partint d'un diccionari, per tal com es manifesten de maneres variades.

La coordinació (mitjançant la conjunció *i* o *o*) uneix unitats que comparteixen el mateix esquema semàntic:

> *La coordination est la variation prototypique de l'isotopie sémantique entre termes car elle unit des termes dont les schémas interprétatifs sont semblables tant par la sélection sémantique des morphèmes que par le type de lien qui les unit.(...) La présence de deux termes au sein d'une coordination dont la conjonction est* et *ou* ou *dénote leur proximité sémantique.*
>
> [Jacquemin, 1997:6-7]

A vegades un segment entre parèntesis és significatiu d'un sinònim formal:

(62)

> *La immunoflorescència directa* ***(****IFD****)*** *permet de manera precoç i específica demostrar la presència de rickèttsies no en teixits infectats només, sinó també en les paparres.*

en què *immunoflorescència directa* és un sinònim, una variant formal, d'*IFD*.

D'altres vegades, els parèntesis indiquen l'equivalent en una altra llengua:

(63)

Per immunotransferència de Western ***(***Western Immunoblot***)*** *es poden identificar, de manera específica, les diverses espècies de* Rickettsi*a.*

(64)

El 1925, Pieri va descriure la taca negra ***(***tache noire***)*** *com una lesió d'inoculació de la malaltia.*

en què *Western Immunoblot* és un equivalent anglès d'*immunotransfèrència de Western* i *tache noire* és un equivalent francès de *taca negra* (i, segurament, indiquen també les llengües en les quals es va descobrir els fenòmens).

D'altres vegades, els mots entre parèntesis evidencien una relació meronímica *tipus de:*

(65)

Clínicament es manifesta com una malaltia aguda febril, amb cefalea, artromiàlgies, exantema macopapulós i lesió d'inoculació ***(****taca negra****)*** *en el 75% dels pacients.*

(66)

Aproximadament un terç dels pacients té manifestacions digestives ***(****vòmits, diarrea o dolor abdominal****)****.*

en què *taca negra* és un tipus de *lesió d'inoculació* i els *vòmits*, la *diarrea* i el *dolor abdominal* són tipus de *manifestacions digestives*.

La relació *tipus de* també es pot expressar d'altres maneres: compartint un mateix nucli:

(67)

> ***febre*** *Q,* ***febre*** *quintana,* ***febre*** *botonosa mediterrània i* ***febre*** *tacada de les Muntanyes Rocalloses*

en què tots els sintagmes són tipus de febre; o amb la conjunció *com*

(68)

> *De vegades, se n'han descrit altres manifestacions clíniques* ***com*** *hemorràgia digestiva altra, pancreatitis, síndrome monocleòtica (...)*

en què la conjunció *com* introdueix exemples de manifestacions clíniques.

Una altra d'aquestes relacions és la de causa-efecte que, en el text es pot transmetre a través dels verbs:

(69)

> *La lesió vascular* ***provoca*** *augment de la permeabilitat vascular que* ***afavoreix*** *l'extravasació de líquid intravascular i* ***pot causar*** *hipovolèmia i hipotensió.*

Totes aquestes relacions de semàntica lèxica són en el text i l'especialista no té cap problema ni per establir-les ni per identificar-les. De fet, són aquests tipus de lligams entre les unitats del text especialitzat els que ajuden el lector o l'oï dor a estructurar el sistema conceptual d'un àmbit concret del coneixement especialitzat. Els SEACAT, però, no solen utilitzar aquest tipus d'informació semàntica del text; tot i que alguns programes recullen la relació d'hiperonímia i cohiponímia a través de la coincidència de nuclis o de complements i els paradigmes hiperonímics per tal d'ajudar l'usuari a elaborar la selecció definitiva d'UT (aquest és el cas de LEXTER o de TERMINO); però, generalment, els SEACAT treballen

amb la UT descontextualitzada, sense posar èmfasi en les relacions contextuals[41].

2.7 Validesa dels patrons morfosintàctics

El propòsit bàsic d'aquest capítol era validar els resultats del treball de recerca presentat el 1996. Així, mitjançant els buidatges, manual i automàtic, de dos corpus textuals hem verificat que:

1. Un 98% de les UTP presenten un dels dos patrons sintàctics següents: N SAdj i NSPrep. El patró més usual és N SAdj materialitzat en la seqüència morfològica NA.

2. En les UTP que inclouen un sintagma preposicional, la preposició que introdueix aquest sintagma és en un 99% *de*, i en les UTP en les quals el sintagma preposicional està determinat, sempre trobem la presència de l'article definit.

Tot i que les proves de buidatge terminològic de textos han estat suficients per demostrar l'objectiu prioritari d'aquest capítol, els resultats dels buidatges, manual i automàtic, han servit també per qüestionar les

[41]Últimament, però, s'estan desenvolupant eines paral·leles que intenten explotar relacions semàntiques a partir dels resultats d'un extractor de terminologia. Així, per exemple, Zellig [Habert, 1998] és una eina d'adquisició de classes semàntiques *"qui permet de visualiser des regroupements de mots fondés sur le calcul des contextes qu'ils partagent. Intervint en aval d'un analyseur syntaxique, Zellig utilise l'ensemble des contextes syntaxiques de chaque mot pour extraire les relations de dépendance binaire (adjectif nom, nom prép nom, etc.) dans lesquelles il figure et calculer les proximités entre mots sur la base des dépendances élémentaires partagées. Le résultat est un graphe qui relie les mots associés par un nombre des contextes important, et qui fournit une première visualisation des systèmes de relations sémantiques caractéristiques du corpus étudié."*
[Fabre i Habert, 1998].

insuficiències dels SEACAT que es basen en l'ús de patrons, tant si són positius com negatius[42].

Així, els patrons morfosintàctics de les UTP són vàlids en el sentit que les USE polilèxiques responen a un dels patrons proposats, però, si aquests patrons són l'única base d'un SEACAT, els resultats confirmen que no són, ni prou exhaustius —hi ha moltes més unitats pertinents amb altres estructures que no es tenen en compte—, ni satisfactoris —no són exclusius de les UTP. Per això, la intervenció de l'humà amb competència cognitiva sobre un camp especialitzat és totalment necessària després de l'aplicació d'un extractor.

Aquestes insuficiències, que hem comentat en l'apartat anterior, es poden sintetitzar en els quatre punts següents:

1. El fet de no reconèixer les UTP del text que encara que presentin una estructura coneguda no es detecten com a candidates a terme i les USE del text que d'entrada ja no es consideren (silenci).
2. El fet de donar com a pertinents unitats que corresponen a una estructura establerta, però que no són UT (soroll).
3. La falta de delimitació correcta de les unitats.
4. La no detecció de les relacions semàntiques entre USE que es poden deduir del text.

[42] *Malgré ces traits marquants et réguliers, les termes sont des entrées lexicales complexes, sujettes aux homonymes, aux ambiguï tés structurales, aux variation morphologiques et syntaxiques, à la récupération dans d'autres entrées lexicales... Bref, à tout ce qui fait du langage un système reflétant la complexité de la cognition humaine et de ses moyens de communication. Ignorer la diversité des termes, leurs dimensions linguistiques et communicatives, les considérer comme des étiquettes stables et invariables est risquer de ne pas identifier, de ne pas en déceler les constantes évolution et le constant renouveau dont les occurrences textuelles sont des traces exploitables.*

[Jacquemin, 1997: 6]

2.8 Recapitulació

En aquest capítol hem verificat que els resultats que havíem defensat en el treball de recerca són generalitzables a les unitats en context. També hem vist que molts SEACAT es basen en patrons estructurals de les UTP (positius o negatius). Hem insistit en el problema que, tant si s'extreuen les UTP per patrons com per no-patrons, aquests sistemes generen molt de soroll i de silenci.

Des del punt de vista de l'extracció automàtica de terminologia, la pregunta que es planteja a partir d'aquestes observacions és: *Per què generen silenci els sistemes basats en patrons?* I, a continuació: *Com es pot filtrar el silenci?* I, paral·lelament: *Per què generen tant soroll els sistemes basats en patrons?* I, a continuació: *com es pot filtrar el soroll?*

Des d'un punt de vista metalingüístic, les preguntes serien les mateixes, però plantejades a l'inrevés: *Què és una UT? Són les UT les úniques que tenen valor especialitzat? Quins tipus d'unitats tenen "valor" especialitzat? Quines unitats interessa recollir per a un treball terminològic?*

I: *Com podem distingir les UTP de les unitats polilèxiques (UP) de la llengua comuna o de les unitats discursives lliures o de les UFE? Es poden caracteritzar lingüísticament? Quin valor tenen els factors extralingüístics?*

Des de la lingüística, s'ha intentat caracteritzar les UT i les UTP emfatitzant les seves peculiaritats morfològiques, sintàctiques, gramaticals, estructurals i en cap d'aquests casos s'ha arribat a trobar uns trets distintius definitius. És cert que s'ha aconseguit proposar unes determinades prioritzacions, però aquestes tendències no són ni exclusives ni suficients per caracteritzar les UT perquè no permeten diferenciar-les

d'altres unitats de la llengua. És possible que per arribar a caracteritzar-les s'hagi de treballar en els seus vessants semàntic i pragmàtic. I, en últim cas, s'hagi de recórrer a explicacions extralingüístiques:

> *Les unitats del llenguatge dotades de referència i incloses a la gramàtica del parlant no són d'entrada i en abstracte ni paraules ni termes, sinó unitats del lèxic de la gramàtica que, en virtut de les característiques de la situació comunicativa deixen que se seleccioni només un determinat feix de trets del conjunt que les descriu. Seguint aquest supòsit, tota unitat dotada de referència pot ser candidata a terme i a paraula, tot i que és clar que algunes unitats només es realitzen com a termes, per tal com sempre es fan servir per denominar el coneixement especialitzat.*
>
> [Cabré, 1998b]

En els dos propers capítols estudiarem i sistematitzarem quins tipus d'unitats produeixen silenci i quins tipus produeixen soroll quan s'aplica un SEACAT basat en patrons morfosintàctics.

En concret, en el capítol tercer buscarem els motius pels quals els SEACAT silencien certes unitats especialitzadament pertinents; i en el capítol quart estudiarem quin tipus de candidats a terme que generen els SEACAT no són UTP.

3. EL SILENCI: USE NO DETECTADES PER UN SEACAT

Hello, darkness, my old friend,
I've come to talk with your again.
Because a vision softly creeping,
Left its seeds while I was sleeping
And the vision that was planted in my brain.
Still remains within the sound of silence.

[The sound of silence, Paul Simon, 1964]

Qui pot obrir camins?
--qui coneix silencis pot obrir camins.

[El crit del silenci, Joan Croses, 1984]

La finalitat principal d'aquest capítol és estudiar per què els sistemes d'extracció automàtica de candidats a terme (SEACAT) basats en patrons morfosintàctics produeixen silenci. Primerament, ens proposem analitzar els tipus d'unitats dels textos especialitzats que un SEACAT hauria de detectar, però que voluntàriament o involuntàriament no ho fa; i explorarem les causes que podrien explicar aquestes limitacions.

Per dur a terme aquests objectius partirem dels resultats dels dos buidatges terminològics del corpus de comprovació que hem descrit en el capítol anterior.

La comparació dels resultats d'aquests dos buidatges ens permetrà saber quines són les unitats que, a parer de l'especialista, haurien d'haver estat detectades per un SEACAT i no ho han estat. En efecte, com hem vist anteriorment, els resultats dels dos buidatges mostren diferències en dos sentits: d'una banda, EXCAT1 no reconeix totes les unitats marcades per l'especialista i, de l'altra, no totes les unitats generades per EXCAT1 l'espescialista les considera unitats terminològiques (UT). Això vol dir que EXCAT1 produeix *silenci* i *soroll* respecte al buidatge manual.

Pel que fa al silenci, és significatiu el fet que gairebé el 50% de les unitats detectades per l'especialista **no estiguin incloses** en la llista de candidats a terme generada per EXCAT1. Aquesta constatació demostra que els extractors basats exclusivament en patrons morfosintàctics generen una gran quantitat de silenci i, per tant, no poden substituir completament el buidatge manual.

Sovint s'ha afirmat que els SEACAT que funcionen amb coneixement lingüístic generen un percentatge relativament baix de silenci (entre un 2% i un 10% segons el corpus). I és així, si el silenci es mesura en relació amb les UT complexes, que són l'únic objecte d'extracció d'aquest tipus de sistemes. Però si s'espera que un SEACAT sigui capaç de detectar no només les **UT complexes**, sinó totes les unitats especialitzades del text, el percentatge de silenci augmenta molt perquè els extractors no preveuen mecanismes per detectar els termes simples.

Limitant-nos inicialment a l'objecte d'extracció dels SEACAT, les unitats terminològiques polilèxiques (UTP), l'anàlisi de les unitats marcades per l'especialista permet diferenciar entre dos tipus de silenci relatius a l'objecte de buidatge del sistema d'extracció automàtica:

- **silenci intrínsec**
- **silenci extrínsec.**

Entenem per silenci intrínsec el conjunt de segments especialitzats que no detecta un SEACAT i que hauria de detectar perquè són unitats terminològiques polilèxiques (UTP). I entenem per silenci extrínsec el conjunt d'unitats especialitzades, que no són UTP, que un SEACAT ignora explícitament perquè no formen part dels seus objectius d'extracció. L'especialista, en canvi, les assenyala com a unitats especialitzades pertinents.

De les unitats marcades per l'especialista que EXCAT1 no ha detectat, unes —les més abundants— són unitats monolèxiques, i unes altres —les menys nombroses— són unitats polilèxiques. Des de la perspectiva d'un SEACAT, el silenci intrínsec és involuntari, i l'extrínsec, totalment voluntari. Consegüentment, des del punt de vista de les finalitats dels SEACAT tradicionals, només hem de valorar negativament el silenci intrínsec. En canvi, des de l'òptica de l'especialista, és molt més significatiu el silenci extrínsec que l'intrínsec.

A continuació, partint de l'anàlisi dels resultats dels buidatges del corpus, establirem quines són específicament les unitats especialitzades que els SEACAT exclouen i explorarem les causes dels dos tipus de silenci.

3.1 Silenci intrínsec

El silenci intrínsec afecta un percentatge molt reduï t d'UT: entre el 10% i el 5% de les UTP d'un text. Les causes d'exclusió d'una UTP poden ser tres:

- errors en la desambiguació morfològica
- superposició de termes (estructures que contenen més d'un terme)
- termes amagats (anàfores discursives).

La causa dels errors de desambiguació depenen de l'etiquetador, del conjunt d'etiquetes que aquest faci servir i de l'encert del desambiguador que el SEACAT utilitzi. Les altres dues menes de silenci intrínsec obeeixen a causes discursives i la importància que es pugui donar al fet que un extractor no les detecti depèn de les finalitats que es proposi el treball terminològic a realitzar.

Per exemple, si el corpus està format només per títols i resums de documents i la finalitat del SEACAT és la indexació d'aquests documents, tant els termes superposats com els amagats són dos problemes reals que es plantegen, atès que el text és tan restringit que difícilment un extractor podrà detectar aquest tipus de termes. En canvi, si el corpus està format per un recull ampli de textos complets i l'objectiu del treball és recollir la terminologia usada en aquests textos, l'extractor podrà detectar la majoria de termes superposats i amagats perquè en alguna part del text els trobarà sense desglossar, completament desenvolupats.

3.1.1 Errors en el processament del text

Els errors que es produeixen en la desambiguació morfosintàctica, com acabem d'exposar, estan condicionats per les característiques del desambiguador lingüístic i/o estadístic que es faci servir. Com més robust sigui el desambiguador menys errors hi haurà en el buidatge terminològic. Actualment, els desambiguadors lingüístics resolen al voltant del 75% de les ocurrències d'un text; el 25% restant se sol desambiguar mitjançant càlculs estadístics. La majoria dels desambiguadors estadístics actuals generen tant sols entre un 3% i un 5% d'error del total d'ocurrències d'un text. Hi ha desambiguadors estadístics molt potents que arriben a generar només un 2% d'error; en aquests casos, però, els corpus d'entrenament solen ser molt restrictius temàticament.

El percentatge d'error en la desambiguació d'un text pot comportar que una UTP no es reconegui perquè estigui etiquetada equivocadament. Per exemple, en el corpus analitzat, les UT *buit popliti* i *immnunocomplexos circulants* són unitats silenciades per EXCAT1 perquè, en aquests casos, els segments *buit* i *immunocomplexos* han estat processats com a adjectius i en aquest text són substantius. D'aquesta manera, les unitats *buit popliti* i *immnunocomplexos circulants*, que són UT formades per un nom i un

adjectiu, se silencien ja que en el corpus etiquetat responen a l'estructura A A que no és una estructura terminològica prevista[1].

3.1.2 Unitats superposades

Les unitats superposades són termes complexos, que alhora contenen unitats simples o complexes que són també termes:

(1)

pneumonitis intersticial és un UT, però també ho és *pneumonitis*

(2)

febre tacada d'Israel és una UT, però també ho són *febre tacada* i *febre.*

En la detecció automàtica dels termes superposats el problema sorgeix quan es vol recuperar totes aquestes UT que formen part d'altres UT més complexes, atès que alguns constituents de les UTP poden ser també UT aï llades:

(3)

prova de Weil-Félix, mètode serològic, estudi histopatològic, sistema mononuclear fagocític, anèmica moderada, adenopatia regional sensible, etc.

[1] Els errors de desambiguació també poden donar lloc a soroll, és a dir poden sobregenerar unitats que no són terminològiques. En el corpus trobem per exemple la frase "E*n les* ***proves analítiques sol*** *haver-hi anèmia*" en què la unitat *proves* ha estat etiquetada com a nom, *analítiques* com a adjectiu i *sol,* en lloc de verb, ha estat etiquetada com a adjectiu. Aquesta errada en la desambiguació del mot *sol* ha propiciat que la seqüència *proves analítiques sol* estigués formada per l'estructura terminològica NAA i per tant el sistema la proposa com a candidat a terme.

En els exemples de (3) veiem que *prova, mètode, estudi, sistema, moderada, regional* i *sensible* no són unitats de significació especialitzada (USE) aï lladament, però la combinació en què s'insereixen sí que és una USE de caràcter terminològic.

En un corpus gran i representatiu, aquest problema el pot resoldre un extractor basant-se en la freqüència d'ús de les unitats, ja que és molt possible que els termes inclosos en altres termes apareguin aï llats en altres parts del corpus i, d'aquesta manera, puguin ser recuperats. Així doncs, podem dir que els termes superposats només provoquen silenci "autèntic" quan els corpus són petits. En el nostre corpus d'anàlisi, el silenci provocat per unitats superposades que al llarg del text no apareixen aï lladament ni una sola vegada és només de 35 USE que representen un 4,5% de totes les unitats especialitzades marcades per l'especialista[2].

Tot i això, les USE incloses en d'altres de més complexes solen ser molt genèriques i per a determinats treballs terminològics no serien pertinents[3]:

(4)

anatomia patològica inclou ***anatomia***[4]
bacteri intracel·lular inclou *bacteri*
coagulació intracel·lular disseminada inclou *coagulació intracel·lular, coagulació disseminada* i ***coagulació***
dolor abdominal inclou ***dolor***
enzim hepàtic inclou *enzim*

[2] El problema és que la majoria de SEACAT no detecten les USE monolèxiques.
[3] Tots els termes simples superposats estan documentats tant en el *Diccionari Enciclopèdic de Medicina* (1990) com en el *Diccionari de la Llengua Catalana* (1993) d'Enciclopèdia Catalana, fet que, teòricament, ens indica que són termes d'abast general, encara que estan marcats amb una etiqueta temàtica relacionada amb el camp conceptual de la medicina.
[4]Hem remarcat en negreta els termes de significat genèric.

> *hemorràgia digestiva alta* inclou *hemorràgia digestiva, hemorràgia alta* i *hemorràgia*
> *meningitis asèptica* inclou *meningitis*
> *síndrome mononucleòtica* inclou ***síndrome***
> *trastorn de la visió* inclou ***trastorn*** i ***visió***

Si aquests termes els busquem en el corpus textual ampli sobre malalties infeccioses de més de 60.000 ocurrències[5], trobem que, dels trenta-cinc, vint-i-sis estan documentats aï lladament i nou no (*anatomia, alteració, fissió, nefritis, microscòpia, hemorràgia alta, coagulació disseminada, enzim específic, insuficiència hepàtica*). Aquesta disminució del silenci en corpus grans permet formular la hipòtesi que:

> *si s'augmenta la dimensió de determinats corpus, els termes superposats tendeixen a deixar de generar silenci perquè en algun lloc del text els seus components terminològics solen aparèixer aï lladament.*

3.1.3 USE amagades

El problema més complex pel que fa al silenci intrínsec és, sens dubte, detectar i extreure les USE de les quals, per raons discursives, s'ha suprimit el nucli o el complement (o una part del complement) i els components que resten es coordinen amb una conjunció copulativa o disjuntiva o bé usant altres recursos gramaticals com l'especificació, la comparació, la condició, l'atribució. Alguns autors [Kister, 1993] anomenen *anàfora* a aquest tipus d'escurçament discursiu. Des d'un altre punt de vista, aquestes unitats amagades es podrien considerar un tipus de variació discursiva.

[5] Per a una descripció més completa d'aquest corpus vegeu l'apartat 2.2 del capítol anterior.

La majoria d'anàfores discursives d'aquest tipus es produeixen per fer més àgil un text, ja que no es repeteix els segments que corresponen a la part compartida. En algunes situacions s'elideix el nucli, en d'altres el complement o una part del complement.

En el corpus d'anàlisi hem detectat 25 termes amagats per una anàfora discursiva, classificats en els grups següents:

14 USE amagades en frases o sintagmes coordinats
conjunció copulativa *i*: 9
conjunció disjuntiva *o*: 5
3 USE amagades en frases o sintagmes subespecificats
2 USE amagades en frases comparatives
1 USE amagada en frases condicionals
5 USE amagades en frases predicatives

Vegem exemples d'anàfores extrets del text sobre malalties infeccioses per Rickèttsia:

USE amagades en frases o sintagmes coordinats amb una conjunció copulativa::

(5)

Des del punt de vista clínic, cal fer el diagnòstic diferencial amb ***malalties víriques i bacterianes***.

malalties víriques

malalties bacterianes

(6)

A diferència d'altres febres tacades, R. akari no aglutina el ***Proteus OX-19, OXK i OX-2****.*

proteus OX-19

proteus OXK

proteus OX-2

(7)

La diferència serològica entre el ***tifus murí*** *i l'****epidèmic*** *s'ha indicat en el diagnòstic d'aquest darrer.*

tifus murí

tifus epidèmic

(8)

Prop del 50% dels pacients presenten ***anèmia normocítica i normocròmica moderades****.*

anèmia normocítica (moderada)

anèmia normocròmica (moderada)

USE amagades en frases o sintagmes coordinats amb una conjunció disjuntiva

(9)

tifus murí o endèmic[6]

tifus murí

tifus endèmic

(10)

La malaltia per esgarrapada de gat és una ***malaltia ganglionar inflamatòria benigna, subaguda o crònica****, que guareix espontàniament, amb fusió de les adenopaties o sense.*

malaltia ganglionar inflamatòria benigna

malaltia ganglionar inflamatòria subaguda
malaltia ganglionar inflamatòria crònica

USE amagades en sintagmes especificatius

(11)

*Amb certa freqüència (7,5%) se'n presenta formes greus que inclouen diverses combinacions d'**insuficiència orgànica greu (neurològica, respiratòria, renal, cardíaca, hepàtica)**.*

insuficiència neurològica greu
insuficiència respiratòria greu
insuficiència renal greu
insuficiència cardíaca greu
insuficiència hepàtica greu

(12)

*S'estableix sobre **vàlvules** natives prèviament malaltes. (**mitral o aòrtica**).*

vàlvula mitral
vàlvula aòrtica

USE amagades frases comparatives

(13)

*La febre tacada de les Muntanyes Rocalloses respon al tractament amb tetraciclines i cloranfenicol tant **per via oral** com **intravenosa**.*

per via oral
per via intravenosa

---

[6] Només documentat en el títol d'un apartat. En aquest cas, s'hi afegeix el problema que

unitat terminològica en frases condicionals

(14)

En el 9% de casos sol haver-hi ***conjuntivitis bilateral****. Quan és* ***unilateral*** *i va acompanyada d'afectació ganglionar regional intensa (síndrome oculoglandular) és molt probable que aquesta sigui la porta d'entrada de la infecció.*

conjuntivitis bilateral

conjuntivitis unilateral

USE amagades en frase predicatives

(15)

La febre botonosa és la rickettsiosi exantemàtica més freqüent als paï sos de la conca de la Mediterrània on la ***malaltia*** *és* ***endèmica****.*

malaltia endèmica

(16)

La ***infecció*** *és* ***asipmtomàtica*** *o* ***subclínica****.*

infecció asimptomàtica

infecció subclínica

(17)

El ***diagnòstic*** *és* ***serològic****.*

diagnòstic serològic

Com es pot deduir dels exemples, en la majoria de casos el constituent que s'anaforitza és el nucli; només en tres casos, dels vint-i-cinc que trobem en el corpus, s'ha anaforitzat el complement.

---

els dos termes coordinats són sinònims.

Un cas especialment delicat el plantegen les frases atributives del tipus *l'**exantema** és **maculós,*** que correspon a una frase desenvolupada com *l'extantema és un exantema maculós.* L'especialista ha marcat *exantema* i *maculós* com a pertinents, perquè diu que *exantema maculós* és UT. Aquesta afirmació es fonamenta en els arguments següents:

- el segment *exantema maculós* té caràcter referencial, és un tipus d'*exantema* (cosa que indica que l'adjectiu en un principi qualificatiu ha adquirit un valor relacional classificador)
- en el paradigma semàntic dels *exantemes,* hi trobem *exantema petequial, exantema papulós, exantema vesicular, exantema maculopapulós,* etc. (cosa que demostra que hi ha una taxonomia establerta)
- tots els tipus d'exantema, formalment, s'han format per un nucli que és l'hiperònim (*exantema*) i un adjectiu que localitza el nucli o que el caracteritza des del punt de vista de la forma.

Així, podríem considerar que *exantema maculós* és UT que en el text que analitzem apareix "amagada".

En el text analitzat, de dimensions molt reduï des, només cinc de les vint unitats anaforitzades apareixen aï llades. De les altres vint unitats amagades que no hi apareixen, catorze les hem trobades en el corpus més ampli de més de 60.000 ocurrències (això significa que un 70% de les unitats amagades es podrien recuperar). Aquest fet indica que la hipòtesi que hem formulat en l'apartat 3.1.2, segons la qual si s'augmenta el nombre d'ocurrències del corpus de buidatge, els termes amagats tendeixen a "aparèixer", resulta ben orientada.

3.2 Silenci extrínsec

Com s'ha explicat en el capítol 1, els sistemes d'extracció automàtica de terminologia de base lingüística dissenyats fins ara[7] no s'han proposat reconèixer les unitats que no són polilèxiques nominals. Així, els SEACAT s'han centrat exclusivament en la detecció de les UTP dels textos, i si reconeixen alguna unitat monolèxica és perquè aquesta forma part d'una de polilèxica. Els autors d'aquest tipus de sistemes justifiquen explícitament aquesta restricció amb arguments com els següents:

1. La majoria de les UT d'un àmbit d'especialització són unitats sintagmàtiques i, per això, és irrellevant de complicar el sistema fent-lo reconèixer els termes simples.

2. Formalment, les unitats monolèxiques especialitzades i les generals no es diferencien i, per això, és impossible discriminar entre les unitats monolèxiques especialitzades i les generals.

3. Les unitats monolèxiques són molt més polisèmiques que les polilèxiques i, per això, treballar amb heurístiques semàntiques que facin referència als termes monolèxics és molt més complicat que no pas introduir coneixement semàntic en les UTP.

En conjunt, és cert que, des d'un punt de vista morfosintàctic, les UTP són més fàcils de detectar que les unitats monolèxiques, atès que presenten una estructura morfosintàctica explícita que és controlable. Però tot i això els tres arguments adduï ts s'han de matisar.

En relació al primer argument, els autors afirmen que el 80% de les terminologies estan formades per UTP perquè parteixen de lèxics establerts, però si partissin de l'ús real dels termes en els textos, aquestes afirmacions ja no serien vàlides. Els percentatges d'unitats monolèxiques i

[7] Vegeu l'anàlisi dels principals SEACAT presentat en els apartats 1.5 i 1.6 del capítol primer.

d'unitats polilèxiques en els dos corpus textuals analitzats són molt similars: en el corpus textual1 (més especialitzat), les unitats polilèxiques representen el 54,26% de les USE marcades pels especialistes i el 45,73% correspon a les unitats monolèxiques; en el corpus textual2 (divulgatiu), les unitats polilèxiques comporten el 52,30% i les unitats monolèxiques el 47,69%. Aquestes xifres indiquen que el conjunt d'unitats monolèxiques especialitzades d'un text temàticament especialitzat no es pot menystenir i, per tant, un bon SEACAT també les hauria de poder extreure.

Els tres arguments també són discutibles perquè, si bé les UT, que són sempre nominals, són les USE més freqüents dels textos especialitzats, els verbs, els adjectius i els adverbis amb un ús temàticament especialitzat també són rellevants en aquests textos.

Per això, creiem que aquestes afirmacions es podrien reconsiderar perquè, encara que sigui cert que les unitats lèxiques simples són força idiosincràtiques i molt polisèmiques (i, consegüentment, és molt difícil discriminar lingüísticament quan una unitat simple s'utilitza amb un sentit especialitzat o general), dins de les unitats monolèxiques hi ha diferents classes de paraules —derivades, compostes, abreujades— que presenten algunes peculiaritats formals en què els SEACAT es podrien basar per detectar una bona part dels termes monolèxics.

L'especialista que no té estudis de lingüística allò que detecta són les unitats de comunicació (lingüístiques i no lingüístiques) que utilitza amb un significat precís en l'àmbit mèdic, tant si aquestes unitats són molt especialitzades com si ho són més poc. Així, des d'un punt de vista comunicatiu, usa tant *fetge*, *inflamació de fetge*, com *hepatitis;* utilitza tant *glàndules salivals* i *secreció de les glàndules salivals* com *PCR*, *Rickettsiaceae* i PG_1. Per a ell, totes aquestes unitats comparteixen la

característica de ser unitats de significació especialitzada, és a dir de ser USE.[8]

En aquest treball, el marc d'interès no és la terminologia, sinó les **unitats sígniques especialitzades que apareixen en els textos de l'àmbit mèdic,** és a dir, tant les unitats lingüístiques com les unitats que no formen part del sistema lingüístic i l'objectiu principal de l'estudi del silenci extrínsec és trobar elements, sobretot lingüístics, que contribueixin a augmentar la cobertura dels SEACAT actuals. Per això, hem distribuï t aquest estudi en apartats que equivalen als diversos tipus d'unitats que EXCAT1 ha silenciat del text1 en relació amb les unitats marcades per l'especialista, a saber:

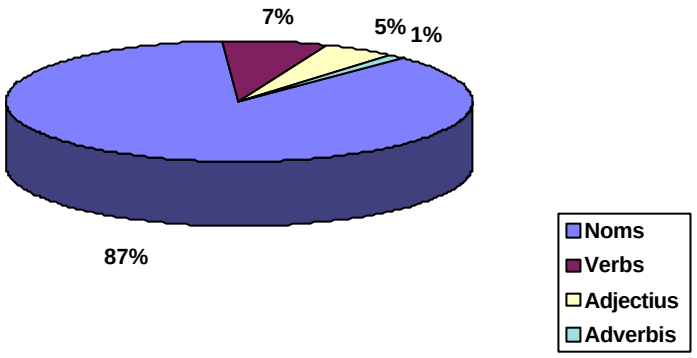
- unitats lingüístiques:
  - unitats monolèxiques (simples, derivades, compostes)
    - nominals
    - verbals
    - adjectivals
    - adverbials
  - sigles
  - *unitats fraseològiques verbals*
- unitats no lingüístiques:
  - símbols
  - noms llatins

Aquestes unitats representen el 48% de les unitats especialitzades del corpus marcades per l'especialista[9].

[8]D'altra banda, recordem que l'especialista intueix que el valor pragmàtic d'aquestes unitats no és homogeni: unes són més genèriques, unes altres les coneix tothom, d'altres

3.2.1 USE monolèxiques

Les USE monolèxiques analitzades representen el 41% de les unitats seleccionades per l'especialista com a pertinents. Segons la categoria gramatical de les USE monolèxiques, aquest percentatge es distribueix de la manera següent:



En aquest apartat ens referirem només als noms, verbs, adjectius i adverbis que compleixin les dues condicions següents:

- apareixen aï lladament en el text de buidatge
- l'especialista els ha marcat com a pertinents.

I no parlarem, en canvi, de les unitats que, encara que siguin especialitzades, en el text que analitzem només formen part d'una unitat polilèxica (*insuficiència cardíaca greu, líquid intravascular, malaltia*

---

s'usen exclusivament en medicina, d'altres només tenen sentit en context, etc.

[9] Hi ha d'altres tipus d'unitats lingüístiques especialitzades temàticament (com ara acrònims, onomatopeies, manlleus adaptats, compostos patronímics, unitats fraseològiques adverbials, etc. [Cabré, 1992], [Riera, 1998]) que, en aquest estudi, no tractarem, atès que ens limitarem a les unitats que l'especialista ha destacat del corpus de buidatge.

endèmica, etc.), atès que aquest tipus de noms i de modificadors els estudiarem en el proper capítol dedicat al soroll.

3.2.1.1 USE monolèxiques nominals

En el corpus analitzat l'especialista ha trobat 272 USE monolèxiques nominals o UT (que representen el 86,07% de les USE monolèxiques del corpus i el 35,47% de totes les USE marcades) que l'extractor automàtic no ha detectat.

Partim de la idea que, si agrupem les UT seleccionades en camps lexicosemàntics, serà més fàcil detectar-ne la sistematicitat, formal i semàntica, perquè, dins de cada camp lèxic, la forma i el significat de les UT tendeixen a ser regulars. Si això fos cert, podríem arribar a demostrar que els models de denominació s'elaboren sobre la base de la selecció sistemàtica de determinades propietats i característiques:

> *En alguns camps d'especialitat es dóna una gran regularitat en l'ús de determinats formants, que no s'ha de transgredir quan es formen noves unitats terminològiques. Així, per exemple, en el camp de la indústria en general:*
>
> - *el sufix -atge és el recurs més freqüent per indicar operacions tècniques*
> - *el sufix -ada és el més freqüent per indicar el resultat d'un procés*
> - *el sufix -ció és el més freqüent per indicar una acció*
> - *-el sufix -dor (a) sol servir en general per designar els aparells que fan una determinada acció.*
>
> [Cabré, 1992: 180]

Encara que sabem—com adverteix Sager (1993: 99)— que moltes vegades "*secuencias de términos bastante regulares se interrumpen por formaciones irregulares que pueden ser explicables históricamente, pero que claramente alteran lo que, de otra forma, podría parecer un conjunto de términos bien estructurado*".

En l'àmbit de les ciències de la salut, com en tots els àmbits especialitzats, les USE monolèxiques es poden classificar en diversos camps semàntics. En el nostre treball hem agrupat les USE monolèxiques que l'especialista havia marcat en el text en classes molt generals:

a. Patologies (malalties, manifestacions biològiques de les malalties, conseqüències de les malalties):

acufen, alcoholisme, anèmia, aneurisma, anorèxia, apirèxia, arítmia, artràlgia, artritis, artromiàlgia, buf, bradicàrdia, brucel·losi, cefalea, cerebil·litis, cianosi, conjuntivitis, convalescència, diarrea, edema, ehrlichiosi, endocarditis, escara, fagolisosoma, febre, fotofòbia, gangrena, hepatitis, hepatomegàlia, hepatosplenomegàlia, hipergammaglobulinèmia, hipertròfia, hipoalbuminèmia, hipocomplementèmia, hipotensió, hipovolèmia, icterícia, immunodeficiència, leptopirosi, leucocitosi, leucopènia, mareig, mastitis, miàlgia, microhematúria, microtromba, microtrombosi, miï tis, miocarditis, monoartritis, nàusea, necrosi, odinofàgia, oligúria, pàpula, pancreatitis, pericarditis, plaquetopènia, pleuropericarditis, pneumònia, pneumonitis, poliartritis, polineuritis, poliradiculoneuritis, proteï núria, pústula, rickettsèmia, rickettsiosi, rubèola, somnolència, tifus, tos, toxoplasmosi, trombocitopènia, tromboflebitis, úlcera, uveï tis, valvulopatia, varicel·la, vasculitis, vòmit, xarampió, zoonosi, xoc.

b. Components del cos humà (localització de les malalties en el cos humà)

aixella, articulació, cap, cara, cervell, conjuntiva, cor, endocardi, engonal, epidermis, extremitats, fetge, intestí, melsa, mucosa, múscul, palmell, paròtide, pell, planta, pubis, pulmó, ronyó, tronc, vas, vesícula, endoteli, nòdul, fibra, fibroblast, histiòcit, anticòs, antigen, hipoproteï na, immunocomplex, immunogen, limfòcit, mastòcit, patogen, sèrum, sang, toxina, cèl·lula, citoplasma, fibrosoma, macròfag, plaqueta.

c. Organismes animals i microorganismes (agents externs de les malalties)

àcar, artròpode, coccobacil, comensal, hoste, ixòdid, larva, microorganisme, nimfa, reserva, rickèttsia, soca, vector.

d.. Proves i exàmens (exploració i diagnòstic diferencial)[10]

analítica, biòpsia, clínica, cultiu, diagnòstic, hemocultiu, hemograma, limfocític, pronòstic, títol.

e. Accions (relacions entre malalt i metge, entre agent i malaltia, entre malalt i tractament, etc.)

abrasió, amputació, cavitació, complicació, confusió, contaminació, disseminació, espollament, extravasació, flexió, hemòlisi, incubació, infecció, inoculació, lisi, microaglutinació, prevenció, rabdomiòlisi, secreció, seroconversió, tumefacció, vacunació, vegetació.

f. Bioquímica i farmacologia: elements, compostos, substàncies orgàniques químiques, fàrmacs (diagnòstic, tractament, intervenció, prevenció)

aldolasa, antibiòtic, bactericida, ciprofloxacina, citocina, cloramfenicol, cotrimoxazole, creatinafosfocinasa, dosi, eritromicina, fàrmac, immunoperoxidasa, infiltrat, josamicina, lindà, macròlid, monodosi, neopterina, ofloxacina, pefloxacina, permetrina, profilaxi, quimioprofilaxi, repel·lent, rifampicina, prostaglandina, roxitromicina, tetraciclina, transaminasa, vacuna.

L'objectiu principal d'aquesta agrupació és trobar elements regulars que proporcionin indicadors perquè un extractor pugui recuperar la terminologia mèdica dels textos sense haver de recórrer als grans diccionaris clàssics de medicina. És cert, com comenta Sager (1993:144), que la *nomenclatura* mèdica està més diversificada que d'altres codis específics, atès que, d'una banda, designa, a més d'objectes reals, condicions, processos i operacions; i, de l'altra, està sotmesa a una actualització contínua:

[10] Les proves i els exàmens mèdics impliquen mètodes i tècniques que inclouem en la mateixa classe semàntica.

> *El lenguaje médico no se ajusta a un criterio lógico uniforme. Hemos visto que ello se debe principalmente al cambio de significado de los términos a lo largo del tiempo, asó como a los problemas planteados por los epónimos, por la proliferación de abreviaturas y por diferentes fenómenos semánticos sobre todo por la polisemia, la homonimia y la sinonimia. La ausencia de criterio uniforme supone serias dificultades para las funciones que el lenguaje médico tiene que desempeñar como instrumento fundamental de comunicación entre todos los que integran la comunidad médica internacional. También condiciona la eficacia de los sistemas de recuperación de la información y documentación médicas, que deben extraer informaciones determinadas de masa cada vez más enormes de libros y folletos, artículos, informes, historias clínicas, documentos sanitarios.*
>
> [López Piñero i Terrada Ferrandis, 1990: 63]

Però tot i aquestes dificultats, que d'altra banda presenten totes les ciències i les tècniques en major o menor grau, segur que hi ha regularitats entre les USE monolèxiques nominals de la medicina que facilitin la tasca d'un SEACAT, i aquest és el propòsit que abordarem a continuació.

És sabut que el lèxic de la medicina es fonamenta en el lèxic llatí i grec. Segons Quintana (1989: 5), en els diccionaris de medicina, aproximadament un 62% del lèxic es deriva del grec. I el nombre d'arrels grecollatines utilitzades en medicina es calcula al voltant d'unes 1.100. En canvi, el nombre d'UT en el camp de les ciències de la salut sobrepassa en escreix les 100.000 unitats especialitzades. Així doncs, com indica López Piñero i Terrada Ferrandis (1990: 29) "*en torno a mil raíces de procedencia griega o latina componen la casi totalidad de los términos médicos.*"

En el corpus analitzat hem documentat que el 66% de les USE monolèxiques presenten un formant, sufix o prefix cultes. Observem també que moltes de les UT monolèxiques formades a la manera culta donen lloc a UTP. Dit altrament, moltes USE monolèxiques són l'hiperònim d'una altra USE. Si això és cert, significa que una gran part de les USE complexes de la medicina són unitats monolèxiques integrades per formants grecollatins: arrels, sufixos i/o prefixos, que alhora constitueixen unitats polilèxiques. Casos com aquests són molt nombrosos en els textos

mèdics (*biòpsia hepàtica, meningitis linfocítica, nefritis interticial, icterícia colestàtica, mononucleosi infecciosa*, etc.).

Per exemple, el mot *meningitis* està format pel formant clàssic *mening-* i el sufix clàssic *-itis*, és a dir, per un radical que indica una part del cos humà (les meninges) i el sufix *-itis* que indica inflamació de la part assenyalada pel formant ; i alhora el lexema *meningitis* forma part també de la UT *meningitis bacteriana*, que és un hipònim de *meningitis*.

Un altre exemple, el segment *hemorràgia digestiva* està format per un substantiu *hemorràgia* (generat a partir del formant culte *hemo-* que vol dir sang i del sufix culte *-ragia* que significa vessament de la substància assenyalada pel formant al qual s'afegeix) i l'adjectiu *digestiva* (format a partir de la unió del sufix relacional *-iu, -iva* al nom deverbal *digestió*).

Cal subratllar que si bé les unitats del primer nivell morfològic no són USE sinó formants cultes, sí que ho són les del segon i del tercer. Tot i així, només el domini de les unitats del primer nivell permet reconèixer i integrar el significat de les unitats dels grups posteriors

Com a conseqüència de les observacions anteriors, podem formular la hipòtesi que un extractor amb un diccionari d'uns 1.000 formants clàssics podria detectar un gran nombre d'USE monolèxiques i d'aquesta manera reduir substancialment el silenci; i paral·lelament, serviria també per reduir el soroll que produeixen les USE complexes. Si aquests formants cultes, a més, portessin associades etiquetes semàntiques el sistema d'extracció podria tenir una qualitat molt més alta per tal com la semàntica ajudaria, com veurem en el proper capítol, a reduir el soroll.

Per il·lustrar alguns elements d'aquesta hipòtesi, presentem amb més detall algunes característiques de les USE monolèxiques nominals que pertanyen a diversos camps lèxics, concretament dels següents:

1. malalties
2. parts del cos
3. substàncies, compostos i elements bioquímics, etc.
4. accions i operacions.

1. En el camp lèxic de les **MALALTIES**, el nucli semàntic i morfosintàctic de moltes UTP és una USE monolèxica. És important notar que, tant si es tracta d'una USE simple com d'una USE composta a la manera culta, totes aquestes paraules de la classe semàntica de les malalties són l'hiperònim d'altres unitats polilèxiques: *anèmia normocítica, anèmia moderada, anèmia monocrònica, meningitis asèptica, meningitis bacteriana, meningitis linfocítica, febre tacada, febre tifoidea, xarampió atípic,* etc.

En aquest camp, el 86% de les unitats recollides presenten o un afix culte o una arrel culta o bé dos formants cultes alhora. En efecte, per bé que hi ha certes patologies que es denominen amb noms simples (*xarampió, vòmit*), la majoria de malalties denominades mitjançant una USE monolèxica s'han format a partir d'un formant culte que significa una part (o òrgan o sistema) del cos humà i un sufix que s'afegeix a aquesta part del cos humà lesionada:

> *Una de las normas que se recomienda seguir en la formación de neologismos es indicar la clase de enfermedad o estado patológico añadiendo un sufijo a la raíz de la correspondiente parte anatómica:*
>
> | | | |
> |---|---|---|
> | *Inflamación* | *itis* | *neur**itis*** |
> | *Estado patológico no inflamatorio* | *osis* | *nefr**osis*** |
> | *Infección no bacteriana* | *iasis* | *filar**iasis*** |
> | *Tumor, tumefacción crónica* | *oma* | *sarc**oma*** |
> | *Estado patológico* | *ia* | *icteric**ia*** |
> | | *ismo* | *reumat**ismo*** |
>
> [López Piñero i Terrada Ferrandis, 1990: 61]

Bernabeu i al. (1995) recullen una seixantena de sufixos cultes que serveixen per formar els noms de moltes malalties.

En el text analitzat, els principals afixos cultes que serveixen per formar els noms de les malalties o les seves manifestacions biològiques són els següents:

> *-itis* (22,5%); *-osi (12,5%); hipo- (5%); -pènia (3,75%); -algia (2,5%); micro- (2,5%); hiper- (2,5%); megàlia (2,5%); micro- (2,5%); -patia (1,25%); -fobia (1,25%), bradi- (1,25%); oligo- (1,25%); trombo- (1,25%); -oma (1,25%).*

Però no totes les USE monolèxiques nominals que es refereixen a malalties estan integrades per formants grecollatins. En efecte, en els textos mèdics també trobem noms de malalties o de signes i símptomes patològics representats per un terme simple. Els noms simples de patologies i de les seves manifestacions biològiques que trobem en el text analitzat (*buf, febre, tos, vòmit*, etc.) són, però, unitats idiosincràtiques i, per tant, difícilment un extractor els podrà detectar. Per fer-ho, caldria una solució lingüísticament més complexa que recorregués a la semàntica de context.

2. Un altre camp lèxic tradicional dins l'àmbit de les ciències de la salut és l'**ANATOMIA HUMANA**. L'anatomia localitza la part del cos en què es dóna una malaltia. Així, hi ha classificacions de malalties que prenen com a eix vertebrador les parts del cos humà [Cárdenas, 1996]. En general, la classificació anatòmica es basa en principis topogràfics (relacions meronímiques *part-tot*) i en principis funcionals. En anatomia, les parts del cos humà més divulgades solen ser noms simples: *braç, cap, cara, cervell, cor, fetge, mà, melsa,*

pell, peu, pulmó, ull[11]. Ara bé, la majoria de parts del cos que no són tan conegudes per un parlant no especialista se solen denominar o bé amb unitats polilèxiques formades per un nom i un complement (en què aquest té la funció d'identificar inequívocament la part del cos a la qual acompanya) o bé amb paraules monolèxiques formades a la manera culta.

3. En els textos sobre temes mèdics, un bon conjunt d'USE monolèxiques —com l'especialista ha marcat en relació al text de buidatge— "provenen" d'altres camps i més freqüentment del camp conceptual de la **BIOQUÍMICA.** Partint de la idea que si els especialistes en medicina les usen específicament vol dir que deuen ser USE pertinents en els textos mèdics. En el text estudiat trobem trenta unitats de bioquímica. Vint d'aquestes unitats s'han format per la unió d'un sufix amb un significat especialitzat. Els sufixos es distribueixen de la manera següent:

-ina: 46,1%
-ol: 3,8%
-ole: 3,8%

Les unitats formades per un o més d'un formant grecollatí representen en aquest camp semàntic el 27% de les unitats; la resta (aproximadament un 24%) equival a paraules simples molt generals com *dosi, fàrmac, vacuna*.

Més del 60% de les unitats d'aquest grup es podrien recuperar amb el control d'un nombre limitat de sufixos[12].

[11] Aquestes paraules que denominen parts del cos humà se solen usar també a través d'una metàfora per denominar realitats molt diferents en un altre domini temàtic (com la mecànica, la construcció, l'enginyeria, la informàtica, l'urbanisme, etc.).
[12] Certament, la nomenclatura de la química és de les més regulars des que el 1787 Guyton de Morveau i Lervosier, entre d'altres, van publicar el *Méthode de nomenclature*

4. Les **ACCIONS** i les **OPERACIONS** tenen un paper important en el discurs mèdic, per tal com serveixen per al·ludir la interacció entre el malalt i els diferents professionals de les ciències de la salut, entre l'agent d'una malaltia i el pacient d'aquesta malaltia, entre el pacient i el tractament, entre el pacient i les tècniques diagnòstiques, etc. Lingüísticament, la manera més natural d'expressar accions és a través d'un verb. En el discurs mèdic escrit, però, hi ha una tendència a nominalitzar els verbs d'acció. Això comporta que en textos mèdics com historials clínics, informes mèdics, peticions d'exàmens s'usin noms provinents d'un verb d'acció.

En aquest capítol, només ens referirem als noms deverbals que apareixen en el text aï lladament. En el proper capítol sobre el soroll reprendrem aquest tema, atès que aquest tipus de noms generen molta fraseologia nominal.

En el text de buidatge, l'especialista ha marcat vint-i-tres noms d'accions o d'operacions. D'aquests, dinou estan formats per un verb (preferentment de la primera conjugació) i el sufix *-ció (-ió, -ació, -ició)* (*afecció, extravasació, infecció, inoculació,* etc.); un dels substantius d'acció ha estat format a partir d'un verb de la primera conjugació i el sufix *-ment*; (*tractament*); i, finalment, tres substantius d'acció estan formats a la manera culta, a través del sufix *-lisi* amb el significat de "dissolució o desintegració ", afegit a

chimique. A partir d'aquesta data la nomenclatura química ha sofert diferents revisions i reformulacions, però avui encara és vigent i depèn de les diferents comissions de la IUPAC (International Union of Pure and Applied Chemistry). En medicina té una importància particular la nomenclatura de la química orgànica que consisteix en "*una raíz que corresponde al esqueleto carbonado, y en sufijos y prefijos que indican los grupos funcionales*" [López Piñero i Terrada Ferrandis, 1990: 69]. Paral·lelament, la IUPAC i la International Union of Biochemistry tenen una comissió compartida que publica normes

un formant culte indicant la part del cos o l'element afectats (*lisi, hemòlisi i rabdomiòlisi*).

Tot i que en el text de buidatge només hem documentat un sufix culte que serveix per formar noms d'accions o d'operacions, Bernabeu i al. (1995) en recopilen quaranta-tres[13].

En resum i després d'examinar quatre camps lèxics de manera molt general, podem concloure que, en medicina, el paper dels afixos —tant catalans com grecollatins— i de les arrels cultes és significatiu i pot ajudar enormement a reduir el silenci que sol generar un SEACAT basat en patrons morfosintàctics.

Però a més de noms, en els textos especialitzats també s'usen USE monolèxiques que pertanyen a altres categories gramaticals —encara que aquestes siguin menys freqüents. A continuació, estudiarem els verbs, els adjectius i els adverbis constituï ts per una sola paraula detectats per l'especialista.

3.2.1.2 USE monolèxiques verbals

L'especialista ha subratllat un total de quinze USE monolèxiques verbals del text de buidatge (3,15% del total d'unitats seleccionades): *aglutinar, contaminar, cultivar descrostar-se, desinfectar, desparasitar, infectar, infestar, injectar, parasitar, prostrar, reepitelitzar-se, remetre, tolerar, vacunar.*

---

sobre nomenclatures i símbols d'interès especial en el marc de la medicina (International Union of biochemistry, 1978).

[13] Com ara *-ectomia* —que significa extirpació— (*amigdalectomia, mastectomia, pneumectomia*, etc.), *-scòpia* —que significa observació— (*abdominoscòpia, gonioscòpia, uranoscòpia*, etc.), *-pèxia* —que significa coagulació— (*adipopèxia, organopèxia, vaginopèxia*, etc.).

Tots aquests verbs pertanyen a la primera conjugació, a excepció de *remetre* que és un verb de la segona conjugació. La majoria són transitius, excepte *remetre,* que s'usa en la seva accepció intransitiva (en el DIEC **remetre** (..) v. intr. Perdre la intensitat. *Remetre la febre, el dolor la tos),* i de *descrostar-se* i *reepitelitzar-se* que són verbs pronominals. La majoria d'aquests verbs es poden nominalitzar mitjançant el sufix *-ció,* excepte *cultivar* que es nominalitza per un procés de conversió sintàctica (*cultiu)* i *tolerar* que ho fa amb el sufix *-ança (tolerança).* En el mateix text d'anàlisi trobem unitats nominalitzades formades a partir d'algun d'aquests verbs: *cultiu, infecció, contaminació, vacunació* i *inoculació.*

Dels quinze verbs que l'especialista ha marcat en el corpus de buidatge, només tres no es troben en el *Diccionari de la Llengua Catalana* (DLC)[14] i d'aquests dotze verbs, n'hi ha nou que en el DLC porten una marca d'àrea temàtica específica[15].

Paral·lelament, gairebé tots aquests verbs apareixen documentats en el *Diccionari Enciclopèdic de Medicina* (1990), excepte: *cultivar, desparasitar, prostrar, reepitelitzar-se. Cultivar, desparasitar, reepitelitzar-se* no estan documentats a cap diccionari, ni general ni especialitzat. I *Prostrar* és l'únic verb d'aquesta llista que, pel seu caràcter general, no forma part de cap diccionari especialitzat[16].

[14] L'accepció semàntica de *cultiu* com una tècnica microbiològica no es recull en el diccionari, per bé que s'inclogui l'entrada **cultiu**.

[15] infectar, infestar: PAT (patologia); injectar, tolerar: MED (medicina); inocular, vacunar: TERAP (terapèutica); desinfectar: FARM (farmacologia); parasitar: BIOL (biologia); contaminar: ECOL (ecologia).

[16] En el DLC, el nombre de verbs amb una marca temàtica afí a les ciències de la salut és de 508, xifra que representa l'1,9% de les unitats lèxiques amb una marca temàtica relacionada amb la medicina:

Àrea temàtica del DLC	unitats lèxiques	unitats verbals
ANAT	302	0
ANAT ANIM	1950	4 (0,1%)
EMBRIOL	225	1 (0,4%)
CIT	285	1 (0,35%)

Des del punt de vista semàntic, podem agrupar els verbs amb valor especialitzat a partir de les seves característiques semanticofuncionals ja que una classificació dels verbs d'aquest tipus podria ser una ajuda per als SEACAT basats en coneixement lingüístic.

Un extractor que utilitzés un analitzador sintàctic amb etiquetes semàntiques facilitaria la detecció dels verbs especialitzats juntament amb els seus arguments. En el cas de no poder disposar d'aquesta eina, la freqüència d'ús dels verbs, les concordances, el context funcional i el fet que hi hagi en el corpus nominalitzacions i altres USE del mateix paradigma també poden ajudar un SEACAT a detectar les USE verbals[17].

FARM	1184	16 (1,35%)
FISIOL ANIM	629	30 (4,76%)
GEN	338	7 (2,07%)
IMMUNOL	14	0
MED	859	57 (6,6%)
OBST	59	0
ÒPT	327	11 (3,3%)
PAT	3709	98 (2,6%)
PEDIAT	23	1 (4,3%)
PEDOL	180	3 (1,6%)
PSIQ	353	5 (1,4%)
TERAP	392	38 (9,6%)
TOXICOL	31	1 (3,2%)
TRAUM	32	7 (21,8%)
BIOL	900	23 (2,5%)
BIOQ	928	4 (0,4%)
BIOTEC	21	0
BOT	6945	47 (0,6%)
ECOL	319	5 (1,5%)
GENEAL	32	0
PSIC	717	19 (2,6%)
QUÍM	1096	98 (8,9%)
QUÍM ORG	1828	17 (0,9%)
ZOOL	2680	15 (0,5%)

En el *Diccionari Enciclopèdic de Medicina* (DEM) el nombre de verbs també és pobre: 2.844, un 3,5% de les UT d'aquesta obra.

[17] Els verbs de caràcter especialitzat han estat poc estudiats.

3.2.1.3 USE monolèxiques adjectives

En el corpus analitzat, l'especialista ha subratllat vint-i-un adjectius que apareixen aï lladament. Tots són adjectius derivats: formats per un nom o un formant culte nominal i un sufix (*citopàtic, epidemiològic, cel·lular, alveolar, cerebral, humoral,* etc.). Aquesta dada indica que en medicina els adjectius simples no són USE, si no s'integren en una USE polilèxica (*groc, -ga* i *greu*: *febre groga, insuficiència cardíaca greu*).

El nom que serveix de base a les USE adjectives autònomes és **sempre** una UT. Els sufixos d'aquests adjectius pertanyen a dues grans classes:

- sufixos prototípicament relacionals[18]
- sufixos prototípicament qualificatius.

Cada un d'aquests tipus es distribueix de la següent manera:

1. Sufixos relacionals

   -ic: 10 (*al·lèrgic, citopàtic, clínic, embòlic, epidemiològic, exantemàtic, histològic, neurològic, purpúric, subclínic*), "relatiu a"

   -ar: 2 (*cel·lular, polimorfonuclear*), "relatiu a"

   -al: 2 (*humoral, petequial*), "relatiu a"

   -í: 1 (*intrauterí*), "relatiu a"

   -il: 1 (*febril*), "relatiu a"

2. Sufixos qualificatius:

   -ós: 3 (*maculós, macupapulós, papulós*), "que té forma de"

L'especialista també ha marcat un participi (*immunitzat*) i un gerundi (*pruent*) que en el text funcionen autònomament; en aquests dos casos observem que el verb del qual deriven és també especialitzat.

Dels vint-i-un adjectius derivats autònoms seleccionats, disset estan documentats en el *Diccionari de la Llengua Catalana* (1993). D'aquests disset adjectius, catorze porten una marca d'àrea temàtica especialitzada afí a les ciències de la salut[19]. La majoria d'aquests adjectius (el 90%) també han estat documentats en el *Diccionari Enciclopèdic de Medicina* (1990). Tres adjectius dels vint-i-un, però, (*embòlic, epidemiològic i macupapulós*) no estan recollits en cap dels dos diccionaris, encara que aquests tres adjectius s'han format a partir dels noms *embòlia, epidemiologia, pàpula* i *màcula,* que estan documentats tant al *Diccionari de la Llengua Catalana* (1993), amb un marca temàtica de medicina, com en el *Diccionari Enciclopèdic de Medicina* (1990), i, per tant, són també USE.

La meitat dels adjectius derivats estan compostos per un **constituent grecollatí**; fet que podria beneficiar el funcionament d'un SEACAT que comptés amb un diccionari de formants grecollatins. Així, per reconèixer les USE adjectives aï llades, els SEACAT podrien recórrer a l'estratègia següent:

- Primerament, podrien comprovar si es tracta d'un adjectiu derivat o compost.

- Si la resposta és afirmativa, podrien comprovar si aquest està integrat per algun formant grecollatí propi de les ciències de la salut.

[18]Per a una classificació dels adjectius terminològicament pertinents, vegeu l'apartat 4.2.2 del capítol proper.

- I, si la resposta és afirmativa, podrien detectar en el corpus totes les USE que comparteixin el mateix lexema de l'adjectiu (*clínic* i *clínica*, *papulós* i *pàpula*, *embòlic* i *embòlia*, etc.), perquè si hi ha un altre USE amb el mateix lexema és una prova que es tracta d'una adjectiu especialitzat.

En el corpus de buidatge, tots els adjectius especialitzats que apareixen aï lladament compleixen dues d'aquestes condicions i, la majoria, totes tres.

3.2.1.4 USE monolèxiques adverbials

Les USE adverbials no són gaire freqüents en els textos especialitzats, per bé que n'hi ha algunes. Tots els adverbis especialitzats monolèxics pertanyen al grup dels adverbis formats a partir d'un adjectiu en femení i del sufix *-ment*, amb el significat "des del punt de vista + A", i a vegades "de manera + A". En el cas d'adverbis especialitzats, l'adjectiu a partir del qual es forma l'adverbi és **sempre** una USE.

En efecte, els tres adverbis marcats per l'especialista amb significat especialitzat segueixen aquest patró: *clínicament, histològicament, immunològicament*. Tots tres s'han format a partir d'un adjectiu en femení (*clínica, histològica* i *immunològica*) documentat en el mateix corpus. En aquest corpus, també hi trobem altres USE que comparteixen la mateixa base: *clínica, histologia, immunologia*[20].

[19] PAT (patologia): 5; MED (medicina): 4; ANAT ANIM (anatomia animal): 2; CIT (citologia): 1; HISTOL (histologia): 1; FISIOL ANIM (fisiologia animal): 1.

[20] En el DLC i en el DIEC no trobem aquests adverbis documentats —perquè formalment són regulars i, per tant, predictibles—, però sí els adjectius i els noms amb la mateixa arrel; en el DLC, a més, aquests noms i adjectius estan marcats amb una marca temàtica: MED (medicina): *clínic, clínica* i *immunològic i immunologia*; i HISTOL (histologia): *histològic i histologia*.

Consegüentment, si un SEACAT vol extreure les USE adverbials d'un corpus especialitzat:

- En primer lloc, ha de detectar només els que estan formats per un adjectiu i el sufix *-ment*.
- I, a continuació, d'aquest grup d'adverbis en *-ment* ha de seleccionar només els que l'adjectiu sigui especialitzat (que estarà documentat en el mateix corpus de buidatge).

3.2.1.5 Síntesi de l'anàlisi sobre les USE monolèxiques

L'anàlisi de les USE monolèxiques marcades per l'especialista ens permet arribar a una sèrie de constatacions, de les quals voldríem retenir les següents:

En primer lloc hem observat que les USE monolèxiques poden ser noms, verbs, adjectius i adverbis. Hem comprovat, però, que la representativitat d'aquestes categories en el corpus textual és molt diferent. Mentre que els substantius representen el 35% del total d'unitats especialitzades seleccionades per l'especialista, els verbs només representen el 3,25%, els adjectius el 2,19% i els adverbis l'1%.

D'acord amb el nostre objectiu d'intentar buscar elements regulars que els SEACAT puguin utilitzar per tal d'augmentar la seva eficiència, hem proposat l'ús d'elements morfològics i semàntics que poden facilitar la discriminació de les USE monolèxiques d'un text:

1. Pel que fa als noms, els hem agrupat en camps lèxics molt generals i hem comprovat que es podien trobar algunes constants en el si de cada camp lèxic. Així, hem comprovat que la majoria de

malalties presenten un constituent grecollatí. Més precisament, hi ha moltes unitats que denominen malalties o estats patològics que s'han format adjuntant un sufix o un prefix cultes a un radical grecollatí que indica una part anatòmica del cos humà.

En general, hem vist que moltes USE monolèxiques de diversos camps lèxics s'han format a partir d'un o més d'un constituent grecollatí. Aquest fet ens porta a la conclusió que els formants clàssics són la base morfològica del lèxic mèdic.

Altres camps lèxics també presenten regularitats pel que fa a les bases i als sufixos més indicatius; aquest és el cas dels noms que expressen accions o operacions, els quals, majoritàriament, s'han format per verbs de la primera conjugació més el sufix nominalitzador *-ció*.

2. Pel que fa els verbs, hem comprovat, doncs, que la majoria pertanyen a la primera conjugació i que estan documentats a les obres lexicogràfiques de referència i hem suggerit que establir el seu context funcional podria facilitar-ne la detecció.

3. Pel que fa als adjectius monolèxics que en el text apareixen aï lladament, hem constatat que tots són derivats formats a partir d'un substantiu de caràcter especialitzat i d'un sufix (majoritàriament relacional).

4. Finalment, **quant a les USE adverbials**, hem constatat que són poques i que totes s'han format a partir d'un adjectiu especialitzat en femení i del sufix *-ment* i el seu significat és, gairebé sempre, "des del punt de vista + A" i, molt rarament, "a la manera + A".

En resum, hem arribat a la conclusió que alguns recursos lingüístics que permetrien reconèixer automàticament una USE monolèxica en els textos de medicina són:

- Els formants, prefixos i sufixos grecollatins
- L'agrupament en classes semàntiques de les USE d'un mateix corpus
- Alguns sufixos que tendeixen a adjuntar-se a determinades bases especialitzades.

Les USE monolèxiques, des del punt de vista de la seva formació, poden ser simples, derivades o compostes. Les USE simples són idiosincràtiques i, per tant, no presenten elements lingüístics per poder-les identificar. Les USE derivades es caracteritzen per la combinació recurrent de determinades bases amb determinats sufixos i/o prefixos. En aquest cas, hem vist com en medicina les operacions i accions, es formen, bàsicament, per la unió del sufix *-ció* a un verb de la primera conjugació. En canvi, els adjectius es decanten pels sufixos relacionals: *-ic, -al, -ar.*

Però les USE monolèxiques més abundants en els textos mèdics són els compostos cultes, perfectament cultes o híbrids. En aquest sentit hem constatat que més del 70% de les USE monolèxiques del corpus presenten un morfema culte. I hem documentat que, en el marc de la medicina, actualment, funcionen uns 1.100 formants cultes.

Hem proposat que un SEACAT es beneficiaria de l'ús d'un diccionari de constituents cultes amb informació formal i semàntica. Hem formulat la hipòtesi que amb aquest diccionari un SEACAT reduiria tant el silenci de les USE monolèxiques com el soroll de les USE polilèxiques. I més, si tenim en compte que, com hem evidenciat, les USE monolèxiques nominals són hiperònims de moltes USE polilèxiques. Hem documentat que entre un

75% i un 80% aproximadament de tot el lèxic de la medicina està constituï t, com a mínim, per un formant culte; diferents autors [Quintana, 1989], [Cárdenas, 1996], [López Piñero i Terrada Ferrandis, 1990], [Bernabeu i al., 1995] avalen aquesta afirmació. Això significa que la base del lèxic de la medicina, tant si es tracta d'una USE monolèxica composta com d'una USE polilèxica, són els formants grecollatins[21].

Com a conseqüència d'aquest fet, també hem formulat la hipòtesi que un SEACAT amb un diccionari de formants podria actuar seguint el mateix procés que segueixen els estudiants de ciències de la salut. De fet, a la pràctica, cap professional és capaç de dominar una quantitat molt gran de lèxic especialitzat (més de 100.000 paraules), però, gràcies al domini del significat dels formants cultes, molts professionals desglossen el significat —a vegades per aproximació— d'una unitat que escolten o llegeixen per primera vegada. El punt clau és, doncs, que l'extractor conegués els mil morfemes grecollatins més habituals en ciències de la salut.

3.2.2 Sigles especialitzades

Al costat de les USE monolèxiques, els especialistes també usen sigles especialitzades, les quals poden ser un obstacle per al reconeixement i l'extracció automatitzats. Les sigles —combinació de les inicials de les diverses paraules i/o formants d'una USE polilèxica— són unitats lèxiques aparentment simples, però si analitzem el seu procés de formació

[21] Aquesta hipòtesi és la mateixa que defensa Quintana (1989: 5) quan diu: *"Si cogemos uno de los grandes diccionarios de Medicina, podemos encontrar allí unos 80.000 términos. De ellos no todos son propiamente médicos, pues los hay de otras ciencias auxiliares (Biología, Bioquímica, Psicología, etc.); específicos de Medicina pueden ser unos 55.000, y de éstos ciertamente que 50.000 se derivan del griego. El número de raíces griegas de que se componen vienen a ser unas 1.000."*

comprovem que tenen un origen complex. Les sigles amb significat especialitzat són presents en els textos temàticament especialitzats, sobretot en aquells que van dirigits a especialistes o aprenents d'especialista[22]. En efecte, en tots els camps científics, tècnics o professionals s'usen, en més o menys freqüència, sigles[23].

En medicina, el nombre de sigles en els discursos mèdics va creixent desmesuradament, el fet que hi hagi al mercat diccionaris exclusivament de sigles mèdiques n'és una prova evident; per posar dos exemples concrets, l'any 1989 Heister (1989), motivat per la dificultat per entendre abreujaments que s'utilitzen en els articles científics[24], publica el

[22] Les sigles poden substituir noms comuns —TNF (factor de necrosi tumoral)— o noms propis —ICS (Institut Català de la Salut). En aquest estudi ens centrarem especialment en les sigles comunes.

[23] Vegem-ne alguns exemples d'àmbits temàtics molt diferents:

- En economia, les sigles proliferen i algunes són utilitzades habitualment pel públic general, per bé que moltes vegades estan tan lexicalitzades que el parlant té dificultats per reconstruir totalment el segment que, en un principi, substituï en:

  (18)

  EO (*Estimació Objectiva*), EDS (*Estimació Directa Simplificada*), IAE (*Impost de les Activitats Econòmiques*), IRPF (*Impost de la Renda a Persones Físiques*), IRPH (*Índex de Referència de Préstecs Hipotecaris*), IVA (*Impost del Valor Afegit*), NIF (*Número Identificació Fiscal*), PIB (*Producte Interior Brut*), TAE (*Tassa Anual Equivalent*), etc.

- La informàtica és potser una de les matèries en què més sigles es fan servir. En aquesta disciplina tècnica, les sigles no adaptades són molt habituals. Certament, gran quantitat de sigles s'usen en anglès i són tan conegudes i utilitzades internacionalment que els especialistes prefereixen no modificar-les a fi de conservar-ne la universalitat:

  (19)

  DOS (*Disk Operating System*), DRAM (*Dinamic Random Acces Memory*), FTP (*File Tansfer Protocol*), PC (*Personal Computer*), RAM (*Random Acces Memory*), ROM (*Read Only Memory*), SIMM (*Single In-Line Memory Modules*), SVGA (*Super Video Graphics Adapter*), URL (*Uniform Resource Locator*), WWW (*World Wide Web*), etc.

- Volem citar també un cas del món professional. És cert que en dominis més professionals les sigles no són massa abundants, això no vol dir, però, que siguin inexistents. En l'àmbit dels esports, les sigles d'organismes —i, per tant, noms propis— són molt corrents (JJOO, FCB, FIFA, FIBA, FITA, UEFA), però també se n'usen algunes que substitueixen unitats terminològiques sintagmàtiques, com:

  (20)

  AF (*Aleví Femení*), BTT (*Bicicleta Tot Terreny*), C-2 (*Canoa de dues places*), GR (*Gran Recorregut*), KO (*Knocking Out*), PPE (*Període Preparatori Específic*) , R1 (*primera Reunió*), RV (*Ruta Verda*), etc.

[24] *"Both as a student and later as practising physician I kept on being tripped up by medical and scientific abbreviations; sometimes this annoyed me. I spent hours with books or in libraries trying to find out the meanings of some of these abbreviations. At*

Dictionary of Abbreviations in Medical Sciences que conté aproximadament 15.000 entrades; l'any 1993 l'editorial Jims de Barcelona edita un recull de les sigles que s'usen en oncologia, aproximadament unes 1.500 d'una sola branca de coneixement de la medicina interna[25].

En segons quines situacions comunicatives es crea una sigla per a qualsevol unitat discursiva, fins i tot per a una frase sencera i, d'aquesta manera, les sigles poden esdevenir conjunts semànticament indesxifrables i sovint polisèmics:

(21)

PH pot voler dir:

hipertròfia prostàtica, hipertensió pulmonar, història personal o història prèvia.

IR pot voler dir:

immunoreactiu, resistència interna, resposta immunològica.

Viana i de la Morena (1998) han analitzat la presència de sigles en les altes mèdiques, document en què tots els apartats s'adrecen al pacient, i

times, even recognised authorities had to admit default on the meaning of an acronym that did no belong to their immediate specialist field." [Heister, 1989: VII].

[25] D'acord amb el pròleg del *Dictionnaire de sigles* publicat per La Maison du Dictionnaire (1992), aquesta proliferació de sigles en diferents dominis especialitzats està motivada, bàsicament, per tres aspectes que caracteritzen la comunicació i la informació avui dia: la rapidesa, la concisió, la condensació. L'abús de les sigles, però, pot donar lloc a efectes contraris al de la concisió i provocar incomunicació i desinformació: *"La multiplication des sigles répond aux impératifs du développement des techniques de communication et d'information qui contraignent à la rapidité, à la concision, à la condensation. Il convient de révéler aussi qu'elle accentue les particularismes et les corporatismes et qu'elle engendre souvent des phénomènes d'incompréhension et d'incommunication. De caractère profondément ambigu, l'existence des sigles recoupe un problème très ancienne qui, de l'antiquité de Platon à la linguistique contemporaine, s'interroge sur le problème du rapport entre le signe et le sens."*

han constatat que són molt abundants sobretot pel que fa als antecedents i a l'exploració, però també en el judici clínic i el tractament:

> (…) resultan más problemáticos las manejadas en el tratamiento con referencias tan complejas como L-X-V (lunes-miércoles-viernes); D-C-C (desayuno-comida-cena); acudirá a C. Ext. MI (acudirá a consultas externas de Medicina Interna); control de TA, Fc, ECG o TAC tras el alta; contraindicación de AINES o debe tomar 02 o Haloperidol X-X-X- gotas; todas estas abreviaturas sin explicación previa en el texto.
>
> [Viana i de la Morena, 1998: 196]

Malgrat tota aquesta diversitat, hem intentat buscar regularitats i elements que puguin ser útils a un SEACAT per al reconeixement i l'extracció de les sigles especialitzades que es troben en els textos mèdics. Per fer-ho, hem partit de l'observació de les sigles que apareixen en el corpus textual *Malalties infeccioses per Rickettsia* i en diferents informes hospitalaris, i hem pogut fer les constatacions següents:

Les generalitzacions que es poden fer per a les sigles i que poden ser d'utilitat en el disseny d'un sistema d'extracció de terminologia són les següents:

1. **Les sigles tendeixen a tenir un caràcter internacional.** La internacionalització de les sigles té conseqüències pràctiques en el disseny d'un SEACAT atès que en aquests casos se sol respectar l'ordre de formació de la llengua en què es crea la sigla, que, en molts dominis temàtics, és la llengua anglesa. Com és sabut, l'anglès respecte de les llengües romàniques segueix un ordre invers en l'especificació del substantiu. Per tant, en aquests casos l'extracció automàtica de sigles no adaptades és més difícil perquè la sigla no respecta l'ordre del segment en la llengua d'arribada i, a vegades, les inicials poden no coincidir:

(22)

BCC: *Carcicoma BasoCel·lular* (*Basal Cell Carcicoma*)

EPS: *Símptoma ExtraPiramidal* (*ExtraPiramidal Syntoms*)
BCDF: *Factor Diferencial de les Cèl·lules B* (*B-Cell Differential Factor*)

2. **Les sigles presenten diferents graus de lexicalització.** Pragmàticament, les sigles tendeixen a adaptar-se al sistema lingüístic d'una llengua i a funcionar com a paraules, a lexicalitzar-se, per bé que no totes presenten el mateix grau de lexicalització. Els dos punts següents mostren nivells diferents de lexicalització:

1. L'objectiu de la sigla és, en principi, substituir una UTP per fer més àgil el discurs. Ara bé, això no és del tot cert perquè —com hem vist—, a vegades, les sigles poden substituir segments superiors a la unitat lèxica. En el corpus examinat, trobem el segment *reacció en cadena de la polimerasa* (PCR) que, des del punt de vista lingüístic, no és una unitat terminològica *strictu senso*, sinó una unitat fraseològica. Les sigles, doncs, no només corresponen a sintagmes lèxics, sinó també a unitats superiors a la unitat lèxica:

(23)

AC x FV (*Arítmia cardíaca per fibrilització ventricular*), BCRDFH[26] (*Bloqueig Complet Anterior a la Branca Dreta del Feix de Hiss*), PCR (*Reacció en Cadena de la Polimerasa*).

És interessant observar, però, que mitjançant la sigla aquests sintagmes fraseològics s'adapten perfectament al sistema

[26] Aquesta sigla pot tenir diferents variants perquè el bloqueig pot ser complet o incomplet, anterior o posterior i la branca, dreta o esquerra; si fem totes les

lingüístic, de tal manera que funcionen com unitats lèxiques nominals:

(24)

El DNA de C. burnettii es pot detectar a la sang o en mostres tissulats mitjançant la PCR.
Es poden detectar anticossos específics mitjançant ELISA[27].

2. Hi ha sigles que poden arribar a produir sèries derivatives: de *sida, sidós, sidòpata* i *sidòtic*; de *làser, laserteràpia;* i fins i tot poden formar part d'unitats sintagmàtiques: *afectació de l'SNC; DNA rickettsià; pH àcid*, etc. Aquest darrer fenomen és habitual en medicina i un SEACAT ha tenir-lo present a l'hora de detectar les USE polilèxiques.

3. **Les sigles solen presentar una tipografia determinada.** En medicina, les sigles gairebé sempre s'escriuen en lletres majúscules, normalment juntes i sense punts (per bé que certs autors i tradicions les separen amb punt). Si una sigla la trobem en minúscula és un signe més del seu grau elevat de lexicalització: *làser, radar, sida*.

Les sigles solen estar formades per seqüències de dues a cinc lletres, encara que, normalment, substitueixen un segment de tres paraules

combinacions tenim vuit possibilitats diferents: BCABDFH o BIABDFH o BCPBDFH o BIPBDFH o BCABEFH o BIABEFH o BCPBEFH o BIPBEFH.

[27] També s'hauria de fer un estudi sobre el gènere i el nombre de les sigles. En teoria, el gènere ha de ser el mateix que el del nucli del sintagma al qual substitueixen i el nombre hauria de ser invariable. Però això és el que aconsellen els manuals d'estils, a la pràctica el gènere pot ser fluctuant i moltes sigles s'usen en plural (LOE i LOES, TAC i TACS), tant per escrit com sobretot oralment. El text que estudiem és un manual universitari i, per tant, ha estat revisat per correctors que vetllen per la utilització normativa de les sigles; tot i així llegim: *Altres tècniques, com la PCR (..).*, però també *S'ha utilitzat la*

referencials. Això és lògic si pensem que no existeixen UT superiors a cinc elements i que la majoria d'UT estan formades per un nom i un o dos complements. En el corpus textual sobre les malalties infeccioses per rickèttsia apareixen 14 sigles, de les quals 13 estan formades per tres lletres, només una està formada per cinc lletres (ELISA: *assaig immunoabsorbent lligat a enzims*). Si fullegem el *Diccionario oncológico. Abreviaturas, siglas y acrónimos* (1993) ratifiquem l'afirmació que la sigla més usada és la formada per tres lletres majúscules.

4. **Les sigles tendeixen a seguir unes pautes de disposició en el text.** Algunes de les sigles, les menys usades, les més discursives, en el primer moment que apareixen en el text ho fan *entre parèntesis*, darrera mateix del segment desenvolupat al qual substitueixen. De fet, això és el que recomanen els manuals d'estil [Caldeiro i al., 1993], [Benavent i Iscla, 1997]. En els textos, però, les sigles molt conegudes pels especialistes no apareixen entre parèntesis. Resulta il·lustratiu citar l'anàlisi del llenguatge científic de les comunicacions presentades en el *IV Congreso Nacional de Documentación Médica* realitzada per Benavent i Iscla (1997). En aquest estudi, els autors recullen 75 abreviatures que no han estat explicades en els textos, cosa que els fa concloure que:

> *En el lenguaje científico, el abuso de abreviaturas, siglas y acrónimos convierten al lenguaje en un instrumento impreciso, con graves problemas para su comprensión, ya que en ocasiones se establecen por simple economía lingüística del creador. Además, evolucionan, aparecen otras nuevas que las sustituyen, pierdan o cambian de significado, se utilizan en nuevas situaciones en la que es difícil reconocerlas o caen en desuso.*
>
> [Benavent i Iscla, 1997: 11]

Segons el grau de lexicalització, les sigles poden conviure amb la denominació desenvolupada (PCR, IFI, IFD) o bé poden haver

---

tècnica del PCR per a detectar DNA (....), i recordem que PCR significa reacció en cadena

substituï t la forma completa (ADN, TAC) fins al punt que se sol desconèixer el segment exacte al qual substitueixen, com és el cas de *làser* (Light Amplification by Stimulated Emission Radiation). La pèrdua de relació entre la sigla i el sintagma que substitueix és el principal problema amb què topen els SEACAT, ja que aï lladament una sigla és idiosincràtica i si s'escriu en minúscules està lexicalitzada i comporta com una unitat simple.

En el nostre corpus, sis de les catorze sigles que hi apareixen no ho fan precedint la unitat totalment desenvolupada, ni tant sols si es tracta de la primera vegada que apareixen al text: DNA, DDT, ELISA, RNA, SNC, VSG. Un llec en medicina pot saber què significa DNA i RNA, però difícilment coneix les altres quatre.

5. **Els redactors de textos sobre temes mèdics abusen de les sigles.** Com comentàvem anteriorment, en segons quins tipus de textos mèdics es produeix un abús en la utilització de les sigles. En els informes mèdics, històries clíniques, exploracions, anamnesis, les sigles es multipliquen; en aquests casos, les sigles internacionals se solen barrejar amb sigles pròpies d'un grup professional o una escola mèdica i, en aquests contextos, no apareixen mai al costat de la unitat sencera a la qual representen. A vegades, fins i tot dins d'una mateixa branca mèdica, els metges d'un centre no poden desxifrar informes escrits per metges d'un altre centre per la quantitat de sigles *ad hoc* que fan servir.

6. **Les sigles són semànticament opaques.** No hi ha cap element de la sigla que sigui semànticament transparent. Des del punt de vista del significat, les sigles són opaques, per bé que expressen molt més del que es pot intuir formalment, ja que el significat de les sigles

---

de la polimerasa, en què el nucli del sintagma és un substantiu femení.

> és el mateix que el del sintagma al qual substitueixen, però entre els dos elements només hi ha una relació d'inicials. Si la sigla, a més, no està adaptada, aquesta relació encara està més poc assegurada. Carbonnier (1980) expressa metafòricament aquesta dificultat per arribar al significat de les sigles en el pròleg del *Dictionnaire des principaux sigles utilisés dans le Monde juridique:*
>
> > *Les juridiques, c'est l'ésotérisme à la puissance. Le langage du droit est déjà, bien souvent, langage d'initiés. S'il se cache derrière une grille de lettres à compléter, il s'ajoute le cryptogramme à l'énigme. Nous voilà perdus.*

Consegüentment, un extractor es pot valer d'estratègies basades en elements lèxics (diccionaris de sigles especialitzades), tipogràfics i de distribució en el text per reconèixer i extreure les sigles especialitzades dels textos.

3.2.3 Unitats fraseològiques especialitzades verbals[28]

En el buidatge que l'especialista ha realitzat del nostre corpus d'anàlisi, no ha hi cap unitat fraseològica especialitzada (UFE) verbal marcada com a unitat especialitzada pertinent. Els únics verbs subratllats per l'especialista són unitats monolèxiques.

Des d'una perspectiva lingüística, però, si analitzem detalladament el text, constatem que hi ha determinades expressions, que tenen com a nucli un verb, que es repeteixen contínuament i que els seus constituents no poden ser substituï ts lliurament: *transmetre una malaltia, tolerar un tractament, un fàrmac, etc.*

[28] En el proper capítol tractarem les UFE nominals, perquè la seva estructura en lloc de silenci sol generar soroll.

Hem preguntat a tres especialistes si consideraven que aquests segments formaven part del lèxic mèdic. La resposta ha estat unànime: *"són paraules que solem usar combinades d'aquesta manera, juntes"*. Aquesta constatació corrobora que en medicina hi ha també fraseologia verbal especialitzada i que aquesta es caracteritza per la freqüència d'ús.

Històricament s'ha tendit a considerar les UFE verbals fora del conjunt de paraules que anomenem *terminologia*. En el supòsit, però, que un SEACAT les volgués detectar, podria utilitzar eines lexicomètriques de mesura de freqüències d'ús dels segments sintagmàtics i d'altres elements lingüístics que analitzarem en el capítol cinquè.

3.2.4 USE no lingüístiques

En determinats textos especialitzats (matemàtiques, botànica, química, zoologia, etc.) l'ús de codis alternatius procedents de sistemes semiòtics complementen els lingüístics. L'especialista quan es comunica fa servir sovint unitats lingüístiques i unitats no lingüístiques, sobretot en la comunicació escrita molt especialitzada:

> *Bien loin que les langues spécialisées soient des sous-systèmes linguistiques autonomisables, il s'agit d'usages socialement normé de plurisystèmes. Les textes scientifiques comportent de façon régulière et prévisible des signes non linguistiques au sein même de leurs énoncés. Même les signes qui appartiennent à des alphabets, comme* **α**, *par ailleurs lettre grecque, dans l'expression radioactivité* **α**, *sont intégrables à des systèmes non linguistiques, des notations propres à telle discipline. D'où le besoin d'une théorie générale des systèmes de signes (d'une sémiotique), pour en pas limiter l'approche des langues spécialisées à une lexicologie des racines grecques, latines et autres. Cette prise en compte de l'intégralité des signes utilisés dans les énoncés spécialisés conduit à se donner des unités terminologiques une définition qui prévoie les cas comme celui de radioactivité* **α** *(...) Mais une théorie des langues spécialisées en peut se fonder que sur une théorie générale des langues.*
>
> [Lerat, 1995: 28-29]

Les unitats no lingüístiques més usuals d'aquesta mena de discursos són o bé símbols internacionals o bé noms llatins que formen part de nomenclatures científiques estandarditzades.

Tot i el seu caràcter aparentment lingüístic, les nomenclatures estan constituï des d'unitats no lingüístiques que no pertanyen al llenguatge natural, perquè han estat creades artificialment. A la pràctica, però, les nomenclatures de la química, de l'anatomia i de les malalties, per posar alguns exemples, no semblen artificials perquè usen recursos vigents en la llengua: sufixos, prefixos, formants grecollatins. En canvi, la nomenclatura de la zoologia, de la botànica, de la bacterologia, de la virologia o el *Sistema Internacional de Mesures* són sistemes que fan servir recursos propis de llengües mortes que es perceben com a aliens a la llengua natural en funcionament:

> *Los lenguajes artificiales creados de esta manera se asocian diversamente con uno o varios lenguajes naturales que se han utilizado en su construcción. El grado de dependencia con otros lenguajes y el área de uso del lenguaje decide la influencia que el lenguaje natural tienen en el sistema de denominación construido.*
>
> [Sager, 1993: 140]

3.2.4.1 Símbols

El *Diccionari de la Llengua Catalana* (DIEC) (1995) defineix el símbol d'una manera molt àmplia: "element que es pren com a signe figuratiu d'un altre per raó d'una analogia que l'enteniment percep entre ells o d'una convenció". En aquest treball, però, només considerarem símbols aquells elements que serveixen per representar les magnituds, quantitats, unitats, operacions i també els que serveixen per representar els elements químics. Aquests símbols són fruit d'un consens internacional abastament debatut. Tots aquests símbols estan estandarditzats en el *Sistema Internacional d'Unitats* (1974) o en la *Taula Internacional d'Elements Químics* (1787(1976)).

No entrarem en valoracions sobre aquest tema, sinó únicament apuntarem que aquests símbols són també presents en alguns textos que parlen sobre qüestions mèdiques. En el nostre text només hem trobat dues unitats simbòliques aï llades: PG_1 i PG_2, que representen dos àcids grassos de la família de les *prostangladines*.

3.2.4.2 Noms científics en llatí

La nomenclatura biològica actual és el resultat dels esforços de classificació dels éssers vius —organismes (plantes i animals) i microorganismes (bacteris, fongs, virus, etc.)— que Linneus va iniciar durant el segle XVIII i que es recull en el *Codi Internacional de Nomenclatura*. Més concretament, la proposta i ús dels zoònims estan regulats pel *Codi de Nomenclatura Zoològica*, la dels classificats com a "plantes" pel *Codi de Nomenclatura Botànica* i els classificats com a "bactèries" pel *Codi Internacional de Nomenclatura Bacteriològica*. Les regles d'aquestes nomenclatures no són vinculants legalment per lleis, sinó que es tracta d'acords voluntaris presos per diferents comissions internacionals. Cal assenyalar també que aquestes nomenclatures no són estables ja que estan sotmeses a canvis i reorganitzacions.

En el text analitzat, hem documentat 35 noms llatins diferents marcats per l'especialista que formen part o de la nomenclatura zoològica o de la bacteriològica. Aquesta xifra representa el 4,93% del total de les USE del text marcades per l'especialista. És important remarcar que ni les unitats monolèxiques verbals, ni les adjectives ni les adverbials superen aquesta xifra.

És lògic que en un text sobre malalties infeccioses es faci referència a determinats organismes perquè la causa de qualsevol malaltia infecciosa és o un bacteri o un virus[29].

Tant els organismes animals com els bacteris tenen un nom científic en llatí, que els identifica inequívocament i internacionalment, i un nom *semicientífic*. En els textos mèdics sobre malalties infeccioses sovint s'usen els noms llatins perquè són la manera més inequívoca d'identificar els agents de la malaltia:

> *El tifus murí és causat per Rickettsia typhi. La reserva n'és la rata peridomèstica (comunament, Rattus norvegicus i Rattus rattus), a partir de la qual es pot transmetre a través de la puça de la rata (Xenopsyllachopis). La puça s'infecta quan ingereix sang de les rates infectades; R. typhi es multiplica a les cèl·lules intestinals de l'artròpode, i és excretada a través de la femta (...)*
>
> [Farreras-Rozman, 1997: 2400]

Quinze dels trenta-cinc noms llatins del text corresponen a zoònims i vint a bacteris. Alguns estan denominats per la família: *Rickettsiaceae*; d'altres pel gènere: *Bartonella, Coixella, Hyalomma, Ixodes, Leptotrombidium, Rickettsia*; d'altres per l'espècie: *Bartonella quintana, Rickettsia tsutsugamushi, C. burnetii, R. burnetii, Rattus rattus;* i, fins i tot, n'hi ha que per la subespècie: *Pediculus humanus corporis.*

En el text trobem també dues expressions llatines, escrites en cursiva, que tenen un significat especialitzat pertinent en medicina: *in vivo* i *in vitro.* I, en canvi, no hem trobat cap altra classe de nomenclatura o d'expressió llatina, per bé que en altres textos especialitzats sobre temes mèdics

[29] En el cas de les malalties infeccioses per *Rickettsia,* l'agent de la malaltia és un microorganisme de la família de les *Rickettsiaceae*. Les *Rickettsies* són bacteris de la família de les *Rickettsiaceae;* dintre del gènere de les *Rickettsies* tenim diferents espècies, que són agents de malalties específiques: *Rickettsia burnetii, Rickettsia quintana, Rickettsia tsutsugamushi,* etc. Les *Rickettsies* necessiten un vector de transmissió que infecti a l'ésser humà. Els vectors de transmissió de les Rickettsies solen ser artròpodes o àcars. Alhora aquests vectors s'instal·len en altres animals que moltes vegades contenen la reserva de *Rickettsia*.

(encara que no sigui molt usual) podem trobar també els noms de les parts del cos humà en llatí:

> *Existeix des de 1895 una Nomina Anatomica. Aquesta primera versió va ser aprovada a Basilea per l'Anatomische Gesellschaft alemana. Aquesta Nomina Anatomica es basa "en unos principios semejantes a los de las otras nomenclaturas normalizadas. Todos los términos de su lista oficial están redactados en latín (aunque se deja a cada país en libertad de traducirlos en idiomas modernos para su uso docente y no especializado), son únicos para cada estructura anatómica, cortos y sencillos en lo posible, e intentan no ser meras expresiones simbólicas, sino tener algún valor informativo o descriptivo."*
>
> [López Piñero i Terrada Ferrandis, 1990: 75].

3.3 Conclusions

En aquest capítol, hem examinat el silenci que generen els SEACAT basats exclusivament en patrons morfosintàctics. Hem vist que hi havia dos tipus de silenci en relació amb l'objecte d'extracció de la majoria d'aquests sistemes: **intrínsec i extrínsec.**

Des del punt de vista del SEACAT, hem arribat a la conclusió que és més important solucionar el silenci intrínsec que no pas l'extrínsec; és a dir, interessa més controlar que no hi hagi cap unitat que, d'acord amb el seu objecte d'extracció —les UTP—, hauria d'haver estat detectada, però no ho ha estat.

Des del punt de vista de l'usuari, en canvi, creiem que el més interessant és trobar criteris i mecanismes perquè un sistema informàtic superi el silenci extrínsec, és a dir tot allò que en un text especialitzat és especialitzadament pertinent —tant si és o no una UTP.

Hem mostrat i exemplificat que el **silenci intrínsec** que produeix un SEACAT està motivat per una de les **causes** següents:

- errors en la fase de desambiguació morfològica
- termes superposats
- termes amagats.

En canvi, la causa del **silenci extrínsec** s'ha de buscar en la definició de l'objecte mateix del sistema d'extracció automàtica. Els dissenyadors d'aquests sistemes solen limitar l'objecte de detecció a l'extracció de la UTP que, per bé que sigui la unitat més freqüent de les unitats especialitzades, no és l'única. La diversitat de les unitats especialitzades (pel que fa a la seva naturalesa, la categoria gramatical i l'estructura) que s'usen en els textos especialitzats condueix a pensar que l'objecte d'un SEACAT ha de ser totes les unitats de significació especialitzada d'un text, i no només les unitats terminològiques polilexemàtiques.

Per això, hem analitzat totes aquelles unitats especialitzades que no eren UTP —és a dir, USE monolèxiques, sigles, unitats fraseològiques, símbols, noms en llatí— per tal de trobar recursos que ajudessin un SEACAT a reduir el silenci extrínsec. En aquesta línia, hem evidenciat que el domini de les ciències de la salut reuneix tres característiques que poden facilitar la tasca automatitzada de reconeixement i extracció de les unitats especialitzades d'un text:

1. L'**ús de nomenclatures normalitzades**. La complexitat de la medicina (en el seu vessant teòric, aplicat i pràctic) justifica l'ús de nomenclatures mèdiques relatives a malalties, procediments diagnòstics, procediments terapèutics. Però també l'ús de nomenclatures de camps lèxics que constitueixen els fonaments de la medicina: la nomenclatura de l'anatomia, de la química, de la zoologia, de la botànica, de la bacterologia, de la virologia. Les primeres no estan tan consensuades com les segones; aquestes

últimes, a més, estan basades en classificacions molt fermament establertes.

2. L'**ús dels formants grecollatins**. El fet que més del 60% del lèxic mèdic estigui basat en un nombre d'arrels, de prefixos i de sufixos grecollatins limitat avala la nostra proposta que un extractor amb un diccionari de formants (la xifra s'estima al voltant dels 1.100 constituents) amb informació semàntica i regles de combinació dels formants reduiria de manera substancial el silenci.

3. L'**ús d'heurístiques de morfologia lèxica**. En concret, ens referim al paper que tenen alguns sufixos patrimonials quan s'adjunten a unes bases determinades en el marc d'un domini semàntic delimitat. Per citar un dels exemples comentats, els únics adverbis amb significat especialitzat són els que s'han format a partir d'un adjectiu de caràcter especialitzat i el sufix *-ment* i que, normalment, signifiquen “des del punt de vista + A”.

Per tant, hem arribat a la conclusió que si un SEACAT conegués la llista exhaustiva de les nomenclatures estandaritzades, l'inventari de formants grecollantins habituals en medicina, convenientment caracteritzats formalment i semànticament, i les regles que relacionen determinats sufixos adjuntats a determinades bases freqüents en medicina, podria aconseguir reduir el silenci que generen els SEACATS actuals.

D'altra banda, hem comprovat que les sigles especialitzades són molts abundants en els textos mèdics i per donar-ne compte un SEACAT hauria d'ancorar-se en la seva tipografia i la disposició d'aquestes unitats en el text.

En resum, al llarg d'aquest capítol hem defensat la idea que, a l'hora de dissenyar un SEACAT, s'hauria de tenir en compte **totes les USE que poden aparèixer en un text especialitzat**, tant si pertanyen al codi de la llengua natural (unitats lèxiques i unitats fraseològiques) com si no hi pertanyen (símbols, nomenclatures), tant si són simples com complexes, tant si són lèxiques com fraseològiques. També hem demostrat que totes les USE ofereixen alguns recursos formals i semàntics per al seu reconeixement.

3.4 Recapitulació

Recapitulant: hem partit de la hipòtesi que els SEACAT actuals que es basen en patrons morfosintàctics no detecten totes les unitats d'un text, de manera que més del 60% de les USE d'un text no són identificades pels sistemes d'extracció automàtica. Hem vist com una part d'aquest silenci depèn de la concepció de l'extractor, però una altra part no està controlada.

A continuació, hem buscat les causes que provoquen el silenciament d'alguns tipus d'USE i hem analitzat el tipus d'USE silenciades a fi de proposar elements morfològics, formals, tipogràfics, discursius i semàntics que poden ajudar un SEACAT a augmentar la seva cobertura. Malgrat tots aquests recursos, pensem que cal aprofundir encara més en:

1. La representació, disposició i relació de les USE en els textos, com les unitats es presenten en els textos i com es relacionen entre elles.

2. Els diversos aspectes semàntics de les USE.

Com és sabut, els SEACAT s'avaluen per la seva eficiència que es valora pels índexs de silenci i de soroll que generen. Si en aquest capítol ens hem centrat en el silenci, en el proper capítol abordarem el soroll.

4. EL SOROLL: SEGMENTS NO TERMINOLÒGICS PROPOSATS PER UN SEACAT COM A UNITATS TERMINOLÒGIQUES

Defender lo real de lo irreal, lo verdadero de lo falso, lo que es, de lo que no es.

Calidescopia [1997: 249]

La terminología teórica y práctica se enfrenta, desde hace tiempo, a dos problemas básicos para el tratamiento de las unidades terminológicas: la detección de términos y la segmentación de unidades terminológicas de estructura compleja. Hasta ahora, la terminología sólo disponía de la consulta con el especialista para decidir sobre la identificación definitiva de un término, sobre su correcta segmentación y sobre la pertinencia de este término en un ámbito de especialidad. A estos dos problemas hay que añadir la complejidad que presenta la automatización de estos procesos.

Projecte DGES PB 96-0293

En el capítol anterior hem analitzat un dels problemes dels sistemes d'extracció automàtica de candidats a terme (SEACAT): el silenciament d'algunes de les unitats de significació especialitzada d'un text. Però, com avançàvem, aquest no és l'únic inconvenient dels SEACAT actuals, els quals, a més de silenci, generen una quantitat important de soroll, és a dir, proposen com a candidats a termes una gran quantitat de *falses* unitats terminològiques. A partir d'un corpus textual, aquests sistemes proporcionen llistes *brutes* de sintagmes presumptament terminològics que l'usuari ha de validar manualment decidint quins segments de la llista rebutja perquè no són termes i quins altres segments o parts de segments reté perquè són unitats terminològiques (UT).

En general, els SEACAT basats en coneixement lingüístic (sobretot en patrons morfosintàctics, que són la majoria) generen uns percentatges alts de soroll: entre el 45% i el 75% dels candidats proposats per aquests programes s'han de rebutjar.

Aquests resultats obliguen a plantejar-se dues qüestions: *Què és el que provoca soroll? Quin tipus d'unitats són les que sistemàticament s'han de "rebutjar"?*

Els objectius d'aquest capítol —que deriven directament d'aquestes dues preguntes— tenen com a finalitat la cerca d'elements que permetin a un SEACAT extreure llistes de candidats a terme més *netes*. En aquest capítol, doncs, ens proposem:

- aclarir les causes que provoquen el soroll, i
- estudiar i delimitar les unitats concretes que l'ocasionen.

Per respondre a aquests propòsits, partirem dels resultats que s'obtenen de l'aplicació al mateix corpus que hem utilitzat als capítols 2 i 3 d'un SEACAT que fa servir patrons positius i d'un altre basat en patrons negatius[1].

Per saber si una unitat identificada per un SEACAT és pertinent com a unitat especialitzada, hem utilitzat quatre criteris que són progressivament excloents entre si:

- En primer lloc, hem comparat els resultats del buidatge automàtic amb els del manual que, prèviament, havia fet un especialista[2].
- Si el primer criteri no ha estat suficient per saber si una unitat és especialitzada o no, ens hem guiat per la nostra competència lingüisticocognitiva.
- Si tot i així, encara no ho podíem determinar, hem consultat obres de referència especialitzada com el *Diccionari Enciclopèdic de Medicina* (1990)

[1] Recordem, però, les principals característiques del corpus textual d'anàlisi i de les eines informàtiques:

- El corpus textual està format per tots els capítols sobre malalties infeccioses del llibre *Medicina Interna* de Farreras-Rozman (1997). Es tracta d'un document molt especialitzat, adreçat tant a professionals com a estudiants de medicina i compost de 60.948 ocurrències.
- Les eines informàtiques utilitzades són EXCAT1 i CERPAT.

[2] El buidatge manual s'ha realitzat sobre el subcorpus: *Malalties infeccioses per* Rickettsia. Per a més informació, vegeu els apartats 2.2 i 2.3 del capítol segon.

- I, finalment i en cas de dubte, hem consultat diversos especialistes en medicina.

Normalment, tractant-se dels SEACAT, quan es parla de soroll es fa referència a segments de discurs integrats per diversos mots que presenten una estructura morfosintàctica idèntica a la d'una UTP, però que, en canvi, no tenen caràcter terminològic: són **unitats disfressades d'UTP que despisten un SEACAT que no té altre recurs per detectar-les que les estructures morfosintàctiques i la freqüència d'ús**. Aquesta restricció neix del fet que el soroll sempre es mesura en relació amb l'objecte de detecció d'un SEACAT que, en general, es redueix a la UTP.

4.1 La causa principal del soroll: les estratègies de detecció

Per analitzar el soroll que generen els SEACAT de base lingüística estàndards, hem tingut en compte una sèrie de premisses i arguments que, de manera encadenada, ens han portat fins a la causa principal del soroll:

1. L'objecte d'extracció dels SEACAT és fonamentalment la **UTP.**
2. El soroll mesura el percentatge de **segments disfressats d'UTP** d'un corpus textual en relació amb el nombre de candidats a UT proposats per un SEACAT.
3. Per reconèixer una UTP, la majoria de SEACAT es basen en **patrons morfosintàctics**[3] i en la freqüència d'ús.
4. Tots els SEACAT generen soroll perquè els patrons morfosintàctics de les UTP **no són exclusius:** hi ha altres

[3] Els patrons morfosintàctics tant poden ser positius (i per tant de les UT) com negatius (és a dir de les unitats que no poden ser mai terminològiques).

unitats no terminològiques, especialitzades o no, que presenten les mateixes estructures morfosintàctiques que les UT.

5. Consegüentment, s'hauria de recórrer a d'altres paràmetres (a part de la categoria gramatical i l'estructura morfosintàctica) si es vol discriminar les UTP reals.

D'aquestes constatacions, se'n pot deduir que el factor principal del soroll prové del concepte mateix d'UT amb què els SEACAT treballen. En realitat, per als SEACAT el terme és una forma exclusiva d'un àmbit especialitzat; i, del seu vessant formal, la majoria de sistemes només tenen en compte l'estructura sintagmàtica. Des del punt de vista lingüístic, però, una UT és l'associació d'una forma i d'un contingut i no només una forma. Per això, l'estructura formal d'una UTP —tot i que és un indici probabilístic— no és un element suficient que serveixi per discriminar-les d'altres classes d'UT.

En la nostra hipòtesi, la causa del soroll rau, doncs, en la incapacitat de discriminar les UTP a partir de les estructures morfosintàctiques, i aquest fet està suscitat per les restriccions de base de les quals parteixen aquests sistemes.

En els apartats següents analitzarem els tipus d'unitats que provoquen soroll. Per realitzar aquest estudi, ens centrarem en les estructures morfosintagmàtiques de les UTP més freqüents en els discursos especialitzats[4], que són en el 98%:

- [N [SAdj]]$_{SN}$
- [N [SPrep]] $_{SN}$

[4] Òbviament, les UT monolèxiques no ocasionen soroll perquè, actualment, no són l'objecte de detecció dels SEACATS. En l'apartat 3.2.1 del capítol anterior hem vist que les USE monolèxiques ocasionaven silenci extrínsec a l'objecte d'extracció d'un SEACAT.

Aquestes dues estructures, com hem vist abans, es poden desglossar en cinc subestructures morfològiques:

- [N[A]$_{SAdj}$]$_{SN}$
- [[N[A]$_{SAdj}$]$_{SN}$ [A]$_{SAdj}$]$_{SN}$
- [N [de N]$_{SPrep}$] $_{SN}$
- [N [de art N]$_{SPrep}$] $_{SN}$
- [N [de N$_{propi}$]$_{SPrep}$] $_{SN}$

Les estructures [N [[de N]$_{SPrep}$ [A]$_{SAdj}$]$_{SPrep}$] $_{SN}$ i [N [[de art N]$_{SPrep}$ [A]$_{SAdj}$]$_{SPrep}$] $_{SN}$ són terminològicament poc productives i, en canvi, produeixen molt de soroll, perquè la majoria d'unitats que presenten una d'aquestes dues estructures, encara que inclouen alguna USE, no són especialitzades. En molts casos, el sintagma nominal que funciona de complement sol ser una UT i, en canvi, el nucli nominal de la seqüència, no. Per aquest motiu, també les tractarem en aquest capítol.

4.1.1 Unitats lingüístiques amb estructures morfosintàctiques idèntiques a les UTP

A priori, les unitats filtrades per alguns d'aquests patrons haurien de correspondre a UT, però això no passa sempre perquè, com dèiem, aquestes estructures no són exclusives. Aquest fet el constatem també en el corpus analitzat en el qual hi ha diferents tipus d'unitats que comparteixen les mateixes estructures:

- UTP (*medul·la òssia*, *meningitis bacteriana*, *malaltia de Brill-Zinsser*, *prova de la immunoperoxidasa*, *sistema mononuclear fagocític*, etc.)

- UFE (*acumulació de líquid extravascular, augment de la permeabilitat vascular, extravasació de líquid intravascular, factor de necrosi tumoral*, etc.)
- combinacions especialitzades recurrents (*radiografia de la mà, massatge a les cervicals*, etc.)
- unitats discursives (UD) (*augment del nombre de casos, dècades dels anys trenta, distribució geogràfica, manera específica, resultats retardats, banys freqüents, color vermellós, temps addicional, estat general, escorxador de Brisbane*, etc.)
- USE no pertinents per a l'àmbit temàtic del text (*blau de metilè, conca mediterrània, treballadors socials, condicions de vida, classe social, canvi climàtic*, etc.)[5].

Des del nostre punt de vista, els tres primers tipus d'unitats tenen caràcter especialitzat perquè són unitats de significació especialitzada (USE), encara que només les UTP tenen valor referencial.

També pot ocórrer que només una part del segment sigui terminològicament pertinent, i en aquest cas, el caràcter pertinent pot tenir-lo el nucli i/o el complement com mostren els exemples següents:

> ***biòpsia*** *de la* ***pell***[6], *presència de* ***rickèttsies***, *peculiaritats* ***clíniques***, *existència de la* ***taca negra***, ***zoonosi*** *de distribució mundial, durada de l'****exantema***, ***tractament*** *del* ***tifus epidèmic***, *prevenció de la* ***malaltia***, *importància* ***epidemiològica***, ***tractament antimicrobià*** *adequat*, etc.

[5] En aquest grup tant incloem les UL com les UF no pertinents per a l'àmbit temàtic del text.
[6] Hem marcat en negreta les UT.

En el cas de *biòpsia de la pell* i de *tractament del tifus epidèmic,* tant el nucli com el complement són aïlladament UT, però la seva combinació no dóna lloc a UT, sinó a combinacions recurrents especialitzades o a UFE.

D'aquestes estructures, les que provoquen més soroll, per ordre ascendent, són:

- [N [[de art N]$_{SPrep}$ [A]$_{SAdj}$]$_{SPrep}$] $_{SN}$
- [N [de art N]$_{SPrep}$] $_{SN}$
- [N [[de N]$_{SPrep}$ [A]$_{SAdj}$]$_{SPrep}$] $_{SN}$
- [N [de N]$_{SPrep}$] $_{SN}$

La presència d'article davant del complement és sovint un indici que la unitat no està del tot lexicalitzada i, per tant, les estructures en què el complement està introduït per un article determinat percentualment tendeixen a generar més soroll que les estructures en què el complement és indeterminat. En canvi, les estructures [N[A]$_{SAdj}$]$_{SN}$ i [[N[A]$_{SAdj}$]$_{SN}$ [A]$_{SAdj}$]$_{SN}$ generen menys soroll, cosa que no vol dir que en generin poc.

Com indica Sager (1993: 121), les UTP contribueixen a la construcció de sistemes terminològics, de tal manera que el nucli de la unitat indica la categoria a la qual correspon el concepte, i el complement determina el criteri utilitzat per a la subdivisió de la categoria. A continuació, analitzarem cada una d'aquestes estructures, fixant-nos per separat en el nucli i en els complements, per tal de cercar elements sistematitzables referents a les estructures de les UTP més productives i al soroll que aquestes estructures generen.

4.2 El soroll de l'estructura [N[A]$_{SAdj}$]$_{SN}$

L'estructura $[N[A]_{SAdj}]_{SN}$ és la més freqüent de les UTP en els textos de ciències de la salut i, en general, en els de tots els dominis especialitzats. És també l'estructura morfosintàctica que transmet més sintèticament un concepte especialitzat. Així, davant d'unitats com *anatema faringi*, *anatema de faringe*, *anatema de la faringe* o *episodi febril* i *episodi de febre* —totes documentades en el corpus—, l'usuari percep com a més lexicalitzades les unitats *anatema faringi* i *episodi febril*, formades per un nom i un adjectiu, que no pas les altres.

En aquesta intuïció del parlant recolza la hipòtesi que l'estructura $[N[A]_{SAdj}]_{SN}$ és l'esglaó últim d'un procés progressiu de lexicalització sintàctica:

frase (nucli V complement) N ~~de~~ → art N ~~N de~~ → N N A →

Les paraules de Sager que reproduïm a continuació il·lustren les característiques d'un procés de terminologització d'un concepte[7]:

> *La evolución de los conceptos está acompañada de fases de denominación, un proceso que se llama terminologización. En el desarrollo del conocimiento, los conceptos de materias científicas y tecnológicas así como los de otras disciplinas, sufren ciertos cambios; en consecuencia sus formas lingüísticas son flexibles hasta que un concepto termina de formarse y se incorpora de forma más o menos permanente en la estructura del conocimiento.*

[7] En molts àmbits d'especialitat, com és el cas de les ciències de la salut, hi ha una voluntat explícita, per part dels mateixos especialistes, que les unitats que serveixen per denominar els conceptes especialitzats siguin, des del punt de vista semàntic, tan descriptives i transparents com sigui possible:

> *El proceso de la observación y descripción científica incluye la designación de conceptos y esto, a su vez, conlleva un nuevo examen del significado de las palabras, junto con el cambio de las designaciones y la acuñación de otras nuevas. Esta preocupación por la manipulación de las formas lingüísticas conduce a un intento de reflejar en el lenguaje elementos del pensamiento y de la percepción. Por lo tanto, la designación dentro de los lenguajes especializados tiene como objetivo la transparencia y la consistencia. A menudo se hacen tentativas para que las designaciones reflejen en su morfología y estructura los rasgos conceptuales o las características principales de los conceptos que representan.*
>
> [Sager, 1993: 91-92]

[Sager, 1993: 95-96]

Això no vol dir, però, que tots els noms dels conceptes especialitzats hagin de seguir sempre totes les fases:

> *Una colocación originariamente libre, como en box for a tool, puede reducirse poco a poco a su forma más concentrada, por ejemplo tool box, tool-box, toolbox. No siempre se siguen todas estas fases de forma sistemática, ni tampoco todos los compuestos llegan a adoptar, a la postre, la forma de una sola palabra ortográfica. A pesar de la brevedad que proporciona este tipo de reducción a una simple yuxtaposición de nombres, hace que se pierda cierta transparencia en la comprensión.*
>
> [Sager, 1993: 104]

L'estructura sintagmàtica formada per un nom i un adjectiu en els textos especialitzats és especialment indicada en les llengües romàniques perquè compleix dues condicions:

- és l'estructura més productiva
- és l'estructura més terminologitzable.

Amb la finalitat d'estudiar el soroll que provoca aquesta estructura, hem extret —mitjançant EXCAT1— tots els segments que en el corpus tractat corresponen a un sintagma nominal integrat exclusivament per $[N[A]_{SAdj}]_{SN}$.

El resultat d'aquesta cerca és que en aquest corpus hi ha 2.750 unitats amb una estructura sintàctica $[N[A]_{SAdj}]_{SN}$, xifra que representa un 17,71% del total d'ocurrències del corpus que presenten una estructura morfosintàctica candidata a UT[8]. D'aquestes unitats, aproximadament

[8]Hem fet la mateixa cerca amb CERPAT i el resultat ha estat que 3.878 segments responien a l'estructura NA. Aquesta diferència de 1.128 unitats respon al fet que EXCAT fa servir una estratègia de detecció diferent a CERPAT. Mentre que EXCAT cerca seqüències terminològiques de longitud màxima i les seqüències són mútuament excloents, CERPAT busca una seqüència en abstracte sense tenir en compte cap aspecte terminològic. Si CERPAT comptés amb un analitzador sintàctic la selecció de les unitats

unes 1.000 (36%) són UT (USE nominals amb caràcter referencial), unes 500 (18%) són unitats discursives (UD) i les 1.250 restants (45%) són USE sense valor referencial, és a dir, UFE nominals o combinacions especialitzades recurrents nominals.

Si analitzem amb més precisió les 2.750 ocurrències trobem efectivament una gran varietat de possibilitats:

1. **Segments que tenen com a nucli una UT formada a la manera culta** (amb formants i/o afixos grecollatins) **seguida d'un complement adjectiu de diversos tipus**: *bronquitis crònica*, *broncograma aeri*, *aspergil·loma sinusal*, *biòpsia ganglionar*, *cefalea intensa*, *diarrea lleugera*, etc. Dintre d'aquest grup, podem distingir entre:

    1.0 Les unitats en què l'adjectiu s'ha format a partir d'un terme: *colitis ulcerosa*, *endocarditis bacteriana*, *hiperplàsia cel·lular*, *parènquima pulmonar*, etc.

    1.1 Les unitats en què l'adjectiu no és especialitzat, en el sentit que aïlladament no té contingut especialitzat sinó que només té valor especialitzat quan forma part d'una USE polilèxica: *parasintèmia perifèrica*, *meningitis subaguda*, *diarrea greu*, *apendicitis aguda*, etc.

2. **Segments que tenen com a nucli una UT simple seguida d'un complement adjectiu de diversos tipus**: *faringe humana, febre alta, febre botonosa, febre tifoide, febre remitent, fetge normal, lòbul esquerre, llavi major, medul·la òssia,*

---

seria més precisa. Actualment, però, no existeixen analitzadors sintàctics per al català, i en altres llengües encara són poc eficients.

òrgan pelvià, etc. A l'interior d'aquest grup podem discriminar entre:

2.1 Les unitats en què l'adjectiu s'ha format a partir d'un terme: *glàndula salival, pelvis renal, nervi òptic, rubèola congènita, tifus endèmic*, etc.

2.2 Les unitats en què l'adjectiu no és especialitzat: *abdomen agut, pell seca, cor dret, dolor intens, dolor local, lòbul esquerre, lòbul inferior, lòbul superior*, etc.

3. **Segments que tenen com a nucli un substantiu "no especialitzat" i com a complement un adjectiu especialitzat**: *control serològic, cavitat nasal, conducte biliar, dada clínica, defensa humoral, dificultat respiratòria, esquema immunològic, fase eritrocítica, manifestacions digestives, medi vaginal, material proteic, marc endèmic, origen hematogènic, origen infecciós, llum ultraviolada, paret toràcica, passatge hepàtic, regió perianal, regió renal, resposta immunitària*, etc.

4. **Segments que tenen com a nucli un substantiu deverbal**: *alteracions immunitàries, afecció cardíaca, complicació extrapulmonar, confirmació diagnòstica, envermelliment faringi, embassament pleural, manifestacions al·lèrgiques, propagació transdiafragmàtica, reacció inflamatòria*, etc. En aquests exemples, observem que l'adjectiu tendeix a ser un adjectiu denominal i que gairebé sempre té caràcter especialitzat.

5. **Segments en què ni el nucli ni el complement són especialitzats.** En aquest cas, observem dos tipus d'unitats diferents:

5.1 USE que en l'àmbit temàtic del text no tenen contingut especialitzat: *aigües residuals*, *animal domèstic*, *classe social*, *congelador domèstic*, *desastre natural*, *medi ambient*, *raça blanca*, *àrea rural*, *gent gran*, etc.

5.2 Unitats discursives (UD): *aparició tardana*, *aspecte interessant*, *aspecte comú*, *canvi reversible*, *criteri següent*, *difusió mundial*, *direcció paral·lela*, *factor fonamental*, *fet característic*, *setmana anterior*, etc.

Les unitats del primer, del segon i del quart grup són sempre UT, tant les constituïdes per un adjectiu especialitzat com aquelles en què l'adjectiu és general. Les ocurrències del cinquè apartat no són UT (algunes són unitats discursives i d'altres són unitats lèxiques que en altres textos poden ser especialitzades, però que en aquest corpus els manca valor especialitzat). Finalment, les ocurrències del tercer grup són *unitats* difícils de classificar, perquè algunes poden ser considerades UT, d'altres combinacions nominals recurrents i unes altres UFE nominals, encara que totes les unitats d'aquest tercer grup són USE, és a dir unitats amb un significat especialitzat. Els segments d'aquest tercer grup són, doncs, els més difícils de classificar sobretot automàticament[9].

4.2.1 Nuclis nominals d'una UTP

Des del punt de vista de la complexitat formal, els nuclis d'una UTP poden ser:

[9] Segons la naturalesa del nucli, Sager (1993: 122) distingeix entre UTP objectes, UTP propietats i UTP processos i operacions. Així, seguint amb l'esquema anterior, observem que, en medicina, les unitats del primer i del segon grup solen ser objectes, les unitats del tercer grup poden ser objectes, però també —encara que en molta menys freqüència— propietats i, finalment, les unitats del quart grup corresponen sobretot a processos, accions i operacions.

- substantius simples
- substantius complexos.

Els substantius simples són difícils de detectar lingüísticament perquè són idiosincràtics. En canvi, totes les UT $[N[A]_{SAdj}]_{SN}$ que tenen com a nucli un substantiu derivat o bé un substantiu construït a partir d'un formant clàssic i/o d'un afix grecollatí es poden controlar més fàcilment.

Les USE monolèxiques, tant simples com derivades o compostes, molt sovint constitueixen el nucli —l'hiperònim o el merònim— de moltes altres USE més complexes. En el *Diccionari Enciclopèdic de Medicina* (1990), per exemple, hi ha 18 UT que tenen com a nucli *bronquitis*, totes subespecificades amb un adjectiu, excepte una que va acompanyada d'un nom propi introduït per la preposició *de*[10].

Si seguim analitzant el *Diccionari Enciclopèdic de Medicina* (1990), confirmem aquesta afirmació: la majoria de mots monolèxics (simples, derivats o compostos cultes) són també el genèric de diverses UTP[11]. Les UT simples són encara molt més productives: el diccionari inclou 320 tipus de *cos*, 153 UTP en què el nucli és el terme *febre* i 1.143 UTP en què el nucli és *malaltia*. I aquesta informació és la que recull només un sol diccionari, perquè la tipologia que apareix en els textos mèdics és encara molt més rica.

[10] *bronquitis asmàtica, bronquitis capil·lar, bronquitis crònica, bronquitis fibrosa, bronquitis sinoï dal, bronquitis hemorràgica, bronquitis infecciosa, bronquitis membranosa, bronquitis obliterant, bronquitis plàstica, bronquitis productiva, bronquitis pseudomembranosa, bronquitis pútrida, bronquitis verinosa, bronquitis vesicular i bronquitis de Castellani* (aquesta última *bronquitis* també és coneguda com a *broncospiroquetosi*).

[11] Vegem-ne exemples, escollits a l'atzar, que mostren la productivitat dels termes compostos clàssics: 50 UT que tenen com a nucli *meningitis;* 11 que tenen com a nucli *hemòlisi;* 103 que tenen com a nucli *encefalitis;* 28 tipus d'*endocarditis;* 42 tipus de *glaucoma;* 67 tipus de *nefritis;* 34 classes d'*esclorosi;* 4 tipus de *galvanòmetre;* 8 d'*oftalmoscopi;* etc.

Si ens centrem en els nuclis nominals, els més productius són els noms derivats i, sobretot, els noms compostos a la manera culta (amb formants i afixos grecollatins). És veritat que els símbols —*ph àcid*—, les sigles —*DNA rickettssià*— i els substantius formats per conversions sintàctiques —*cultiu cel·lular*— també poden ser el nucli d'una UTP, però aquests casos són molt menys nombrosos perquè, majoritàriament, les UT de l'àmbit de la medicina es formen a partir de formants cultes.

4.2.1.1 Els formants grecollatins

En ciències de la salut, com hem vist en el capítol anterior, un nombre reduït d'afixos i de formants grecs i llatins serveixen per formar un nombre molt elevat de termes no sintagmàtics:

> *En torno a mil raíces de procedencia griega o latina componen la casi totalidad de los términos médicos de origen clásico y de los neologismos.*
> [López Piñero i Terrada Ferrandis, 1990: 29]

Si observem la classificació semàntica que Bernabeu i al. (1995) fan dels formants clàssics pertinents en medicina, deduïm que la xifra més elevada de formants clàssics (tant d'arrels com d'afixos) correspon a parts anatòmiques dels cos humà seguits dels que fan referència als humors, les excrecions, les secrecions, les funcions orgàniques i les malalties. La resta de formants es reparteixen, bàsicament, entre els que es refereixen a animals, vegetals, elements i compostos químics, fenòmens i agents físics, objectes, operacions i accions, formes, colors, nombres, mides i quantitats.

Les ciències biomèdiques, com a domini especialitzat, disposen d'un vocabulari extens, que es pot, però, aprendre i deduir si es coneix un nombre limitat d'arrels, prefixos i sufixos grecs i llatins. A partir d'aquests termes cultes, se'n poden construir molts altres. Aquesta hipòtesi pressuposa que si un professional de les ciències de la salut arriba a

dominar aquest conjunt de formants (que es calcula al voltant de les 1.200 unitats entre afixos i arrels) podrà:

- reconèixer i entendre la majoria d'USE del seu domini especialitzat, i
- construir termes nous adequats a un paradigma conceptual concret.

Com molt bé explica Stanaszek i al. (1996: XIII),[12]:

> *La identificación de un término a través de su análisis estructural implica determinar el significado de cada uno de sus componentes que revelarán tanto la definición exacta de la palabra como el significado que transmite:*
>
> *Miopatía = mio (raíz que significa músculo)*
> *+ patia (sufijo que significa enfermedad)*
>
> *Algunas palabras comprenden más de una raíz, cada una de las cuales retiene su significado básico. Dichas palabras son muy comunes en la terminología médica:*
>
> *Osteoartritis= Inflamación de las articulaciones de los huesos*
> *Osteo= Hueso (raíz)*
> *Arthro= Articulación (raíz)*
> *Itis= Inflamación (sufijo).*

Per tot això, defensem la mateixa hipòtesi que hem proposat en el capítol anterior: si un SEACAT compta amb un vocabulari d'uns mil formants i

[12] Aquesta és la raó per la qual, en el marc de les ciències de la salut, existeixen diversos manuals adreçats als professionals —[López Piñero i Terrrada Ferrandis, 1990], [Love i Davis, 1990], [Quintana, 1989], [Navarro, 1996], [Bernabeu i al., 1995], etc.— dedicats a l'ensenyament dels termes "mèdics" a partir del sistema de formació de paraules amb formants grecollatins. En la introducció d'un d'aquests manuals, llegim *"La terminología médica es el conjunto de términos utilizados por los profesionales de la medicina en todo el mundo. Un cálculo preciso de su volumen plantea serias dificultades, pero resulta orientador saber que los diccionarios médicos generales más importantes incluyen entre 40.000 y 100.000 vocablos. Los especialistas en educación médica estiman que los estudiantes del período preclínico deben aprender alrededor de 15.000, cifra incomparablemente superior a la del vocabulario de un curso básico de un idioma extranjero, que no suele llegar a las 5.000 palabras. Este es uno de los factores que explica la importancia creciente que durante las últimas décadas se está concediendo en casi todos los países a la enseñanza de la terminología médica"*. [López Piñero i Terrada Ferrrandis, 1990: XI]

un nombre reduït de regles de combinació, podrà simular el procés de descodificació que utilitza el professional i, d'aquesta manera, reduir substancialment el soroll produït per les estructures morfosintàctiques, en aquest cas corresponents a possibles UTP.

4.2.1.2 Nuclis nominals de caràcter no especialitzat

Paral·lelament a aquests substantius de base culta, en els textos mèdics trobem també un conjunt força important de nuclis nominals que **aïlladament** no són especialitzats, però que segons l'adjectiu amb el qual es combinen poden constituir una UTP. En aquests casos, l'adjectiu serveix per donar caràcter especialitzat a la unitat i classificar-la taxonòmicament. Ens referim a substantius com *prova*, *paret*, *aparell*, *espècie*, *sistema*, *dada*, *imatge*, *massa*, *patró*, *regió*, *torrent*, *zona*, etc.

Aquestes unitats són polisèmiques, tant perquè segons el context d'ús poden tenir diversos sentits especialitzats, com perquè poden tenir, a més dels sentits especialitzats, un sentit general. Així, segons l'adjectiu que les modifica poden ser UD o USE d'un domini concret[13].

Aquests noms de caràcter no especialitzat, però, combinats o modificats per un adjectiu especialitzat classificador es converteixen en una UTP[14]:

USE pertinents en el textos mèdics	**UL o UD no pertinents des del punt de vista especialitzat en els**

[13] *sistema aritmètic* (UT des del punt de vista de les matemàtiques), *sistema respiratori* (UT des del punt de vista de la medicina), *sistema operatiu* (UT des del punt de vista de la informàtica), *sistema elèctric* (UT des del punt de vista de la electricitat), *sistema autonòmic* (UT des del punt de vista polític) *sistema filonià* (UT des del punt de vista de la geometria), *sistema expert* (UT des del punt de vista de la intel·ligència artificial), *sistema local* (UT des del punt de vista de l'astronomia), *sistema barat* (unitat discursiva), etc.

[14] Independentment del fet que en discurs, per reducció anafòrica, es pugui obviar l'adjectiu.

	textos mèdics
paret cel·lular, paret endotelial, paret ventricular	paret groga, paret mestre
patró alveolar, patró enzimàtic	patró modèlic, patró sintàctic
índex bacterià, índex cerebral, índex endèmic	índex elevat, índex alt, índex baix
torrent circulatori	torrent caudalós
sistema circulatori, sistema excretor, sistema immunològic, sistema linforeticular, sistema nerviós	sistema local, sistema automàtic, sistema elegant
regió iliocecal, regió renal, regió cervical, regió endèmica	regió muntanyosa, regió geogràfica, regió freda, regió petita

Aquests exemples ens serveixen per introduir el proper apartat en què valorarem la importància de l'adjectiu en les UTP.

4.2.2 Complements adjectivals pertinents en una UTP

> *Dependiendo de la naturaleza del núcleo, el determinante sirve para especificar con mayor detalle, indicar un fin, los medios mediante los cuales se lleva a cabo una operación, el objeto al que se aplica un proceso, o el tiempo, el lugar u otras circunstancias que llegan a convertirse, de este, en un rasgo distintivo integral del nuevo concepto.*
>
> [Sager, 1993: 122]

La naturalesa de l'adjectiu és molt complexa perquè, com és sabut, participa de dues categories gramaticals: el nom i el verb. Des d'un punt de vista morfològic, els adjectius tenen similituds amb els substantius, però, des d'una perspectiva sintàctica, s'assemblen als verbs. Si l'adjectiu

comparteix unes característiques amb els noms i unes altres amb els verbs és lògic preguntar-se la qüestió que es planteja Wierzbicka (1986) després de fer un repàs crític a diferents posicions teòriques:

What is, the "raison d'être" of adjectives as special word class?

Aquesta autora pensa que el que fa original l'adjectiu respecte del verb i del substantiu és l'atribució. Per a Wierzbicka, el fet que caracteritza l'adjectiu, des del punt de vista semàntic, és que afegeix un tret al substantiu que acompanya. Soler (1997: 61), en una tesi sobre la representació dels desajustaments [N + Adj], intenta justament establir els diversos valors amb què un adjectiu pot modificar un nom.

Com és sabut, existeixen criteris de molt diversa naturalesa per caracteritzar l'adjectiu, per bé que, des del punt de vista semàntic, els autors han tendit a diferenciar els adjectius entre adjectius relacionals i adjectius qualificatius. El fet que, tradicionalment, s'hagin caracteritzat els adjectius com a paraules que denoten qualitats o propietats de les entitats a les quals modifiquen ha induït Bosque [1993: 10] a dir que *"(...) con bastante frecuencia se olvida que no dejan de ser adjetivos aquellos que no representan nociones. Los adjetivos que se suelen llamar "relacionales", no "predicativos", "clasificatorios", "denominales" y "referenciales", entre otras denominaciones, se caracterizan precisamente porque no son calificativos, es decir no denotan cualidades o propiedades de los sustantivos, sino por el hecho de que establecen conexiones entre esas entidades y otros dominios o ámbitos externos a ellas con las cuales sitúan o clasifican a los sustantivos sobre los que inciden"*.

Aquesta dicotomia presentada en dos pols totalment oposats no dóna compte de la realitat, que és molt més ambigua per tal com hi ha adjectius que poden ser classificats com a qualificatius i com a relacionals:

> *Es evidente que no podemos interpretar de igual manera el adjetivo* musical *cuando aparece en SN como* sonido musical *y cuando aparece en* crítica musical. *En el primer caso,* musical *es un adjetivo calificativo, por lo que denota una cualidad o una propiedad del sonido, pero en el segundo es relacional o clasificativo, puesto que nos habla de una clase de crítica, es decir nos introduce en un dominio (el de la música) en relación con el cual hay que entender la crítica.*
>
> [Bosque, 1993: 11]

Per això, Bosque (1993) i Soler (1997), des d'una perspectiva estrictament semàntica, prefereixen parlar, no de tipus d'adjectius, sinó de valors (relacional o qualificatiu) que poden adquirir els adjectius segons el nom amb el qual es combinen i el context i àmbit temàtic en què apareixen:

> *Las diferencias entre los adjetivos calificativos y los adjetivos relacionales se manifiestan en la morfología, la sintaxis y el léxico e incluso una parte de ellas tiene su origen en nociones de naturaleza pragmática y existen casos de ambigüedad entre la interpretación calificativa y la interpretación relacional de un adjetivo.*
>
> [Bosque, 1993: 14]

Des del punt de vista terminològic, la funció de l'adjectiu en les UT amb estructura $[N[A]_{SAdj}]_{SN}$ és **sempre classificar el nom**, subespecificar-lo de la classe més genèrica a la qual pertany. Aquesta és una de les raons per la qual assumim la idea de Bosque que els adjectius poden presentar dos valors no excloents entre si. I dependrà del context (i del paradigma conceptual —afegim nosaltres—) perquè s'activi un valor o un altre.

En aquesta línia, Maingueneau [Maingueneau i Salvador, 1995] descriu aquesta dualitat mitjançant els conceptes encunyats per Milner de *classificatorietat* i *no-classificatorietat*:

> *Emprar un adjectiu de manera classificatòria (com que hi ha adjectius mixtos i contextos molt particulars, és preferible parlar d'usos que d'adjectius classificatoris) equival a fer entrar els referents en classes delimitables portadores d'informació. Usar un adjectiu de manera no classificatòria és, de fet, avaluar un objecte. El significat dels adjectius no permet, en aquest darrer cas, establir categories discretes (ço és, classificar un conjunt), sinó que es vincula essencialment a l'acte d'enunciació concret on s'inscriu.*
>
> [Maingueneau i Salvador, 1995: 132]

En les UT, l'ús activat és exclusivament el relacional perquè la funció de l'adjectiu en una UT és sempre classificar el substantiu al qual s'adjunta.

Els adjectius que *per defecte* són relacionals es diferencien de la resta d'adjectius per motius principalment semàntics, però també d'ordre morfosintàctic. Morfològicament, aquests adjectius solen ser derivats, formats a partir d'un nom i un sufix determinat (*-al*, *-ar*, *-ic*, *-isme*, etc.). Sintàcticament, no tenen naturalesa predicativa i, per tant:

- no poden funcionar com a atributs en frases copulatives
- no poden ser predicats en clàusules reduïdes
- no apareixen en posició prenominal
- no es poden coordinar amb adjectius qualificatius
- requereixen adjacència amb el substantiu.

I semànticament:

- no accepten una modificació de gradació[15]
- manifesten gairebé totes les relacions semàntiques que permeten els complements amb *de*.

Encara que, com assenyala Bosque (1993: 20), aquestes característiques dels adjectius relacionals per defecte són tendències significatives que tenen un valor estadístic més que sistemàtic, ja que en dominis especialitzats les recategoritzacions són molt freqüents.

Així, els adjectius, majoritàriament simples, que tendeixen a activar el valor qualificatiu, en tant que unitats especialitzades no apareixen mai aïlladament, però sí que poden formar part d'una UTP i és en aquest

moment que desactiven el valor qualificatiu adquirit per defecte i activen el valor relacional. Això significa que els adjectius qualificatius poden, en un domini concret, adquirir el valor d'adjectiu classificador:

> *Una expresión tal como* trocha ancha *utilizada por un ingeniero del ferrocarril se convierte en una expresión terminologizada; es un término de un lenguaje especializado mientras que en el lenguaje general no siempre se considera una forma lexicalizada.*
>
> [Sager, 1993: 96]

En aquesta mateixa línia Bosque (1993:20) cita un exemple del paper del sufix *-oso* (considerat tradicionalment qualificatiu) en el lèxic d'algunes matèries cientificotècniques:

> *Tampoco son muchos les adjetivos relacionales entre el numeroso grupo de adjetivos terminados en* -oso. *Entre las excepciones están los términos técnicos que se pueden usar como adjetivos relacionales, además de como adjetivos calificativos, en los casos en los que la presencia de la materia denotada por el sustantivo del que derivan puede ser distintiva de alguna clase natural, lo que les permite adquirir propiedades clasificatorias:* glanduloso, nervioso, arenoso, oleaginoso, *etc.*
>
> [Bosque, 1993: 20]

A més, cada àrea de coneixement compta amb un conjunt d'adjectius qualificatius que poden ser constituents d'una UTP. I, quan aquests adjectius formen part d'una UTP, es converteixen en adjectius classificadors i subespecifiquen el substantiu al qual modifiquen.

En medicina, per exemple, els adjectius que denoten el grau d'intensitat, el color o la forma, moltes vegades funcionen com a classificadors:

> *cefalàlgia moderada, cefalàlgia lleu, defalàlgia aguda, cefalàlgia intensa, cefalàlgia pulsativa, edema benigne, edema maligne,* etc.; *substància blanca, substància gris, taca negra, tifus groc, febre groga, glòbul blanc, glòbul vermell, màcula blava,* etc.; *cos estriat, cos esponjós, peu pla, peu buit, pronador quadrat, pronador rodó,* etc.

---

[15] Bosque nota que és lògic que els adjectius relacionals no acceptin aquesta propietat perquè no denoten qualitats, sinó classes o individus.

Un indici per reconèixer que un adjectiu està actuant d'adjectiu classificador és la presència en textos especialitzats d'adjectius en relació de contrast que modifiquen un mateix substantiu:

> *maxil·lar superior/maxil·lar inferior, intestí gros/intestí prim, paladar tou/paladar dur, jugular anterior/ jugular posterior, palatí anterior/palatí mitjà/palatí posterior, artèria faríngia ascendent/artèria faríngia descendent, ronyó esquerre/ronyó dret.*

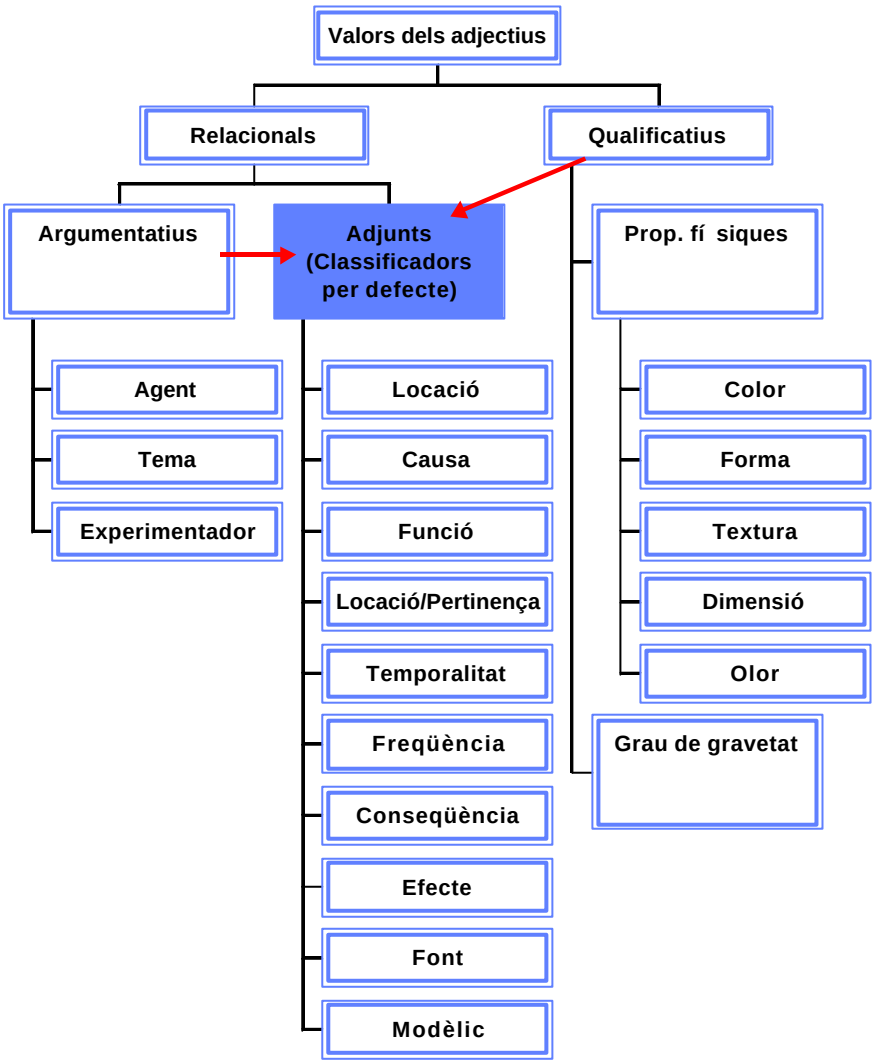
Ni *superior, inferior, gros, prim, tou, dur, anterior, mitjà, posterior, ascendent, descendent, esquerre* ni *dret,* són adjectius que, autònomament, siguin especialitzats ni classificadors; només quan es combinen amb un terme (que pertany a un paradigma conceptual), el subclassifiquen i converteixen la combinació $[N[A]_{SAdj}]_{SN}$ en un tipus de la classe a la qual pertany el seu nucli nominal.

Consegüentment, podem afirmar que no tots els adjectius que formen part d'una UTP poden funcionar aïlladament amb un significat especialitzat, dit d'una altra manera, no tots els constituents d'una UTP són USE autònomes. La característica més important perquè un adjectiu sigui constituent d'una UTP és que aquesta unitat formi part d'una tipologia, d'una classificació conceptual. Un dels trets bàsics dels adjectius relacionals és, doncs, el fet que creen una subclasse de la classe a la qual pertany el nom:

> *Parece acertado considerar que una de las características de los adjetivos relacionales, por lo que se refiere a su incidencia en el nombre, es que éste resulta subclasificado.*
>
> [Soler, 1997: 92]

En el cas que l'adjectiu d'una seqüència $[N[A]_{SAdj}]_{SN}$ no funcioni de classificador significa que aquesta seqüència no és terminològica, sinó discursiva o fraseològica.

A continuació hem elaborat una classificació dels adjectius del corpus que formen part d'una UTP amb estructura [N[A]$_{SAdj}$]$_{SN}$ a partir de les propostes de Soler (1997), Bosque (1993) i Lorente (1994):



Aquesta classificació ens permet prioritzar la informació específica que aporten aquests adjectius al nom, informació que serveix per discriminar semànticament aquest nom respecte de les unitats de la seva mateixa classe. A continuació, hem exemplificat les diverses classes d'adjectius que poden formar part d'una UTP amb els termes del corpus i, en alguns casos,

per tal d'il·lustrar amb més claredat les diferents classes, amb unitats d'altres textos i de diccionaris:

Adjectius relacionals classificadors:

- **Adjectius que indiquen la localització del "N" en el cos humà:**

  - components del cos humà + sufix relacional: endotelial; renal; cutani, -ània; cerebral; nasal, retrocervical; laringi; fetal; medul·lar; cardíac, -a; hepàtic, -a, endocrí, -ina, muscular; salival; etc.

    Exemples: *embòlia cerebral, embòlia pulmonar, fibrosi pulmunar, fístula cutània, trombosi cerebral, adenopatia retrocervical, sondatge esofàgic, úlcera anal*, etc.

  - no anatòmics: anterior; central; superior; bilateral; distal; dorsal; dret, -a; esquerre, -a; extern, -a; frontal; central; global; inferior; intern; lateral; local; dreta; transversal; medial; mitjà, -ana; parcial; posterior; proximal; supí, -ina; unilateral; primer, -a; segon, -a.

    Exemples: *bronquièxtasi central, extremitats inferiors, adenopatia regional, lòbul superior, pulmó dret, pulmó esquerre*, etc.

- **Adjectius que indiquen la localització/pertinença del "N"[16]:** alveolar; femoral; lingual; nasal; palatí; testicular; etc.

  Exemples: *artèria alveolar, artèria bucal, artèria testicular, bíceps femorals, cavitat nasal, nervi palatí, nervi toràcic, papil·la lingual, vena testicular*, etc.

- **Adjectius que indiquen la causa del "N":** al·lèrgic, -a; anèmic, -a; bacterià, -ana; reumàtic, -a; rickettsià, -ana; tífic, -a; víric, -a; etc.

[16] Els adjectius que indiquen la pertinença i els adjectius que indiquen la localització respecte del cos humà presenten característiques semàntiques molt semblants i, a vegades, es fa difícil diferenciar-los, és possible que entre uns i altres hi hagi una diferència genètica (tothom neix amb una cavitat nasal, però l'embòlia cerebral és una malaltia aquirida).

Exemples: *febre reumàtica, febre tífica, malaltia vírica, pneumònia vírica, verola rickketsiana, pneumònia bacteriana, taquicàrida anèmica,* etc.

- **Adjectius que indiquen la temporalitat o la freqüència del "N":** precoç; crònic, -a; constant; diürn; habitual; incessant; intensiu, -iva; intermitent; nocturn; permanent; periòdic, -a; persistent; primaveral; recent; remitent; únic, -a; tardà, -ana; terminal; transitori, -òria.

  Exemples: *conjuntivitis primaveral, cura intensiva, diagnòstic precoç, febre remitent, pneumonitis crònica, infecció recent,* etc.

- **Adjectius que indiquen la funció del "N":** antidepressiu, -iva; estimulant; patogen; respiratori, -òria; sensitiu, -iva; motor, -a, obturador, -a; gustatiu, -iva; òptic, -a; tàctil.

  Exemples: *sistema respiratori, nervi auditiu, neurona sensitiva, placa motora, nervi obturador, papil·la gustativa, psicofàrmac antidepressiu, psicofàrmac estimulant,* etc.

- **Adjectius que indiquen la font amb què funciona el "N":** magnètic, -a, neutrònic, -a; químic, -a; radiològic, -a, tèrmic, -a; ultrasònic, -a; etc.

  Exemples: *anàlisi química, anàlisi tèrmica, pelvimetria radiològica, pelvimetria ultrasònica, intura alcohòlica, radiografia electrònica,* etc.

- **Adjectius que indiquen la conseqüència que ocasiona el "N":** congestiu, -iva; dolorós, -osa; febril; irritatiu, -iva; etc.

  Exemples: *adiposi dolorosa, lesió irritativa, seborrea congestiva, taquicàrdia febril,* etc.

- **Adjectius modèlics** que indiquen la relació respecte d'un estàndard establert: normal; anormal; típic, -a; atípic, -a; atòpic, -a; estrany, -a; artificial; patològic, -a; etc.

  Exemples: *pneumònia atípica, dermatitis atòpica, pneumotòrax patològic, ronyó artificial,* etc.

Adjectius argumentatius que, en el domini de la medicina, adquireixen el valor de relacionals classificadors:

- **Adjectius que indiquen l'agent:** alcohòlic, -a; climàtic, -a; químic, -a; solar.

  Exemples: *neuritis alcohòlica, meningitis química, icterícia solar, nefritis solar,* etc.

- **Adjectius que indiquen el tema[17]:** adjectius derivats integrats per una base de caràcter terminològic que es combinen amb un nom deverbal[18].

  Exemples: *concentració bactericida, disseminació hematogènica, infecció cutània, infecció endotelial, lesió histopatològica, pressió venosa, tractament antimicrobià,* etc.

- **Adjectius que indiquen l'experimentador del "N":** infantil; professional; rural; senil; etc.

  Exemples: *demència juvenil, demència senil, paràlisi infantil, malaltia professional, malaltia rural[19]*, etc.

Adjectius qualificatius que, en el domini de la medicina, adquireixen el valor de relacionals classificadors:

- **Adjectius que indiquen el grau de gravetat del "N"**: advers; agut, -uda; ascendent; favorable; benigne, -a; descendent; maligne, -a; moderat, -ada; intens, -a; secundari, -ària; nociu, -iva; estable; primari, -ària; progressiu, -iva;lleuger, -a; lleu; greu; pulsatiu, -iva; superficial.

  Exemples: *anèmia moderada, cefalea intensa, infecció aguda, infecció secundària, sífilis primària, sífilis secundària, diarrea lleugera, enteritis greu*, etc.

- **Adjectius que indiquen la morfologia del "N":**

  - forma o composició:

[17] Això no vol dir que totes aquestes combinacions, en què el nucli és un substantiu deverbal i l'adjectiu té valor relacional, siguin adjectius temàtics; ja que, per exemple, també poden ser complements: *sobreinfecció bacteriana* (causa), *reacció al·lèrgica* (conseqüència), i en aquests casos tendeixen a tractar-se d'UT.

[18] Si aquestes seqüències són terminològiques, significa que l'adjectiu ha abandonat el valor argumental i ha prioritzat el valor classificador. Aquestes unitats són molt nombroses en els textos especialitzats.

[19] En aquests dos darrers exemples l'experimentador no és del tot explícit ja que l'experimentador no és el medi rural, sinó les persones que viuen en un medi rural; s'ha produït una sinècdoque.

    - arrel culta: gonio-, giro-, cilinfro-, cubo-, cif-, acanto-, -forme, etc (Bernabeu i al. (1995) en recullen 32).
    - arrel no culta: recte, -a, tacat, -ada, pla, -ana, els adjectius amb el sufix -ós, -osa o amb el sufix -il, (*artèria callosa, berruga plana, diarrea aquosa, eritema nodós, leucodistròfia esponjosa, màcula eritematosa, peu pla, teixit adipós, tifus febril*, etc.)
- dimensió:
    - arrel culta: braqui-, brevi-, pen-, hemi-, micro-, macro- etc. (Bernabeu i al. (1995) en recullen 25).
    - arrel no culta: mitjà, -ana; major; menor; curt, -a; llarg, -a; gros, -ossa; prim, -a; gegant, -a; (*articulació grossa, cèl·lula gegant, intestí prim, intestí gros*, etc.)
- color:
    - arrel culta: alb-, cian-, cromat-, polio-, rubr- (Bernabeu i al. (1995) en recullen 15).
    - arrel no culta: blau, -ava; groc, -oga; blanc, -a; negre, -a; gris, -a; vermell, -a;[20] *(escara negra, febre groga, màcula blava, substància grisa, glòbul vermell, tifus groc,* etc.*).*
- textura:
    - arrel culta: picn-, malac-, higr-, lepto-, xer- etc. (Bernabeu i al. (1995) en recullen 10).
    - arrel no culta: tou, -ova; dur, -a; fi, -ina; espès, -a; elàstic, -a; sec, -a; humit, -ida; transparent, -a (*pus espès, bena elàstica, paladar dur, paladar tou*, etc.).

Fins aquí, hem presentat una possible classificació dels adjectius que operen en les UT dels textos de medicina, per bé que som conscients que un estudi exhaustiu de la semàntica adjectiva hauria d'incloure la semàntica nominal per poder determinar amb més precisió en quin aspecte semàntic del nom incideix l'adjectiu. Així, per exemple les malalties tendeixen a subclassificar-se, per ordre de freqüència, per:

- el lloc del cos humà on es manifesta la malaltia (*hepatitis lúpica*)
- la causa o tipus de causa que l'ha provocat (*hepatitis alcohòlica, hepatitis vírica*)

[20] Els adjectius que indiquen color i que poden ser classificadors es redueixen bàsicament als colors següents: *blau, groc, vermell* (i la gamma dels morats), *blanc, negre, gris*.

- la morfologia més característica que presenta (*hepatitis groga, hepatitis vermella, hepatitis grisa*)
- la seva gravetat (*hepatitis aguda*)
- la freqüència en què es manifesta la malaltia (*hepatitis crònica*).

En canvi, els adjectius que acompanyen instruments utilitzats en la pràctica mèdica se solen subclassificar per la funció (és a dir per la utilitat per la qual es fan servir) o/i la morfologia (*sonda gàstrica, sonda uretral, sonda acanalada, sonda colzada*). Els components del cos, a partir de la localització i/o de la funció i/o de la morfologia (*orella interna, orella externa, vas sanguini, nervi olfactiu, nervi òptic, òrgan respiratori, paladar tou, paladar dur*).

La prioritització d'un punt de vista o d'un altre a l'hora de subespecificar una determinada classe de conceptes depèn més de factors extralingüístics que de criteris lingüístics:

> *Els criteris aplicats a l'establiment d'una classificació poden ser múltiples; una classificació de malalties, per exemple, pot obeir a criteris d'afinitat anatòmica, patogènica, etiològica, epidemiològica, terapèutica, pronòstica o, fins i tot, a criteris de costos. Ara, com que les classificacions elaborades per una escola que no és la pròpia tenen —sempre, inevitablement— un defecte o altre, les malalties —o les flors de la muntanya o les obres del pensament filosòfic o les begudes alcohòliques— també poden ésser ordenades —amb la intenció de fugir de controvèrsies— alfabèticament, de la A a la Z.*
>
> [Casassas i Ramis, 1995: 35]

Encara que en el cas de les malalties la gran majoria tendeixen a classificar-se per la localització i/o la causa que les ha originat (Bernabeu i al., 1995), són també factors extralingüístics (pragmàtics, culturals, personals, ideològics, perceptuals, etc.) els que permeten denominar un mateix concepte de diferents maneres: *fractura de Bennet* i *fractura dels boxadors,* per exemple, són ontològicament idèntics, en el primer cas, però, el terme ha prioritzat el descobridor d'aquesta fractura i, en el segon cas, el grup de persones que la solen patir. A vegades, aquesta diversitat de

denominacions d'una mateixa realitat respon a una visió de l'objecte més científica, més descriptiva: *tendó del calcani* i *tendó d'Aquiles*; *sarcoma de cutani helangiectàctic múltiple* i *sarcoma Kaposi*, etc. La variació terminològica (conceptual i formal), per bé que els especialistes tendeixen a evitar-la, és un fenomen habitual, espontani, inherent a qualsevol llengua natural.

A part de les característiques morfosintàctiques, Bosque (1993) proposa també intentar restringir les interpretacions possibles dels adjectius classificadors[21]. En aquest sentit, la introducció d'elements semàntics al costat dels morfosintàctics seria una informació útil perquè un SEACAT pogués extreure les UTP d'un text.

4.2.3 Nuclis nominals no terminològics: *organitzadors del discurs*

En el textos especialitzats, al costat de les UT, trobem també un tipus de noms que poden ser el nucli de segments amb estructura $[N[A]_{SAdj}]_{SN}$ que hem anomenat substantius organitzadors del discurs.

Si aquests noms van acompanyats d'un adjectiu especialitzat els hem anomenat *paratermes* (*problema asmàtic*), perquè són unitats que són al costat dels termes, amb la funció "d'esponjar cognitivament" el text i, a la vegada, proporcionar informació semàntica, pragmàtica i extralingüística

[21] *Aunque no creemos que la gramática necesite asumir los mecanismos de cálculo asociados a las conexiones de significado hiperespecíficas, sí creemos en cambio que el léxico debe marcar las que llamaremos interpretaciones semánticas exclusivas de ciertos adjetivos relacionales. Necesitamos prever que en SN como* curación manual *no va a significar "curación de las manos" (cf. en cambio* curación cutánea*) y también que un análisis del aire en un tubo de ensayo no es un* análisis aéreo *(cf. en cambio* análisis bacterial*). Estas son, paradójicamente, interpretaciones permitidas a priori en una aproximación pragmática, pero la gramática debe excluirlas adecuadamente. una forma de hacerlo es lograr que el léxico restrinja individualmente la interpretación de estos adjetivos relacionales, de forma que algunos adjetivos relacionales serán exclusiva o intrínsecamente "instrumentales" (manual), "locativos" (aéreo). Vistas así las cosas, la tarea del léxico es estos casos es la de restringir las interpretaciones posibles de un grupo de adjetivos relacionales.*

[Bosque, 1993: 34]

sobre les USE: en el cas de *problema asmàtic,* per posar un exemple, el paraterme *problema* informa que l'*asma* es percep com una dificultat que s'haurà d'intentar solucionar.

Aquests substantius ajuden, doncs, a fer progressar el discurs, però, des del punt de vista estructural, provoquen **molt de soroll** en qualsevol tipus de text especialitzat perquè tenen una estructura idèntica a la d'una UTP.

Alguns exemples d'organitzadors del discurs que hem documentat en el corpus amb estructura $[N[A]_{SAdj}]_{SN}$ són els següents:

> *Les **parts següents** del sistema nerviós estan implicades primariament en l'execució de moviments: (…)*
>
> *Un altre **factor fonamental** que determina l'aparició de formes greus és el retard en la instauració del tractament.*
>
> *Els **elements claus** de la immunitat són les defenses de l'hoste per les cèl·lules mediades per les cèl·lules T, incloent-hi l'interferó gamma.*
>
> *Durant els **dies següents**, les manifestacions reflecteixen l'afecció sistèmica.*
>
> *La formació d'aneurismes a l'arrel aòrtica s'ha descrit com un **fet característic.***
>
> *El **diagrama petit** de **l'angle inferior esquerre** il·lustra (…)*

4.2.4 Complements adjectivals no pertinents

Hi ha, com ja dèiem en l'apartat anterior, un conjunt d'adjectius que mai no són constituents i que acompanyen els noms que hem anomenat organitzadors del discurs:

> *apartat següent, capítol inicial, característica principal, categories següents, criteris següents, dies diferents, factor fonamental, focus inicial, setmanes anteriors, tret fonamental, tret principal,* etc.

Aquests adjectius, que solen modificar els noms que hem tractat en l'apartat anterior, no són adjectius amb poder subespecificador, classificador i, per tant, no poden constituir cap UT. Però ocasionen una gran quantitat de soroll.

També ocasionen soroll gairebé tots els participis i els gerundis que no provenen d'un verb especialitzat: *produït, determinat, derivat, acumulat, elegit, aproximat, adquirit, indicat, adequat, apropiat*[22], etc.; i, encara que el participi o el gerundi provingui d'una USE verbal, la combinació $[N[A]_{SAdj}]_{SN}$ no és terminològica sinó freqüent:

> *pacient afectat, organisme infectat, macròfag parasitat, dona embarassada, fons infiltrat, múscul adolorit, paràsit inoculat, sensori disminuït, positivitat encreuada, poliovirus atenuat, pus espès, mort sobtada, mortalitat elevada, febre elevada, febre moderada, micoplasma atenuat, vacuna inactiva,* etc.

[22] *base proporcionada, dosi ajustada, fàrmac elegit, fàrmac indicat, lloc freqüentat, manera indicada, motius desconeguts, moment determinat*, etc.

A vegades, però, per factors extralingüístics una unitat d'aquest tipus pot esdevenir terminològica, com és el cas de *febre tacada, aigua destil·lada* (que són les dues úniques que hem trobat en el corpus)[23].

Finalment, també hem observat que els adjectius derivats formats per combinació d'una arrel verbal i el sufix *-able*, en medicina, provoquen generalment soroll: *destacable, notable, possible, diferenciable, recomanable, susceptible*[24]. Tot i així n'hi ha alguns que combinats amb termes poden ser considerats una USE polilèxica, com per exemple *diagnòstic favorable, satura absorbible.*

4.2.5 El soroll de l'estructura [[N[A]$_{SAdj}$]$_{SN}$ [A]$_{SAdj}$]$_{SN}$

L'estructura [[N[A]$_{SAdj}$]$_{SN}$ [A]$_{SAdj}$]$_{SN}$ és la que causa menys soroll de totes les estructures del corpus susceptibles de ser una UTP: només un 12% del total d'unitats amb el patró [[N[A]$_{SAdj}$]$_{SN}$ [A]$_{SAdj}$]$_{SN}$ són segments no pertinents des del punt de vista especialitzat. La resta d'ocurrències es distribueix de la manera següent:

- el 53% de les ocurrències amb una estructura [[N[A]$_{SAdj}$][A]$_{SAdj}$]$_{SN}$ són UTP
- el segon adjectiu del 34% de les ocurrències amb una estructura [[N[A]$_{SAdj}$]$_{SN}$ [A]$_{SAdj}$]$_{SN}$ produeix soroll. D'aquest percentatge, el 30.5% de les ocurrències es redueixen a UT amb una estructura [N[A]$_{SAdj}$]$_{SN}$ i, per tant, s'ha de rebutjar sistemàticament l'últim adjectiu de la seqüència.

[23] Algunes vegades, aquest tipus de participis o de gerundis indiquen la freqüència de la malaltia o el procés fisiològic, la seva gravetat o la morfologia d'un aparell o d'una malaltia, però el fet que siguin combinacions fraseològiques o terminològiques només depèn de factos extrínsecs a la lingüística.

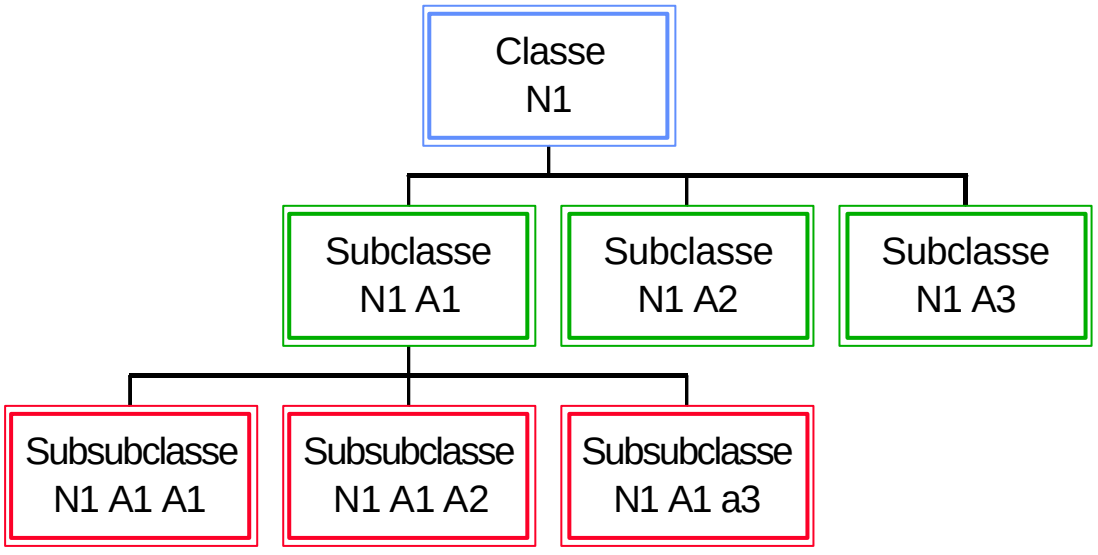
[24] *membre susceptible, població susceptible, mobilitat notable, risc probable, nombre variable, morfologia variable, mida considerable,* etc.

Si analitzem només les ocurrències amb estructura [[N[A]$_{SAdj}$]$_{SN}$ [A]$_{SAdj}$]$_{SN}$ que no són vàlides, podem dividir-les a partir dels elements que han provocat el soroll:

- tots els constituents del segment (12%)
- el nucli del segment (21%)
- el segon adjectiu (13%).

En aquests casos, hem comprovat que en aquesta estructura el primer adjectiu no és mai causa única del soroll.

Les unitats (tant pel que fa als noms com als adjectius) que originen soroll en l'estructura [[N[A]$_{SAdj}$]$_{SN}$ [A]$_{SAdj}$]$_{SN}$ són les mateixes que provoquen soroll en l'estructura [N[A]$_{SAdj}$]$_{SN}$, que ja hem comentat en l'apartat anterior. La diferència entre aquests dos patrons rau en el segon adjectiu de l'estructura [[N[A]$_{SAdj}$]$_{SN}$ [A]$_{SAdj}$]$_{SN}$, el qual té la funció de subespecificar una subclasse formada per [N[A]$_{SAdj}$]$_{SN}$. L'esquema següent mostra aquesta jerarquia que s'estableix a partir dels adjectius, els quals contribueixen a la construcció de sistemes conceptuals:



Una de les característiques dels adjectius classificadors és que requereixen adjacència estructural amb el substantiu que els regeix. Per tant, en el cas de l'estructura [[N[A]$_{SAdj}$]$_{SN}$ [A]$_{SAdj}$]$_{SN}$, el segon adjectiu ha de complementar no només el substantiu, sinó tot el sintagma nominal ([N[A]$_{SAdj}$]$_{SN}$) que el precedeix, perquè si només complementés el nom (i no el sintagma nominal format pel nom més l'adjectiu) deixaria de ser un adjectiu classificador:

> *Una de las propiedades fundamentales de los adjetivos relacionales es el simple hecho de que su interpretación semántica no es intrínseca, sino que depende de su relación posicional con el núcleo al que complementan. Los adjetivos relacionales significan, pues, por el lugar que ocupan.*
> [Bosque, 1993: 39]

Així, per tal que una unitat formada per un nom i dos adjectius esdevingui una UT, el dos adjectius han de tenir valor classificador. En les estructures [[N[A]$_{SAdj}$]$_{SN}$ [A]$_{SAdj}$]$_{SN}$ dels textos mèdics, el segon adjectiu es comporta com el de l'estructura [N[A]$_{SAdj}$]$_{SN}$ (4.2.2). Per exemple, en el terme *alveolitis al·lèrgica extrínseca* el primer adjectiu indica la causa d'aquesta malaltia i el segon adjectiu la zona afectada. En canvi, en la UT *aspergil·losi broncopulmonar al·lèrgica* ens trobem amb el cas contrari perquè els punts de vista de classificació s'han intercanviat: el primer adjectiu ens indica la localització de la malaltia i el segon la causa.

L'ordre dels adjectius classificadors depèn d'aspectes extralingüístics que propicien una prioritització d'una determinada propietat per sobre d'altres a l'hora de discriminar UT d'una mateixa classe:

> *gangli cervical superior* (localització i localització), *gangli cervical inferior* (localització i localització), *cambra ocular posterior* (localització i localització), *nervi auricular major* (localització i morfologia), *septe interalveolar dental* (localització i localització), *vàlvula aurículo-ventricular dreta* (localització i localització), *bacil*

fecal alcaligen (funció i morfologia), *òrgan reproductor femení* (funció i agent), etc.

Tot i així, hem observat algunes **tendències** en la formació de les paraules d'una mateixa classe: les patologies solen modificar-se primer per la localització de l'afectació i després per la causa que les ha produït, o viceversa. A vegades, les classificacions internacionals recomanen certes regularitats a l'hora de denominar un concepte, aquest és el cas de la *Nomina anatomica*, que de fet és una nomenclatura no natural:

> *Todos los términos de la Nomina anatomica son únicos para cada estructura anatómica, cortos y sencillos en lo posible e intentan no ser meras expresiones simbólicas, sino tener algún valor informativo o descriptivo. Por ello muchas partes anatómicas relacionadas desde el punto de vista topográfico llevan el mismo adjetivo y los demás adjetivos son pares de opuestos para permitir la formación de antónimos. Por supuesto, los epónimos están excluidos, como casi todas las nomenclaturas internacionales.*
>
> [López Piñero i Terrada Ferrandis, 1989: 75]

Fins aquí hem analitzat el soroll que provoca l'estructura sintagmàtica més productiva en el lèxic de les ciències de la salut ([N[A]$_{SAdj}$]$_{SN}$); en l'apartat següent estudiarem l'altre patró freqüent dels textos especialitzats: [N [SPrep]] $_{SN}$.

4.3 El soroll de l'estructura [N [SPrep]]$_{SN}$

L'estructura [N [SPrep]]$_{SN}$ representa aproximadament el 18% de les UTP dels textos mèdics. Aquesta estructura es materialitza prioritàriament en les cinc subestructures següents, que són les que tindrem en compte per estudiar el soroll:

- [N [de N]$_{SPrep}$]$_{SN}$
- [N [de N$_{propi}$]$_{SPrep}$]$_{SN}$

- [N [de art N]$_{SPrep}$]$_{SN}$
- [N [[de N]$_{SPrep}$ [A]$_{SAdj}$]$_{SPrep}$]$_{SN}$
- [N [[de art N]$_{SPrep}$ [A]$_{SAdj}$]$_{SPrep}$]$_{SN}$

L'estructura [N [SPrep]]$_{SN}$ és la que causa més soroll de les estructures que poden ser UTP quan s'aplica un SEACAT. Això vol dir que en els textos especialitzats hi ha molts segments no terminològics que corresponen a aquesta estructura; concretament, el 80% de les unitats amb aquesta estructura no són UT, cosa que no significa que no puguin ser especialitzades. Així, quan ens trobem amb un sintagma format per un substantiu i un sintagma preposicional introduït per la preposició *de*, pot tractar-se:

- d'una UT
- d'una UFE
- d'una combinació especialitzada recurrent
- d'una UL o UF
- d'una UD.

O bé pot passar que una part del segment sigui una UT i la part restant provoqui soroll. En aquest darrer cas, hem observat que el soroll, en més del 80% dels casos, està provocat pel nucli d'aquesta estructura i, en canvi, el sintagma nominal introduït per la preposició *de* sol ser una UT:

N de (art)	N (A)
soroll	UT

El soroll del primer nom pot estar motivat per la presència d'un:

- quantificador

- paraterme
- organitzador de l'estructura del discurs.

Tot i que també pot ocórrer que el nucli de l'estructura i el nucli del SPrep siguin terminològics, però la unitat resultant no; en aquests casos sol tractar-se d'una UFE o una combinació nominal recurrent: *biòpsia de pell* o *radiografia de tòrax*, en què tant *biòpsia* i *pell* com *radiografia* i *tòrax* són UT, però el resultat de combinar-se dóna lloc a col·locacions nominals.

4.3.1 Les col·locacions nominals

En els textos sobre ciències de la salut, trobem dos tipus diferents de col·locacions nominals que responen a l'estructura [N [SPrep]]$_{SN}$:

- unitats en què tant el nucli com el complement són USE, però l'estructura sencera no és una UT
- unitats en què el nucli és un substantiu deverbal, especialitzat o no, i el complement és una USE.

Al primer grup, hi pertanyen unitats com:

> *radiografia del tòrax, radiografia de la mà, radiografia del colze, radiografia de la pelvis,* etc.; o *diagnòstic de mononucleosi infecciosa, diagnòstic de toxoplasmosi aguda, diagnòstic de toxoplasmosi cerebral, diagnòstic de toxoplasmosi ocular, diagnòstic de tifus exantemàtic, diagnòstic de bronquitis crònica, diagnòstic de bronquitis infecciosa,* etc.

Són exemples del segon grup unitats com:

> *injecció d'heroïna, producció de limfocines, retenció d'orina, prevenció de la toxoplasmosi, introducció de la vacuna, instauració*

> *del tractament, tumefacció de la paròtide, aparició de brotades febrils, extravasació del líquid intravascular, infecció d'aneurismes aòrtics*, etc.

Observem la diferència entre els nuclis nominals del primer grup i els del segon: mentre que en els primers no hi ha indici d'un verb, en les unitats del segon el nucli és sempre un substantiu deverbal[25].

Per aquesta diferència, considerem que el primer tipus d'unitats (amb un nucli que no prové d'un verb) són **combinacions nominals especialitzades recurrents[26],** i el segon tipus (amb un nucli deverbal), **unitats fraseològiques especialitzades (UFE) nominals.** La característica principal de les combinacions nominals recurrents és només la freqüència d'ús en què es dóna una determinada combinació de paraules especialitzades, i la característica de les UFE nominals és, a més d'aquesta alta freqüència d'ús, la condició que el primer nom s'ha format a partir d'un verb[27].

---

[25] La fraseologia, etimològicament, està vinculada a la paraula *frase* [Cabré, Estopà, Lorente, 1996] i la *frase* —llegim en el *Diccionari de lingüística* (1992)— és "la unitat mínima de comunicació que relaciona un subjecte amb un predicat". Si apliquem aquesta condició a les seqüències anteriors, deduïm que només les unitats del segon grup estan en relació amb un predicat ja que el nucli d'aquestes unitats és un nom que prové d'un verb.

[26] Anomenades *col·locacions* per Sinclair (1991).

[27] Per exemple, les combinacions del tipus *radiografia de la pelvis* o *diagnòstic de bronquitis infecciosa* no són termes perquè no representen cap concepte d'una classificació mèdica. En aquests tipus de combinacions especialitzades el que és realment recurrent és el **patró semanticoformal**:

> *radiografia + de + art + una part del cos humà, diagnòstic + de (+ art) + nom d'una malaltia,* etc.

A un SEACAT que treballés amb informació semàntica (encara que fos amb etiquetes semàntiques preestablertes) li podria ser de gran utilitat comptar amb regles de combinació semàntica que indiquessin quins tipus de complements tenen valor classificador o combinatori amb un determinat substantiu:

- **diagnòstic + nom d'una malaltia** = una combinació freqüent: *diagnòstic d'artritis, diagnòstic de meningitis, diagnòstic de càncer de pròstata,* etc.
- **diagnòstic + adj. classificadors instrumentals que indiquen el mitjà que permet determinar un diagnòstic** = una UTP: *diagnòstic clínic, diagnòstic citològic, diagnòstic radiològic*, etc.

En alguns casos, la nominalització d'un verb derivat d'un substantiu amb significat especialitzat pertinent també és una UT: *infecció, injecció, infiltració,* etc.; aquestes unitats quan se subespecifiquen a través d'un adjectiu o d'un nom propi introduït per la preposició *de* esdevenen UTP. En el cas que siguin modificades per un sintagma nominal introduït per la preposició *de*, però, solen ser UFE.

4.3.2 Les unitats polilèxiques

En el corpus hem trobat molt poques unitats polilèxiques sense caràcter especialitzat que presentin l'estructura [N [SPrep]] $_{SN}$ (*blau de metilè, estació de l'any, fil de niló, rellotge de sorra*); representen el 0.016% del total de les ocurrències del corpus i el 0,3% dels segments amb aquesta estructura.

En principi, en els textos analitzats aquestes paraules no tenen valor especialitzat. Semànticament, a través dels contextos d'ús, només sabem que no estan relacionades directament amb cap unitat de l'esquema conceptual de les malalties infeccioses:

> *L'addició d'una tinció supravital, com el* ***blau de metilè*** *tamponat contrasta els detalls nuclears i redueix les possibilitats de confusió amb els leucòcits fecals.*
>
> *Els ous recuperats en els excrements són fàcilment confusibles amb els de S. haematobium, encara que a vegades s'hi poden reconèixer factors distintius, com la incubació de l'esperó terminal i la presència d'estrangulació en el miocardi que li confereix un aspecte de* ***rellotge de sorra****.*
>
> *Hi ha altres mètodes diagnòstics que consisteixen a demostrar la presència de trofozoï ts en mostres de contingut duodè mai obtingudes amb un* ***fil de niló*** *(enterotest) o recórrer a la biòpsia intestinal, si bé el rendiment d'aquests mètodes és escàs.*

Però aquest treball s'hauria de fer gairebé USE per USE o bé per classes d'USE i aquestes combinacions s'haurien de considerar tendències més que regularitats sistemàtiques.

En altres textos especialitzats, però, aquestes unitats poden tenir un ús especialitzat: per exemple, *blau de metilè* segur que és una UT en un catàleg de colorants o de pintures i *fil de niló* en textos sobre qüestions tèxtils. Tant *estació de l'any*, *fil de niló* com *rellotge de sorra* són, a més, mots molt divulgats i formen part del vocabulari general de la majoria de parlants.

4.3.3 Les unitats discursives nominals

Les unitats discursives (UD) apareixen en tot tipus de textos i, lògicament, també en el textos especialitzats. Ara bé, les UD nominals amb una estructura [N [SPrep]] $_{SN}$ ocasionen molt soroll quan s'aplica un SEACAT basat només en patrons morfosintàctics. D'entre les UD descrites per aquesta estructura morfosintàctica, constatem diferents tipus de nuclis:

a) **organitzadors de l'estructura del discurs**, com *inici del capítol, final del paràgraf*.
b) **quantificadors**, com *meitat dels malalts, majoria dels animals, resta de casos*, etc.
c) **paratermes**, com *característiques de l'exantema, causa de la coinfecció, procés d'immunitat, grup de zoonosis*, etc.

Una característica que comparteixen aquests tres tipus d'unitats és que gairebé sempre el complement està determinat per un article i, en canvi, sabem que la majoria dels complements preposicionals que integren les USE estan formats per un SN amb article indeterminat.

a) Els organitzadors de l'estructura del discurs amb un patró [N [SPrep]]$_{SN}$ (*a l'inici del capítol, al final del llibre, al llarg de l'article*, etc.) són molt poc

freqüents en els textos especialitzats (representen el 0,7% dels segments amb una estructura [N [SPrep]]$_{SN}$ del corpus analitzat), i, en canvi, són molt més freqüents amb l'estructura [N [SAdj]]$_{SN}$, com hem vist anteriorment[28].

b) Els sintagmes nominals amb una estructura [N [de SN]$_{SPrep}$]$_{SN}$ en què el nucli nominal té valor quantificador són molt abundants en els textos especialitzats i causen força soroll.

Presentem a continuació els quantificadors del corpus que responien a aquest patró, completats amb els que hem trobat en altres textos especialitzats utilitzats en l'etapa d'exploració[29] (hem marcat en negreta els més freqüents):

conjunt de	**majoria de**
meitat de	milers de
multitud de	**nombre de**
parell de	poc de
quantitat de	**resta de**
sèrie de	taxa de
terç de	total de
totalitat de.	

c) Els paratermes són unitats que acompanyen els termes i els fan de suport[30] ([Darbelnet,1979], [Lerat, 1995)] [Chetouami, 1997]). Segons aquests autors, totes les temàtiques científiques, tècniques o professionals

[28] Chetouami (1997) analitza el vocabulari que en els textos no forma part del conjunt d'UT i classifica aquests mots en organitzadors del discurs denotatius i funcionals, i, dintre d'aquests dos grups, distingeix entre els que tenen finalitats pedagògiques i els que donen informació sobre els termes.

[29] Textos sobre asma [del Hoyo, 1985], sobre epilèpsia [Oller i Ferrer,1979] i sobre neurologia [Adams i Victor,1984].

usen un *vocabulaire de soutien* determinat. Els paratermes es troben a l'interior d'un segment més llarg en el qual ocupen la posició del nucli sintàctic. En l'apartat 4.3, vèiem que aquest tipus de noms poden formar part de l'estructura [N [SAdj]]$_{SN}$, ara veiem que també poden formar part de l'estructura [N [SPrep]] $_{SN}$ en la qual, normalment, introdueixen una UT:

> *la causa del* ***xarampió***[31]*,* *l'episodi d'****infecció****,* *l'interior de la* ***cavitat nasal.***

Encara que, a vegades, també ocupen una part del complement:

> ***asma*** *de naturalesa* ***bronquial****,* ***citoplasma*** *d'aspecte* ***escumós.***

D'altres, el sintagma del qual formen part no té caràcter especialitzat:

> *factor de risc, anys d'evolució.*

Indirectament, aquestes unitats proporcionen informació sobre les UT del seu context i relacionen conceptualment les UT d'un text entre si.

A continuació, presentem els paratermes que hem trobat en els textos analitzats, agrupats pel tipus d'informació que transmeten a les UT que acompanyen:

Informació topogràfica:

exterior de	lloc de

[30] I moltes vegades expressen les relacions conceptuals entre les UT.

focus de	part de
interior de	zona de
localització de	àrea de

Informació sobre la freqüència:

absència de	predomini de
manca de	presència de

Informació metalingüística:

concepte de	nom de
denominació de	paraula de
idea de	terme de
noció de	

Informació temporal:

any de	setmana de
mes de	temps de
moment de	vida de
període de	

Informació sobre el paradigma conceptual:

[31] Hem marcat amb negreta les USE.

cas de	mena de
classe de	ordre de
espècie de	sistema de
gènere de	tipus de
grup de	

Informació relacional:

activitat de	característica de
causa de	conseqüència de
efecte de	essència de
estat de	factor de
origen de	procés de
raó de	responsable de
signe de	similitud de
símptoma de	

Informació morfològica:

forma de	objecte de
mida de	volum de

Informació sobre individus:

adult de	dona de
home de	nen de
infant de	persona de

Altres:

estudis de	escola de
fet de	ús de

Segments que s'interposen entre una UT:

N d'aspecte A	N de manera A
N de base A	N de naturalesa A
N de forma A	N de tipus A

4.3.5 El soroll de la subestructura [N [de $N_{propi}]_{SPrep}]_{SN}$

Al llarg de tot aquest apartat ens hem referit només als segments amb una estructura $[N_{comú}$ [de (art) $N_{comú}]_{SPrep}]_{SN}$, però, com ja hem apuntat diverses vegades durant el treball, en medicina proliferen els epònims, és a dir les UT formades per un nom, la preposició *de* i un nom propi antropònim o topònim.

Aquests noms propis indiquen o bé la persona que ha descobert una determinada malaltia, estat fisiològic, substància, aparell, operació, etc., o bé el nom d'un lloc (muntanya, poble, zona, país) on sol donar-se el N, d'on prové una determinada substància, on es va detectar un virus, etc.:

> *cèl·lules de Corti, òrgan de Corti, bastonets de Corti, pinces de Babcok, pinces de Hales, pinces de Hunter, mètode de Giménez, operació de Barraquer, malaltia de Hodgkin, febre de Malta, febre de les Muntanyes Rocalloses*, etc.

A vegades, el nom propi pot fer referència a una realitat més sociohistoricocultural: personatges mitològics, històrics o molt coneguts, i fins i tot, novel·lescos, tal com reflecteixen els exemples següents:

> *complex d'Edip* (Edip era el rei llegendari de Tebes que va matar al seu pare i es va casar amb la seva mare; és protagonista de diverses tragèdies clàssiques gregues).
>
> *síndrome de Pickwick* (Samuel Pickwick era el personatge principal de *The Posthumous Papers of the Picwick Club*, novel·la de Charles Dickens).
>
> *orella de Mozart* (el compositor de música W. Amadeus Mozart presentava una anomalia congènita consistent en la fissió de les branques de l'hèlix i l'antehèlix).

El nom propi té la funció de subespecificar el nom comú al qual acompanya, de classificar-lo alhora que dóna una informació sociohistòrica del nom que el regeix. En aquest sentit, el nom propi estableix una relació semblant a la dels adjectius classificadors de "pertinença" o de "possessió" respecte del nom que modifica, malgrat que no dóna cap informació intrínseca sobre aquest nom, sinó només informació extrínseca.

En els textos especialitzats de medicina, un segment que presenti l'estructura següent:

nom terminològic + preposició *de* + nom propi

pot tractar-se:

- d'una UT (aproximadament el 80% de les ocurrències que en un text especialitzat presenten aquesta estructura), o
- d'una UD (aproximadament el 20% de les seqüències que en un text especialitzat equivalen a aquesta estructura).

En el cas dels epònims que corresponen a UT, el nucli de la unitat és una UT, encara que conceptualment pugui ser molt genèrica (*malaltia, síndrome, trastorn, símptoma, operació, tractament)* i a vegades polisèmica (*tècnica, prova, mètode, signe, índex*, etc.).

En el text analitzat hem trobat 85 termes formats amb aquesta estructura: 78 en què el nom propi és un antropònim i 7 en què el nom propi és un topònim:

> *tinció de Giemsa, tinció de Gram, tinció de Wright, mètode de Giménez, malaltia Brill-Zinsser, malaltia de Chagas, cèl·lules de Kupffer, tècnica de Warthin-Starry, sarcoma de Kaposi, reacció de Wassermann, prova de Coombs, prova de Paul-Bunnell, fenomen de Raynaud, cultiu de Foley, síndrome de Katayama, síndrome de Reye, síndrome de Duncan, signe de Theodor, signe de Koplik, taca de Koplik, limfoma de Burkitt*, etc.

> *febre de les Muntanyes Rocalloses, tifus de la paparra de Kenya, tifus de la paparra de Kenya, febre tacada d'Israel, febre d'Astrakhan*, etc.

Apareixen també altres segments amb estructura [N [de $N_{propi}]_{SPrep}]_{SN}$ que no són UT i en què el primer nom tampoc no és terminològic. Així, hi ha un grup de mots que són locatius o ubicatius seguits d'un nom propi que sistemàticament causen soroll (*illa, estat, país, continent, ciutat, llac, muntanya, república, centre, escola, àrea, zona, etc.*; *nord, sud, oest, est*).

En el text analitzat, hem recollit els següents:

> *ciutat de Mèxic, estat de Rondònia, escola de Marsella, illa de Nantucket, nord de Tailàndia, república de Myanmar, sud de Mèxic, sud de Moçambic.*

I també ocasionen soroll mots com *cas, malalt, diagnòstic* seguit d'un nom propi:

> *cas d'Alzheimer, malalt de Parkinson, diagnòstic de Down*, etc.

El problema més difícil de discriminar és el de casos en què es pot confondre un genèric amb la malaltia d'un individu particular:

> *Gastaut, en un treball sobre l'**epilèpsia de Dostoievski** argumenta que aquest escriptor no va demenciar tot i haver sofert més d'un miler de crisis.*

L'usuari llec (com l'ordinador) no sap distingir si aquest segment —*epilèpsia de Dostoievski*— és un una UT, és a dir un tipus d'epilèpsia, o bé és una unitat discursiva, l'epilèpsia que pateix un individu particular. Tot i que aquests casos no siguin gaire habituals en els textos molt especialitzats, documents professionals i textos de caràcter més divulgatiu. Així, segons el tipus de corpus al qual apliquem un SEACAT haurem de tenir en compte aquesta variable que podria provocar més soroll de l'habitual.

4.3.6 Estructures més complexes que es redueixen

Un fenomen que s'ha de tenir controlat quan es dissenya un SEACAT és el fet que moltes unitats que aquests sistemes solen generar com a candidates a terme són unitats massa complexes que l'usuari ha de reduir.

Recordem que les UTP no solen tenir més de dos complements i que l'habitual és que només en tinguin un; per tant, davant d'estructures de dos o més de dos complements, és molt probable que una part del segment sigui una USE i la resta provoqui soroll. Per això, les estructures amb un complement complex o amb més de dos complements s'han de reduir a estructures més senzilles.

En efecte, només el 0,5% de les ocurrències del corpus amb estructura [N [[de N]$_{SPrep}$ [A]$_{SAdj}$]$_{SPrep}$] $_{SN}$ i l'1% amb estructura [N [[de art N]$_{SPrep}$ [A]$_{SAdj}$]$_{SPrep}$]$_{SN}$ són UT.

Vegem-ne alguns exemples:

> *infecció d'**aneurismes aòrtics,** forma de **conjuntivitis purulenta,** obstrucció de l'**intestí prim,** tipus de la **dermatitis exfoliativa**,* etc.
>
> *casos de **tripanosomiasi gambiana, components de la medul·la òssia,** existència de la **taca negra,** setmanes de la **premalaltia neurològica,** interior dels **fagolisomes cel·lulars,** mida dels **ventricles cerebrals**,* etc.
>
> ***zoonosi** de distribució mundial, **grup de risc** principal,* etc.
>
> ***equistosomiasi** del **sistema nerviós**, **histologia** dels **ganglis limfàtics**, **tromboflebitis** de les **extremitats inferiors**, **cèl·lules** de l'**endoteli vascular**, **cèl·lules** de l'**epiteli intestinal**, **cèl·lules** de la **paret intestinal**, **cèl·lules** dels **sediment orinari**,* etc.

> *àrees del sud-est asiàtic, zones de bosc primari, persones de raça blanca, percentatge de resultats positius, canvi de color observable*, etc.

El soroll d'aquestes unitats segueix les mateixes pautes el generat per l'estructura [N [SPrep]]$_{SN}$. El significatiu és el fet que davant d'un d'aquests dos patrons podem trobar-nos en tres situacions diferents:

- una primera possibilitat en què el segment sigui una combinació recurrent nominal
- una segona opció en què la seqüència sigui una UFE nominal
- una tercera situació en què només una part del segment (el complement preposicional format per un sintagma [N [SAdj]]$_{SN}$) sigui una unitat pertinent des del punt de vista terminològic.

Hem constatat que en gairebé cap cas trobem una UT amb una d'aquestes dues estructures:

> [N [[de N]$_{SPrep}$ [A]$_{SAdj}$]$_{SPrep}$] $_{SN}$
> [N [[de art N]$_{SPrep}$ [A]$_{SAdj}$]$_{SPrep}$] $_{SN.}$

De fet, en tot el corpus només n'hem trobat dues:

> *síndrome d'esplenomegàlia tropical, síndrome de coagulació intravascular.*

4.4 Conclusions

La conclusió principal que resulta de l'estudi del soroll —una de les limitacions més importants dels SEACAT— és que ni els esquemes

morfosintàctics ni la freqüència d'ús són suficients per discriminar les UTP dels textos d'especialitat. Al llarg d'aquest capítol, creiem haver demostrat que els patrons estructurals són un filtre massa permissiu per identificar les UT d'un domini determinat. Si s'usen patrons que fan referència només a la forma de les UT, hem comprovat com els SEACAT proposen com a candidats a terme tant unitats erròniament delimitades com segments sense interès terminològic. A més, es presenten les UT, les UFE i les combinacions recurrents sense cap mena de distinció, pel fet que les seves estructures morfosintàctiques poden coincidir:

> *Sur le plan de la reconnaissance des unités complexes, les principaux problèmes proviennent du fait que l'analyse purement syntaxique ne fait que très rarement appel à la sémantique et, bien souvent, de manière très partielle. Une analyse de la structure de surface ne permet pas de distinguer le terme du syntagme de discours lorsqu'ils possèdent une structure syntaxique identique comme* panier d'osier *(terme) et* panier de pommes *(syntagme de discours).*
>
> [Drouin, 1997:46]

Hem vist que en els textos d'especialitat hi ha diversos tipus d'unitats que presenten una estructura idèntica a les UTP i així hem comprovat que les estructures $[N\ [SAdj]]_{SN}$ i $[N\ [SPrep]]_{SN}$ poden correspondre a:

- una UT
- una UFE
- una combinació especialitzada recurrent
- un paraterme
- una unitat lèxica de caràcter noterminològic en el text
- una unitat fraseològica no especialitzada
- una UD.

També hem constatat que, per als nostres objectius, és interessant de poder recollir dels textos especialitzats els tres primers tipus d'unitats, perquè els tres són unitats de significació especialitzada (USE).

Consegüentment, hem demostrat que la causa del soroll —que s'estima entre el 40% i el 75%— rau bàsicament en el fet que les estructures de les UT no són exclusives d'aquest tipus d'unitats.

Partint, doncs, de la diversitat d'unitats que poden presentar una estructura idèntica a una UT, hem buscat altres recursos que permetin a un SEACAT reduir el soroll. Per fer-ho, hem centrat el nostre estudi en les dues estructures més freqüents de les UTP —[N [SAdj]]$_{SN}$ i[N [SPrep]] $_{SN}$— i hem explorat per separat les característiques dels seus nuclis i dels seus complements. D'aquestes anàlisis hem extret diverses conclusions que hem presentat al final de cada apartat i de les quals volen destacar les següents:

a. Quant a l'estructura [N [SAdj]]$_{SN}$:

a.1. Hem exposat que podien ser nucli d'una UTP amb estructura [N [SAdj]]$_{SN}$:

- un terme
- un nom deverbal, terminològic o no
- un nom no terminològic.

a.2. I hem vist que podien ser complements d'una UTP:

- una USE adjectiva
- un adjectiu de caràcter no especialitzat.

a.3. De les combinacions que es podien establir entre aquests tipus de noms i d'adjectius, en resulten les unitats següents:

- Si un terme es combina amb un adjectiu de caràcter especialitzat dóna lloc a una UT
- Si un terme es combina amb un adjectiu qualificatiu pot donar lloc a una UD o a una UT, segons el valor de l'adjectiu dins el domini conceptual de la medicina
- Si un nom deverbal, terminològic o no, es combina amb un adjectiu de caràcter especialitzat pot donar lloc a una UT
- Si un nom deverbal terminològic es combina amb un adjectiu de caràcter no especialitzat dóna lloc a una UD o a una UT, segons el valor de l'adjectiu en el domini mèdic
- Si un nom deverbal no terminològic es combina amb un adjectiu de caràcter no especialitzat dóna lloc a una UD
- Si un nom no terminològic es combina amb un adjectiu de caràcter especialitzat pot donar lloc a una UT, una UD, una combinació recurrent
- Si un nom no terminològic es combina amb un adjectiu de caràcter no especialitzat dóna lloc a una UD.

a.4. També hem comprovat que l'adjectiu que integra una UT sempre té la funció de *classificar* el nom que acompanya, i hem analitzat els diferents tipus d'informacions que l'adjectiu aporta al nom.

b. Quant a l'estructura [N [SPrep]] $_{SN}$:

b.1. Hem arribat a la conclusió que els segments més freqüents amb aquesta combinació són col·locacions ja sigui

amb un nucli deverbal (UFE) o estrictament nominal (combinacions especialitzades recurrents).

b.2. Hem observat que la majoria de paratermes i col·locacions nominals dels textos especialitzats de medicina es descriuen per l'estructura [N [[de art N]$_{SPrep}$ [A]$_{SAdj}$]$_{SPrep}$]$_{SN}$.

c. Hem observat també que els noms i els adjectius que poden formar part d'un paraterme no són els mateixos per a tots els tipus de dominis especialitzats.

d. Finalment, hem comprovat que tant en l'estructura [N [SAdj]]$_{SN}$ com en la [N [SPrep]]$_{SN}$ hi ha segments, que hem anomenat *organitzadors del discurs*, que provoquen sistemàticament soroll.

Les reflexions plantejades durant tot aquest capítol i l'anterior susciten dues qüestions polèmiques que són l'embrió dels dos propers capítols:

a) Una primera qüestió parteix del fet que entre els candidats a terme que un SEACAT proporciona hi ha unes unitats que són clarament UT, d'altres que clarament provoquen soroll, però n'hi ha unes altres que plantegen molts dubtes, fins i tot als propis especialistes, que no són del tot sistemàtics en els seus criteris de buidatge. En l'anàlisi dels buidatges dels corpus vam constatar unitats amb una estructura idèntica a una UTP no especialitzades: UL, les UF i les UD; i unitats especialitzades per bé que no totes són UT *strictu sensu:* les UFE (*complicació extrapulmonar, augment de la permeabilitat vascular, manifestacions al·lèrgiques, afecció cardíaca*, etc.), les combinacions especialitzades recurrents (*radiografia del braç, diagnòstic de*

toxoplasmosi cerebral, tractament de la febre groga, etc.) són també unitats de significació especialitzada[32].

Aquesta constatació ens porta a qüestionar-nos sobre les unitats que han de ser objecte de detecció per part d'un SEACAT (qüestió que tractarem en el capítol següent), i a plantejar-nos, a més, si el soroll és discriminatori, dit altrament ens demanem si el conjunt de segments que ocasionen soroll és el mateix per a tots els col·lectius d'usuaris de terminologia. Aquestes són les qüestions que intentarem resoldre en els propers capítols.

b) La segona qüestió posa en dubte si els SEACAT haurien d'apropar-se més a la noció d'USE i d'UT, la qual es pot definir com l'associació d'una forma i d'un significat especialitzat, i no només d'una forma. Tant els recursos que els SEACAT utilitzen (patrons morfosintàctics, càlculs estadístics, etc.) com els que hem proposat al llarg d'aquest capítol i de l'anterior (diccionari de formants, regles de combinació, xarxes paradigmàtiques, atenció a la tipografia, vocabulari de suport, etc.) són molt útils per extreure les USE d'un text, però no són suficients. Pensem que aquests recursos basats en els aspectes formals dels termes haurien de complementar-se amb els seus aspectes semàntics. Qualsevol informació semàntica relativa a les USE suposa una millora considerable en el filtratge de les seqüències extretes a partir d'esquemes morfosintàctics. Pensem, doncs, que la incorporació de la semàntica i de les relacions entre la sintaxi i la semàntica són, com veurem en el capítol següent, totalment necessàries per detectar i delimitar les USE.

[32] I fins i tot els paratermes (*causa d'artritis, factor de mortalitat, problema asmàtic*, etc.) poden tenir un interès especialitzat especial.

4.5 Recapitulació

En aquest i l'anterior capítol, hem estudiat detalladament les limitacions dels SEACAT que funcionen amb patrons morfosintàctics, limitacions manifestes en la seva aplicació. Hem comparat els resultats del buidatge automàtic amb els resultats del buidatge manual d'un mateix corpus textual i hem establert les diferències. Les dades que hem analitzat confirmen el fet que en els textos d'especialitat hi ha quatre conjunts diferents d'unitats:

- un conjunt d'unitats que, en un text, són unitats lèxiques o fraseològiques de caràcter no especialitzat, i per tant ocasionen soroll
- un conjunt d'unitats que són clarament discursives, i per tant també ocasionen soroll
- un conjunt d'unitats referencials que són clarament UT
- un conjunt d'unitats que poden ser UT, combinacions especialitzades recurrents, UFE, paratermes i, fins i tot, UD.

Els dos primers blocs d'unitats ocasionen soroll. El tercer grup d'unitats no genera soroll, però pot generar silenci. Finalment, el quart conjunt d'unitats amaga una gran diversitat de possibilitats.

A partir d'aquestes observacions, en el proper capítol, ens qüestionem quines són les unitats pertinents en ciències de la salut que un extractor hauria de detectar i, consegüentment, quines són les condicions amb les quals aquest sistema hauria de comptar per detectar les unitats pertinents i eliminar les que no ho són.

5. PROPOSTA DE MILLORA D'UN SEACAT CLÀSSIC: EL SEACUSE

> Localizar y extraer informaciones de masas cada vez más enormes de publicaciones, informes, historias clínicas, documentos sanitarios, etc. resulta escasamente eficaz si no se controlan fenómenos semánticos del lenguaje médico (...) Esta deficiencia se refleja en los dos indicadores básicos de un sistema de recuperación de información: tasas muy bajas de precisión y de exhaustividad.
>
> [López Piñero i Terrada Ferrandis, 1990: 95]

L'objectiu d'aquest capítol és **proposar elements i estratègies,** bàsicament de naturalesa lingüística, perquè un SEACAT augmenti el grau de precisió i d'exhaustivitat en el reconeixement de les unitats de significació especialitzada (USE). Dit d'una manera més simple, a **partir de l'anàlisi** de les mancances quantitativament i qualitativament significatives dels SEACAT clàssics, ens plantegem quins haurien de ser els components amb què hauria de comptar un SEACAT per tal que fos més eficient, més precís i més exhaustiu que els **sistemes actuals.**

En els tres capítols anteriors, hem comparat els buidatges manuals i automàtics d'un mateix corpus textual i hem arribat a la conclusió que els extractors automàtics generen uns llistats d'unitats que, respecte del buidatge fet per especialistes, són incomplets i provoquen molt soroll. A continuació, hem analitzat les característiques de les unitats que provoquen els baixos índexs de precisió i exhaustivitat d'aquests sistemes. A mesura que estudiàvem aquestes mancances, hem començat a trobar elements que ajudarien a millorar el funcionament dels SEACAT, elements que en aquest capítol presentem en forma de proposta.

5.1 Objecte d'extracció: les USE en les ciències de la salut

Abans d'abordar els elements i les estratègies de reconeixement, cal que definim l'objecte d'extracció d'un SEACAT, és a dir cal que especifiquem quines unitats dels textos mèdics volem que un sistema automàtic recuperi, perquè, com ja hem dit els SEACAT tradicionals s'han limitat a extreure, dels textos especialitzats, només les **unitats terminològiques polilèxiques (UTP).** Aquesta restricció ha implicat reduir el reconeixement automàtic a un tipus de paraules: les unitats terminològiques (UT)[1] i, dins d'aquesta categoria, a un únic tipus d'estructura: les unitats sintagmàtiques. En aquest treball, però, defensem la idea que, perquè els sistemes d'extracció automàtica siguin més útils, han de detectar una diversitat més gran d'unitats, que encara que no siguin pròpiament UT, tenen un interès especialitzat.

Des dels seus orígens la terminologia ha considerat que el terme —entès com una unitat lèxica nominal del llenguatge natural— era la seva unitat de base. En l'última dècada, però, alguns canvis d'orientació en la terminologia teòrica han permès d'eixamplar l'interès per les unitats de significació especialitzada d'un text més enllà de les UT.

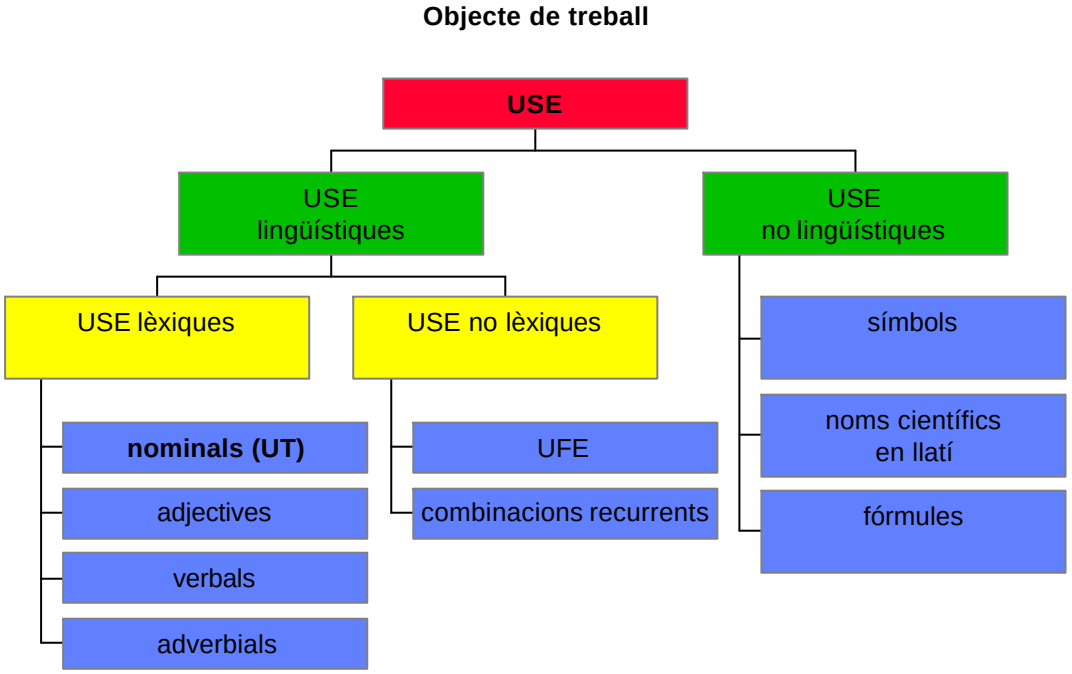
En aquesta línia, els resultats del nostre treball de recerca de 1996 sobre les unitats que formen part dels diccionaris de medicina, juntament amb l'estudi de la selecció de les UT pertinents d'un text feta per especialistes en el capítol segon d'aquest treball, reforcen la idea que la UT no és l'única unitat de significació especialitzada pertinent dels textos especialitzats.

D'acord amb aquesta nova perspectiva, la unitat que és objecte d'estudi de l'extractor que proposem no pot reduir-se ni a la UTP ni a la UT en general, sinó que ha d'abastar totes les unitats que anomenem ***unitats de significació especialitzada*** (USE), que inclouen tant les unitats especialitzades de categories gramaticals diferents que formen part del llenguatge natural, com les unitats que formen part de llenguatges artificials; i dins de les unitats que són llenguatge natural, abraça des de les

[1] Recordem que hem definit les UT com a USE nominals referencials.

UT simples a les complexes, des dels noms als verbs, adjectius i adverbis, des de les unitats lèxiques a les unitats fraseològiques especialitzades (UFE); i, finalment, dins de les unitats de sistemes artificials, comprèn tant els símbols nominals i els noms llatins propis de nomenclatures consensuades com les fórmules complexes[2] [3].

A continuació reproduïm l'esquema de les USE que són l'objecte de l'extractor que proposem, que ja hem presentat en la introducció:



Aquest canvi d'objecte d'extracció ens fa redifinir els sistemes d'extracció automàtica a candidats a terme (SEACAT) fins ara presentats i orientar l'objectiu del nostre treball en la recerca i selecció dels elements i estratègies

[2] En aquest treball i seguint el buidatge que del corpus han fet els especialistes, les USE iconogràfiques i les UFE adverbials i adjectives no seran objecte d'estudi.

[3] Consegüentment, les unitats que dels textos especialitzats hauria de recuperar un extractor pertanyen a dos tipus diferents: a) USE de naturalesa lingüística (*lent, medul·la, maternitat, esporació, medul·la òssia, comptagotes, lentímetre, cardiomegàlia, medul·litis, augment de la pressió arterial, TAC, SNC, p.c.*); b) USE de naturalesa no lingüística o artificial ($C_7H_6F_3NO_2$, C, g, PG_1, B_{12}, *Galium verum, Ballota album*. Cal tenir present, de totes maneres, que la representativitat de les USE en els textos especialitzats és desequilibrada, per tal com hi trobem moltes més unitats lingüístiques que unitats no lingüístiques.

que requereix un **sistema d'extracció automàtica de candidats a unitats de significació especialitzada** (SEACUSE).

Per delimitar les USE pertinents que un extractor hauria de poder detectar, ens hem basat en els buidatges fets per diversos especialistes sobre el corpus textual de medicina, els resultats dels quals hem presentat en els capítols 2, 3 i 4. El punt de vista del professional en medicina ha estat reforçat per l'anàlisi dels tipus d'unitats que figuren en els diccionaris especialitzats d'aquest domini.

5.1.1 USE de naturalesa lingüística

La classificació de les unitats extretes dels buidatges dels textos especialitzats ens permet diferenciar entre USE lingüístiques i no lingüístiques. Dintre de les lingüístiques, distingir entre lèxiques i sintàctiques. Les USE lèxiques poden ser, des del punt de vista de la seva composició, monolèxiques (*mà, hipersensibilitat, nervi, pleural, serològicament, pedicoquirometria, vacunar*) i polilèxiques (*malaltia de Hanot, malaltia vírica, nervi alveolar inferior, nervi auditiu, reacció en cadena de la polimerasa*).

Des del punt de vista de la categoria gramatical, les USE monolèxiques poden ser noms (*peu, abaixallengües, quimioteràpia, patognòmia*), verbs (*injectar, vacunar, desparasitar, desinfectar, infectar, hidratar*), adjectius (*mononèfric, -a, cutani, -ània, nasal, alveolar, queloïdal*) o adverbis (*clínicament, immunològicament, radiològicament, biològicament*). Les USE polilèxiques, en canvi —segons el criteri dels especialistes—, sempre són noms (*diagnòstic clínic, insuficiència cardíaca greu, malaltia de Brill-Zinsser, reacció de fixació de complement*). Així, si un extractor es basa en el punt de vista de l'especialista, no ha de tenir en compte les USE verbals, adjectivals ni adverbials que podrien trobar-se en els textos, mentre que sí que ha de tenir

en compte les unitats monolèxiques nominals, adjectives, verbals i adverbials, i les polilèxiques nominals.

Morfològicament, les USE monolèxiques nominals pertinents en medicina poden ser unitats simples (*febre, mà, nasal, os, peu*), unitats derivades (*ossi, ossada, nasal, enfebrar-se, febril*), compostos patrimonials (*abaixallengües, comptagotes*) o compostos cultes (*tifus, argiròfil, leucocitoblast, pericardi, nefritis, cirrosi, anèmia*). En contrast, les USE monolèxiques verbals són sobretot derivades (*desinfectar, cicatritzar, reepitelitzar, vacunar*) i les USE monolèxiques adverbials (*clínicament, mèdicament, immunològicament*) o adjectivals (*al·lèrgic, -a, atròfic, -a, clínic, -a, medul·lar, ossi, tiroïdal*) sempre són unitats formades per derivació.

Pel que fa a la semàntica, les USE de l'àmbit mèdic formen part de diverses classes conceptuals: malalties, estats patològics, manifestacions biològiques de malalties; parts, substàncies i components del cos humà; plantes medicinals, plantes comestibles i plantes verinoses; bactèries, enzims i protozous; animals comestibles, animals verinosos, animals paràsits i animals portadors; elements, compostos químics i fàrmacs; processos patològics; accions i operacions; efectes d'accions; i propietats.

Al costat de les unitats monolèxiques i de les polilèxiques, i a cavall entre les unes i les altres, trobem un tipus d'unitats que també és pertinent de recollir en medicina: les sigles (*ADN, LTH, LTT, MAO, SIDA*). Des del punt de vista morfosintàctic, són unitats simples que procedeixen d'una concatenació d'unitats, i que funcionalment es comporten com els substantius.

I frontereres entre les UTP i les unitats discursives (UD), trobem també en els textos mèdics les unitats fraseològiques especialitzades (UFE) i les combinacions recurrents especialitzades nominals, que també pot ser

pertinent de recollir-les (*tractament de l'hepatitis, radiografia del peu esquerre, augment de la permeabilitat vascular*)[4].

Des del punt de vista lingüístic, doncs, un SEACUSE que vulgui assolir un nivell d'eficiència raonable i adequar-se a allò que esperen els especialistes, hauria de reconèixer totes les UT, les unitats verbals, adjectivals, adverbials de caràcter especialitzat, les sigles, les UFE nominals i les combinacions recurrents nominals especialitzades.

5.1.2 USE de naturalesa no lingüística

A més de les USE lingüístiques, un extractor també hauria de poder reconèixer les USE no lingüístiques que es troben en els textos especialitzats. Partint sempre de l'enfocament de l'especialista en medicina, les USE que pertanyen a nomenclatures artificials es redueixen als noms científics en llatí, símbols i fórmules, i més concretament als següents:

- els noms llatins de les parts del cos que formen part de la *Nomina anatomica* (*arteria femoralis, vena femoralis, lobus anterior, lobus inferior*).
- els noms llatins dels zoònims (*Rattus rattus, Diphyllobothrium latum*)
- els noms llatins de les plantes, tant de les comestibles com de les verinoses i les medicinals (*Melissa officinalis, Artemisia herba-alba*)
- els noms llatins de les bactèries (*Mycobacterium tuberculosis, Rickettsia australis, Rickettsia conorii*)
- els símbols dels elements o compostos químics, orgànics i inorgànics (*Ra, Na, F, B_1, B_2, B_3*)
- els símbols del Sistema Internacional d'Unitats (*L, J, s, A*)
- les fórmules dels compostos químics (CH_2O, H_2O, $C_8H_{10}N_2S$).

[4] De fet, els sistemes clàssics, encara que involuntàriament, ja solen detectar aquestes

5.1.3 En síntesi

En aquest treball ens hem proposat de dissenyar un extractor més exhaustiu i eficient que els sistemes actuals de reconeixement automàtic de terminologia; i una de les diferències importants de la nostra proposta respecte dels sistemes existents és la delimitació de l'objecte d'extracció, que en el nostre cas abasta no només les unitats referencials —és a dir les UT—, sinó qualsevol unitat de significació especialitzada que des d'una òptica d'especialista sigui pertinent de recollir dels textos especialitzats, amb l'excepció d'iconografies, fotografies, dibuixos i esquemes que depassen l'abast general d'aquest treball. La finalitat del sistema que proposem és la de fornir a l'usuari, en una primera fase, una **llista de segments *candidats* a USE tan refinada com sigui possible.**

A mode de resum, presentem un quadre amb les USE que un SEACUSE que es proposi d'aconseguir un nivell d'exhaustivitat alt hauria de recuperar dels textos de l'àmbit de les ciències de la salut:

unitats pel fet que presenten les mateixes estructures que les UTP.

Unitats de significació especialitzada (USE)

1. *USE lingüístiques*

1.1 USE lingüístiques monolèxiques

1.1.1 simples

1.1.1.1 **nominals**

1.1.1.2 verbals

1.1.2 derivades

1.1.2.1 **nominals**

1.1.2.2 verbals

1.1.2.3 adjectives

1.1.2.4 adverbials

1.1.3 **compostes patrimonials nominals**

1.1.4 **compostes cultes nominals**

1.1.5 **sigles**

1.2 USE lingüístiques polilèxiques[5]

1.2.1 **unitats terminològiques polilèxiques** (UTP)

1.2.2 unitats fraseològiques especialitzades nominals (UFE)

2. *USE no lingüístiques*

2.1 símbols

2.2 noms científics en llatí

2.3 fórmules químiques

5.2 Elements i estratègies per detectar les USE pertinents en medicina

[5] El quadre verd indica les unitats que detecten els SEACAT clàssics i, en canvi, el quadre blau les unitats que creiem que hauria de recollir un sistema basat en els buidatges manuals fets per especialistes. En negreta hem remarcat les USE referencials, és a dir les UT.

Una vegada definit l'objecte d'extracció, ens proposem d'establir els principals elements de les USE de l'àmbit de les ciències de la salut en els quals es basaran les estratègies que permetran que un SEACUSE les recuperi de manera automàtica. Aquestes estratègies se seleccionaran segons el tipus d'USE de què es tracti. Per això, en primer lloc, presentarem els trets principals que caracteritzen cada tipus d'USE i, en segon lloc, les estratègies específiques que utilitzarà un extractor per detectar les unitats de cada tipus, encara que algunes són comunes a més d'un tipus:

- USE monolèxiques simples
- USE monolèxiques complexes: derivats, compostos i sigles
- USE polilèxiques: UTP i UFE
- símbols i fórmules
- noms científics en llatí.

5.2.1 USE monolèxiques simples

5.2.1.1 Elements de reconeixement

Les USE monolèxiques simples —nominals o verbals— són difícils de reconèixer automàticament perquè el seu caràcter especialitzat és totalment idiosincràtic. Són, doncs, un tipus d'unitats que no posseeixen característiques morfològiques ni sintàctiques explícites, que permetin detectar-les automàticament. Consegüentment, per poder-hi accedir, cal recórrer a altres estratègies lexicogràfiques i/o contextuals.

L'anàlisi de corpus lexicogràfics i textuals demostra que el nombre de termes simples en relació amb altres unitats especialitzades —unitats monolèxiques complexes i unitats polilèxiques— és molt baix. Tot i així, és interessant d'extreure dels textos les USE simples, perquè són una peça clau del lèxic especialitzat almenys per dues raons: en primer lloc, perquè són la base de

termes formats per derivació, i, en segon lloc, perquè són el nucli de moltes unitats sintagmàtiques, de les quals són l'hiperònim o el merònim[6].

Però la identificació automàtica de les USE simples no resulta fàcil ni formalment ni semànticament, perquè moltes corresponen també a unitats del lèxic comú i, com a tals, figuren en els diccionaris de llengua general, raó per la qual podria semblar que no tenen caràcter especialitzat, tot i que, semànticament indiquen conceptes específics en medicina[7].

5.2.1.2 Estratègies d'extracció automàtica

Dues són les estratègies que pensem que poden detectar les USE monolèxiques simples: la primera és l'ús d'un **diccionari d'USE simples** nominals i verbals desplegat morfològicament en què cada entrada porti associada una etiqueta de classe semàntica. I la segona consisteix a utilitzar una estratègia més complexa basada en els elements que sobre les USE facilita el seu **context lingüístic**: ens referim a recursos lingüístics del text que indiquen la presència d'USE (metatermes, paratermes, termes genèrics, connectors, patrons fixos, marques tipogràfiques, etc.).

Les investigacions recents en detecció terminològica automàtica atorguen un paper central als corpus textuals, com assenyalen Habert i al. (1997: 114-115): "*On est passé d'une conception logique à une conception distributionnelle selon laquelle le sens d'un mot et plus largement d'une unité textuelle peut se décrire par les contextes dans laquelles il figure (...) La plupart de travaux en sémantique de corpus reposent sur l'idée que le sens se construit en contexte mais aussi par le contexte*." En aquesta línia s'inclou també el treball de

[6] Per posar uns exemples: el *Diccionari Enciclopèdic de Medicina* (1990) inclou 170 unitats sintagmàtiques en què el nucli és el terme simple *os*, 70 en què el nucli és *taca* i 26 que tenen com a nucli *ull*.

[7] Com, per exemple, manifestacions biològiques de les malalties (*taca, gra, grip, xoc, tic, tel*); o parts i components del cos humà (*call, ull, cama, peu, mà, braç, cara, cap, cor, dit, pit, pèl, pols, buf, coll, cos, cul, dent, fel, gen, tronc).* Observem també que molts termes simples en medicina són monosíl·labs.

Pearson (1998), el qual, per localitzar automàticament els termes d'un text, es val de la informació que el context li proporciona a través d'indicis lingüístics que assenyalen la presència de termes.

5.2.2 USE monolèxiques complexes: derivats, compostos, sigles

Les USE monolèxiques complexes que un sistema hauria de poder extreure automàticament dels textos són unitats derivades, compostos patrimonials, compostos cultes i sigles.

5.2.2.1 USE derivades

5.2.2.1.1 Elements de reconeixement

De l'anàlisi de les dades del corpus textual, hem pogut comprovar que les USE derivades pertinents en medicina són noms adjectius, verbs o adverbis. I perquè un SEACUSE les pogués reconèixer, hauria de tenir presents tres peculiaritats d'aquest tipus d'unitats: en primer lloc i des del punt de vista lexicomorfològic, que les USE derivades es formen a partir d'una base lèxica que és sempre una USE; en segon lloc, que aquestes bases es combinen amb un nombre d'afixos molt més limitat que els afixos que s'adjunten a les bases lèxiques no especialitzades; i en tercer lloc, el fet que la base lèxica de les USE derivades sempre està relacionada o bé amb un terme simple o bé amb un formant clàssic pertinent en medicina.

D'acord amb aquestes observacions, sembla imprescindible que un extractor hagi d'associar les USE que comparteixen la mateixa base lèxica i relacionar aquesta base amb una USE simple i/o amb un formant grecollatí.

Les bases de les USE derivades (noms i verbs) poden ser USE nominals (*postinfart: infart, prepart: part, subencèfal: encèfal*, etc.; *receptar: recepta, remeiar: remei, respirar: respir, saturar: satura, sondar: sonda, vacunar: vacuna*, etc.), USE verbals (*circulació: circular, desintoxicació: desintoxicar:*

tòxic, destil·lació: destil·lar, infiltració: infiltrar, intubació: intubar: tub, etc.; *desenguixar: enguixar, desparasitar: parasitar, reepitelitzar, epitelitzar*, etc.) o USE adjectives *(sensibilitat: sensible, immunitat: immune, toxicitat: tòxic, desenguixar: enguixar, desparasitar: parasitar, desintoxicar: intoxicar*, etc.).

Pel que fa als adjectius monolèxics derivats, les nostres dades constaten que majoritàriament tenen com a base un nom terminològic que pot ser simple (*febril: febre, gripal: grip, nasal: nas, ossi: os, petequial: petèquia*) o compost (*immunològic: immunologia, neurològic: neurologia, endotelial: endoteli, pectoral: pectoral, cromosòmic: cromosoma*)[8]. A més dels adjectius formats a partir de termes, hi ha un grup d'adjectius que corresponen a participis d'un verb que alhora està format per una base especialitzada nominal o adjectiva (*empestat, -ada (empestar), immunitzat, -ada (immunitzar)*).

Finalment, els únics adverbis pertinents des del punt de vista terminològic formen part del grup dels *adverbis en -ment*[9], creats a partir d'un adjectiu especialitzat que gairebé sempre és un adjectiu derivat que conté formants cultes (*clínicament, biològicament, immunològicament, histològicament*).

5.2.2.1.2 Estratègies d'extracció automàtica

L'estratègia que proposem que un SEACUSE segueixi per detectar les USE monolèxiques derivades pertinents consisteix, primer, a agrupar les unitats del text que presentin la mateixa base lèxica i, després, a relacionar aquesta base lèxica amb una USE simple o amb un formant culte pertinent en el domini mèdic.

[8] En algunes ocasions el nom és un formant grecollatí: *maternal* prové del llatí *mater, mèdic, -a* prové de *medicus* i, algunes vegades, el sufix també és culte (per exemple en els adjectius formats a partir del sufix *-itis* o *-leg).*

[9] Pearson (1998: 142-143) també s'interessa pels adverbis que anomena *focalitzadors (commonly, especially, exceptionally, frequently, generally, mainly, etc.).* Aquest tipus d'adverbis tenen la funció de "*render a statement generally applicable or to restrict the scope of the reference*".

reepitelitzar[10]*, epitelitzar, epiteli* ⇒ *epiteli*

vacunar, vacunació, vacuna ⇒ *vacuna*

orofaringi, -íngia[11]*, orofaringe, faringe* ⇒ *oro-, faringe*[12]

lesionar, lesionat, -ada, lesió ⇒ lesió

epidèrmia, epidermis, dermis, ⇒ *epi-*[13]*, -dermi*[14]

histologia, histològic, -a, histològicament ⇒ *histo-*

Amb aquest procediment es poden detectar les USE monolèxiques derivades del text, tenint en compte que estan relacionades o amb una USE simple o amb un formant grecollatí. Per dur a terme aquest procés, caldrà que l'extractor compti amb un diccionari de formants cultes i un diccionari d'USE simples que actuïn com a filtres.

[10] A més, en el cas dels verbs, com proposàvem a l'apartat 3.2.1.2 del capítol tercer, seria interessant que un SEACUSE utilitzés un analitzador sintàctic amb etiquetes semàntiques per poder detectar patrons argumentals dels verbs especialitzadament significatius i accedir d'aquesta manera a la fraseologia verbal, fraseologia que, tot i que des del punt de vista de l'especialista, no és pertinent de recollir, des d'altres aproximacions sí que ho és. Es tracta de casos com:

- algun agent (microorganisme) X alguna part del cos humà: *contaminar, infectar, infestar, parasitar.*

[11] Els adjectius relacionals tant a WordNet com a EuroWordNet és lliguen al nom a partir del qual s'han format. Són un apèndix dels noms i no tenen una entrada independent.

[12] I a més també hi estarien relacionats els termes integrats pel formant *oro-*: *orocinasa, orolingual, oronasal*, etc.

[13] I també les paraules que estiguin formades pel formant *epi-*: *epidèmia, epicauma, epigastri,* etc.

[14] I també *gerodèrmia, helodèrmia, crisodèrmia, espasmodèrmia,* etc.

5.2.2.2 Compostos patrimonials

5.2.2.2.1 Elements de reconeixement

En el domini de la medicina els compostos patrimonials no són gens utilitzats i encara menys en els textos escrits adreçats a especialistes; els pocs noms formats per composició patrimonial s'usen en discursos orals o en textos de divulgació, i, generalment, denominen instruments (*comptagotes, comptaglòbuls, abaixallengües, tirapits, tirallet, portaagulles, portacames, portacuixes, portadrenatge, portaplaquetes*) o plantes medicinals (*matafaluga, matafoc, matallums, mataparent, matapoll*).

Hem observat també que tots els compostos patrimonials documentats presenten la mateixa estructura morfosintàctica: $[V\ [N]_{SN}]_N$, en la qual el nom és l'argument temàtic del verb. I, en el cas que el compost es refereixi a un instrument, el nom d'aquesta construcció és sempre un terme pertinent en medicina (*tira**pits**, abaixa**llengües**, porta**plaquetes***, etc.)[15].

5.2.2.2.2 Estratègies d'extracció automàtica

Per detectar els compostos patrimonials un extractor pot relacionar el segon component amb termes simples del diccionari o amb els derivats del corpus i comprovar si tenen caràcter especialitzat en l'àmbit mèdic. Aquest és el cas dels noms compostos d'instruments en què hem comprovat que el seu segon constituent és sempre una UT: *abaixa**llengües,** compta**glòbuls,***

[15] Semànticament, els noms d'instrument que segueixen aquesta estructura es poden descriure com a "*instrument que serveix per* $V_{funció}$ + $N_{objecte}$", per tant són compostos de nucli exocèntric, el predicat dels quals pot interpretar-se literalment a partir del significat del seus components. En canvi, el significat dels noms de plantes no és literal, sinó sempre d'ordre metafòric. Alguns casos amb significat componencial de nucli extrínsec: ***comptaglòbuls:*** *instruments que serveix per comptar glòbuls;* ***abaixallengües:*** *instrument que serveix per abaixar la llengua;* ***portaplaquetes:*** *instrument que serveix per portar plaquetes.* Però, en el cas dels noms vulgars de la biologia, el nom sol tenir relació amb una metàfora: ***matallums****: planta herbàcia de la família de les crassulàcies, suculenta, que fa nombroses rosetes de fulles bassals, semblants a petites carxofes, i tiges floríferes erectes, i espècies afins, que viuen en*

*compta**gotes,** tira**pits,** porta**agulles**, porta**cuixes**, porta**drenatge**, porta**plaquetes.***

Pel que fa als noms vulgars de les plantes —tot i que, cal tenir present que la possibilitat de trobar-ne un en un text científic és ínfima—, l'estratègia de relacionar les bases lèxiques no és vàlida perquè es tracta de bases que literalment no són especialitzades ja que són fruit d'un procés de metaforització que l'extractor no pot interpretar. En aquests casos, doncs, l'extractor ha d'utilitzar una via més idiosincràtica com, per exemple, un diccionari.

5.2.2.3 Compostos cultes

5.2.2.3.1 Elements de reconeixement

Els resultats de l'anàlisi duta a terme en els capítols anteriors ens ha permès confirmar que en medicina els compostos a la manera culta[16] són molt nombrosos i són la base de la majoria del seu lèxic especialitzat. Com afirmen López Piñero i Terrada Ferrandis (1990), al voltant d'unes mil arrels de procedència grega o llatina i d'uns vuitanta afixos clàssics componen la gairebé totalitat del lèxic de la medicina. Això significa que, en ciències de la salut, amb un nombre relativament petit d'elements grecollatins es genera un nombre molt elevat d'unitats.

Reforçant aquesta idea, diversos estudis, [Quintana, 1989] [López Piñero i Terrada i Ferrandis, 1990] [Love i Davis, 1990] [Bernabeu i al., 1995] [Navarro, 1996], han constatat que el professional de la medicina no coneix la totalitat de paraules mèdiques, però que, en canvi, és capaç de desxifrar-ne el

murs, roques i teulades. El nom popular és degut al fet que físicament recorda un matallums o apagallums, instrument per apagar els ciris.

[16] Els compostos cultes inclouen diferents variants segons la combinació de formants: compostos formats amb dos formants grecs, amb dos formants llatins, amb un formant grec i un de llatí, amb un mot català i un formant llatí, amb un mot català i un formant grec, amb dos formants grecs i un de llatí, etc.

significat sense haver-les vist o escoltat mai. I és capaç de fer-ho perquè coneix els mecanismes de formació de la majoria de paraules mèdiques i té interioritzats un conjunt de formants grecs i llatins que li permeten codificar i descodificar els significats de les paraules que s'usen en aquest àmbit.

5.2.2.3.2 Estratègies d'extracció automàtica

Per extreure els compostos cultes, d'acord amb la lògica que caracteritza el coneixement que un especialista té de la terminologia mèdica, un extractor pot simular l'estratègia del professional usant **un diccionari d'uns 1.100 formants grecollatins pertinents en el domini mèdic** i un nombre reduït de regles de combinació dels formants [17].

Però a més de la utilitat dels formants clàssics per a la detecció de les USE compostes a la manera culta, els formants grecollatins poden servir també per, en primer lloc, comprovar si una base lèxica és especialitzadament pertinent o no. Així, la detecció dels formants cultes permet establir una cadena de reconeixement que comença amb la detecció del formant culte i acaba amb el reconeixement de les USE polilèxiques:

> formants cultes ⇒ derivats o/i compostos ⇒ unitats sintagmàtiques
>
> hepat(o)- + *-itis* ⇒ *hepatitis* ⇒ *hepatitis vírica, hepatitis crònica agressiva, hepatitis epidèmica, hepatitis fulminant, hepatitis supurada*, etc.

[17] Les regles de combinació de formants haurien de fer referència a les vocals d'enllaç de les arrels grecollatines. Normalment, la *-o* és la vocal d'enllaç en el cas dels formants de la llengua grega i la *-i* en el dels formants de la llengua llatina. En els formants híbrids sol prevaler l'origen del segon formant, encara que en les paraules de creació recent se sol tendir a utilitzar sempre la *-o*.

sero- (ser-), -logia, sèrum, serologia, serològic, -a, serològicament, seroaglutinació, seromucós, serologia, sèrum d'Hayem, sèrum hemàtic, sèrum polivalent[18], etc.

En segon lloc, la possibilitat de relacionar totes les USE que comparteixen la mateixa arrel amb un terme simple o amb un formant clàssic no només facilita la tasca d'extracció automàtica, sinó que també ajuda l'usuari a seleccionar les unitats pertinents d'entre les USE candidates proposades pel sistema, perquè la productivitat d'una determinada arrel morfosemàntica en els textos de medicina és un element més per decidir si una unitat és o no especialitzada.

En conseqüència amb aquesta polivalència, per saber si una base lèxica és especialitzada seria molt útil i productiu que un SEACUSE construís sèries de mots que compartissin almenys una arrel o formant pertinent, mitjançant un diccionari de formants grecollatins i un diccionari d'USE simples.

5.2.2.4 Sigles

5.2.2.3.1 Elements de reconeixement

Les sigles, com hem analitzat en l'apartat 3.2.2 del capítol tercer, presenten algunes característiques que les singularitzen:

1. Sovint tenen un **caràcter internacional.**

2. Són semànticament **opaques** ja que, formalment, l'única relació que mantenen amb el sintagma al qual substitueixen és la grafia de les inicials.

[18] En el DEM (1990) figuren 203 accepcions de *sèrum*.

3. Presenten una **forma tipogràfica singular,** ja que s'escriuen en lletres majúscules, normalment juntes, sense punts ni espais en blanc[19].

4. Solen estar formades per combinacions de **dues a cinc lletres**, que, normalment, substitueixen un segment de tres paraules referencials[20].

5. Apareixen habitualment en el text ***entre parèntesis***, darrera mateix de tot el segment desenvolupat al qual substitueixen[21].

5.2.2.4.2 Estratègies d'extracció

Els SEACUSE tenen dues vies possibles per detectar les sigles d'un text:

- utilitzar un diccionari de sigles
- recórrer a alguns aspectes del text (disposició en el text de les sigles, tipografia, determinació, etc.).

La primera opció presenta, en el nostre cas, dos inconvenients: en primer lloc el fet que no disposem d'un diccionari de sigles mèdiques per al català, diccionari que, d'altra banda, si es construís hauria de ser molt ampli (recordem que l'any 1989 un diccionari de sigles mèdiques per a l'anglès ja en contenia més de 15.000 [Heister, 1989]); i, en segon lloc, el fet que utilitzar un diccionari de sigles no permetria detectar les sigles neològiques, i això seria un problema important en un domini professional amb tanta innovació lèxica com les ciències de la salut.

[19] Si una sigla apareix en minúscula (signe del seu grau elevat de lexicalització: *làser, radar, sida*), es tractarà com una unitat simple.
[20] Això és lògic si pensem que no existeixen UT que incloguin més de cinc lexemes i que la majoria estan formades per un nom i un o dos complements.
[21] Les sigles molt conegudes pels especialistes no apareixen entre parèntesis i aquesta pèrdua de relació entre la sigla i el sintagma que substitueix és el principal problema amb què topen els SEACUSE, ja que aïlladament una sigla és totalment idiosincràtica.

La segona via de detecció de les sigles sembla, des del punt de vista lingüístic i informàtic, més àgil. Implica recórrer a heurístiques pragmaticolingüístiques que aprofitin els elements característics de les sigles que hem comentat en l'apartat anterior.

La nostra opció, però, s'inclina per utilitzar una estratègia mixta en què els dos tipus de recursos —el diccionari i les heurístiques— siguin complementaris. Així, el SEACAT utilitzarà un **diccionari mínim de reconeixement** format per les sigles internacionalitzades més usades en l'àmbit mèdic (que són les que probablement no aniran entre parèntesis) i un **conjunt d'instruccions** simples basat en aspectes gràfics per detectar la resta de sigles, neològiques o no.

5.2.3 USE polilèxiques: UTP i UFE

Els SEACAT tradicionals, com ja hem presentat, detecten totes les UTP explícites dels textos especialitzats mitjançant un conjunt de patrons estructurals. En general, l'exhaustivitat, pel que fa aquest tipus d'unitats, és molt elevada. Hem descrit en el capítol tercer com aquests sistemes només deixen de reconèixer les unitats implícites del text, és a dir les unitats superposades que inclouen més d'una UT, i sobretot les *unitats amagades* per anàfora discursiva. No obstant això, en aquell moment vam concloure que, segons els paràmetres de la cerca (el corpus i la finalitat del buidatge), el fet de no recuperar aquestes unitats no constituïa un problema, perquè sovint, si en un lloc del text una unitat lèxica apareix anaforitzada o superposada, en un altre lloc del corpus aquesta mateixa unitat apareix explícitament.

Per tant, els resultats obtinguts de l'anàlisi dels principals extractors automàtics permeten dir que, en general, la recuperació de les USE polilèxiques no genera silenci, però, en canvi, genera molt de soroll per tal com les estructures morfosintàctiques que fan servir aquests sistemes per

extreure les USE polilèxiques no són exclusives d'aquestes unitats. S'estima que entre el 45% i el 75% dels candidats proposats per aquests sistemes s'ha de rebutjar, de manera que l'estratègia de reconeixement de les USE polilèxiques basada només en patrons morfosintàctics constitueix un filtre massa permissiu.

Per donar-li més restrictivitat és necessari proposar altres enginys de reducció del soroll que filtrin els segments discursius que presenten estructures anàlogues. En aquesta via, alguns autors han proposat estratègies que complementen els filtres estructurals: estratègies estadístiques (Bourigault, 1993), estratègies sintàctiques (Jacquemin, 1994), estratègies semàntiques (Naulleau, 1998), estratègies contextuals (Pearson, 1998). La nostra proposta consisteix a utilitzar estratègies de diversa naturalesa que, combinades, primer eliminin les unitats discursives i després facilitin la distinció entre UTP, UFE i combinacions recurrents.

Com fan la majoria dels sistemes d'extracció, centrarem la detecció de les USE polilèxiques en les dues estructures més productives, que són la base de totes les USE polilèxiques dels textos de medicina i, en general, de tots els textos especialitzats en llengües romàniques, que són:

$[N[A]_{SAdj}]_{SN}$
$[N\ [de\ (art)\ [N]_{SPrep}]_{SN}$.

5.2.3.1 $[N[A]_{SAdj}]_{SN}$

5.2.3.1.1 Elements de reconeixement

L'estructura $[N[A]_{SAdj}]_{SN}$ és la més productiva dels textos especialitzats, però alhora és una de les que ocasiona més soroll per tal com superficialment no es pot saber si és especialitzada o no. L'anàlisi del corpus textual de medicina mostra que el caràcter especialitzat o discursiu d'una unitat amb estructura

$[N[A]_{SAdj}]_{SN}$ depèn de la naturalesa especialitzada o no del nom i de l'adjectiu que la integren. En aquesta línia, distingim quatre possibilitats:

$[N_{esp}\ [A_{esp}]_{SAdj}]_{SN}$ = UT

$[N_{no\ esp}\ [A_{esp}]_{SAdj}]_{SN}$ = UT

$[N_{esp}\ [A_{no\ esp}]_{SAdj}]_{SN}$ = UT o UD

$[N_{no\ esp}\ [A_{no\ esp}]_{SAdj}]_{SN}$ = UD o UL

D'aquestes possibilitats inferim els resultats següents:

a) Si el nucli d'una unitat és una USE i l'adjectiu que el complementa també, el resultat és sempre una UTP:

> *antibiòtic bactericida, coll uterí, degeneració lenticular, dermatitis actínica, embòlia cerebral, inhibició enzimàtica, injecció endodèrmica, intervenció quirúrgica, llengua leucoplàstica, neurosi gàstrica, punció abdominal, tractament mèdic*, etc.

b) Si el nom de la unitat no és una USE, però l'adjectiu que el classifica sí que ho és, la combinació resultant és també una UTP[22]:

> *bossa serosa, canal coclear, capacitat pulmonar, punt alveolar, timbre nasal, vas limfàtic*, etc.

c) Si, en canvi, el nom és una USE i l'adjectiu no ho és, es pot tractar o d'una UTP (*augment alt, aplicació ràpida, berruga plana, desaparició ràpida, difusió mundial, disminució lenta, edema maligne, eritema simple, febre groga, laringitis aguda, nefritis local, població nova*, etc.) o d'una unitat

discursiva (UD), sense cap caràcter especialitzat (*hepatitis rara, radiografia dolenta, injecció forta, pacient diferent, limfopatia generalitzada, reacció adversa, reacció local, tractament actiu*, etc.). En aquest segon cas, el sistema hauria de proposar com a terminològics els nuclis nominals de cada unitat discursiva i eliminar el soroll que genera l'adjectiu que els modifica.

d) Finalment, si un segment no està format per cap USE, la unitat resultant és o una unitat de discurs lliure (*element clau, estudi recent, mecanisme possible, viatger internacional*) o una unitat lèxica no pertinent per a l'especialitat (*animal domèstic, conca mediterrània, continent americà, zona urbana*) i, per tant, cap dels dos casos és pertinent perquè els reculli un extractor.

5.2.3.1.2 Estratègies d'extracció automàtica

Una vegada establertes les combinacions amb estructura $[N[A]_{SAdj}]_{SN}$ que interessa que un extractor proposi i les combinacions amb la mateixa estructura que hauria de silenciar, hem de formular quina és l'estratègia que hauria d'aplicar el sistema per poder assolir una extracció adequada de les USE polilèxiques que presenten aquesta estructura.

D'acord amb les condicions establertes en l'apartat anterior, el primer que hauria de poder fer el sistema és determinar el caràcter especialitzat o no dels constituents d'una unitat. I, per determinar si un nom o un adjectiu són especialitzats, el SEACUSE pot fer servir els mateixos mecanismes que utilitza per saber quan una USE monolèxica simple, derivada o composta és pertinent[23].

[22] Aquesta combinació és molt productiva quan la UTP pertany a la classe semàntica de les parts del cos humà, però el nucli de la unitat es refereix a una part del cos que rep el nom d'un objecte del món amb el qual manté un grau de semblança.

[23] Així, per saber si una USE simple és terminològica un SEACUSE hauria de fer servir un diccionari d'USE simples amb etiquetes semàntiques; en canvi, per saber si una USE derivada o composta és terminològica hauria de relacionar les paraules que comparteixen una mateixa arrel amb un terme simple de medicina o amb un formant grecollatí pertinent

Una vegada establert el caràcter especialitzat o no especialitzat d'una unitat, l'extractor podrà proposar com a candidates a USE totes les unitats del text que responguin a la combinació $[N_{esp}\ [A_{esp}]]_{SN}$ i rebutjar les unitats que presentin l'estructura $[N_{no\ esp}\ [A_{no\ esp}]]_{SN.}$

A més d'establir el caràcter especialitzat dels seus constituents, per detectar correctament les unitats que responguin a la combinació $[N_{no\ esp}\ [A_{esp}]]_{SN,}$ el sistema haurà de distingir si es tracta d'una unitat de discurs o d'una UT. D'acord amb la hipòtesi que aquesta estructura només és terminològica si el nom no forma part del conjunt de noms quantitatius, ni del d'organitzadors de l'estructura del discurs, ni del de paratermes de medicina, el sistema hauria d'acudir a la consulta de filtres constituïts per aquest tipus de paraules per poder proposar que la unitat és UT o per rebutjar-la en el cas de provocar soroll.

Finalment, la combinació més problemàtica d'aquest grup és $[N_{esp}\ [A_{no\ esp}]_{SAdj}]_{SN}$ perquè després de determinar el caràcter especialitzat del nom i el caràcter no especialitzat de l'adjectiu, el sistema necessita estratègies complementàries per saber si la unitat resultant és terminològica o no; dit altrament, necessita saber quins són els adjectius que, en complementar determinades classes de termes, converteixen la unitat en especialitzada.

L'únic mecanisme per desambiguar aquestes unitats i saber si el segment final és terminològicament pertinent és utilitzar mecanismes basats en la semàntica combinatòria. Aquesta estratègia implica que els diccionaris que l'extractor utilitzi, tant el d'USE simples com el de formants cultes, han d'incloure etiquetes semàntiques que permetin elaborar patrons semanticoformals que actuïn de filtre restrictiu. En l'apartat 5.3.1

---

en medicina. El fet que l'arrel d'una unitat sigui en el diccionari de termes simples o en el diccionari de formants cultes prova que la unitat és especialitzada, i, en el cas contrari, que la unitat no és especialitzada. Es tracta d'estratègies que ja hem plantejat en els apartats corresponents (5.2.1, 5.2.2 i 5.2.3).

treballarem sobre les restriccions de les USE polilèxiques que pertanyen a les principals classes de noms pertinents en medicina.

5.2.3.2 [N [de (art) [N]$_{SPrep}$]$_{SN}$

5.2.3.2.1 Elements de reconeixement

Al costat de l'estructura [N[A]$_{SAdj}$]$_{SN}$, l'estructura[N [de (art) [N]$_{SPrep}$]$_{SN}$, malgrat que no és tan productiva com la primera, és la que provoca els índexs més alts de soroll. Aquesta estructura, per les característiques que presenta el complement (sintagma preposicional sovint determinat), té menys força lexicalitzadora que l'anterior i, per això, en els textos especialitzats, juntament amb UTP d'aquesta estructura, trobem també unitats discursives, fraseològiques i combinacions recurrents.

La pertinència d'una unitat amb estructura [N [de (art) [N]$_{SPrep}$]$_{SN}$ depèn, com en el cas de l'estructura [N[A]$_{SAdj}$]$_{SN}$, de la naturalesa especialitzada o no dels seus constituents i, en alguns casos també del caràcter eventiu del nucli de la unitat i del fet que el nom nucli del sintagma preposicional estigui subcategoritzat com a propi o com a comú. D'acord amb aquestes variables, distingim les combinacions següents:

[N$_{esp}$ [de [N$_{patronímic}$]]$_{SPrep}$]$_{SN}$ = UT
[N$_{esp}$ [de (art) [N$_{no\ esp}$]]$_{SPrep}$]$_{SN}$ = UT o UD
[N$_{esp}$ [de (art) [N$_{esp}$]]$_{SPrep}$]$_{SN}$ = UT o combinació recurrent
[N$_{deverbal}$ [de (art) [N$_{esp}$]]$_{SPrep}$]$_{SN}$ = UFE
[N$_{deverbal}$ [de (art) [N$_{no\ esp}$]]$_{SPrep}$]$_{SN}$ = UD
[N$_{quant}$ [de (art) [N]]$_{SPrep}$] $_{SN}$ = UD
[N$_{paraterme}$ [de (art) [N]]$_{SPrep}$] $_{SN}$ = UD
[N$_{no\ esp}$ [de (art) [N$_{no\ esp}$]]$_{SPrep}$] $_{SN}$ = UD o UL
[N$_{no\ esp}$ [de (art) [N$_{esp}$]]$_{SPrep}$] $_{SN}$ = UT

Aquesta taula de combinació ens permet arribar a les següents conclusions:

a) L'estructura $[N_{esp}$ [de $[N_{patronímic}]]_{SPrep}]_{SN}$ dóna com a resultat una UT i és, en medicina, l'estructura amb complement preposicional més freqüent[24]: *amígdala de Luschka, banda de Broca, distròfia de Fuchs*, *flegmó de Monat, histerectomia d'Amman, malaltia de Miller, membrana de Browman, òrgan de Corti, tendó d'Aquiles,* etc.

b) Si el primer nom d'una seqüència [N [de (art) $[N]_{SPrep}]_{SN}$ és un terme però el segon no, aleshores la combinació resultant pot ser o una unitat discursiva (*zoonosi de distribució mundial, xeringa de l'agulla llarga,* etc.), per bé que aquesta combinació discursiva és molt poc freqüent en els textos molt especialitzats, o bé una UTP (*filtre de vidre, crani de torre, fetge d'acordió, febre dels pantans*, etc.).

c) Si els dos noms de la seqüència són terminològics, segons les característiques semàntiques dels seus constituents podem tenir o una UTP (*hipoglucèmia d'esforç, síndrome d'immunodeficiència,* etc.) o una combinació especialitzada recurrent (*paràlisi del nervi bucal, paràlisi del braç esquerre, picada de paparra,* etc.). Hem constatat que, normalment, en el primer cas hi ha absència d'article i en el segon, l'article hi sol ser present[25].

d) Una de les combinacions més productives en medicina és la construïda per un nom deverbal i un sintagma preposicional introduït per la preposició *de* que té com a nucli un terme, per bé que aquesta seqüència és sempre una UFE:

[24] A partir de les dades hem pogut constatar que el nom propi que especifica un terme sempre apareix indeterminat, perquè en català i en castellà la indeterminació reforça el caràcter genèric de la unitat. Semànticament, la seqüència sempre pot parafrasejar-se de la manera següent:

> *Y ha descobert o ha patit X, en què Y és un investigador o un pacient famós o el primer pacient i X la cosa descoberta o patida.*

[25] En l'apartat 4.3.1 del capítol anterior ja hem fet referència a aquesta combinació amb l'exemple *radiografia de la pelvis.*

> *absorció del sèrum, acumulació de líquid intravascular, augment de la permeabilitat vascular, elevació de la creatinafosfocinasa, instauració del tractament, prevenció de la malaltia, tractament de la febre botonosa*, etc.

Hem constatat també que el sintagma preposicional de la majoria de les UFE nominals que presenten aquesta estructura està determinat per l'article definit.

e) Si trobem una seqüència [N [de (art) [N]$_{SPrep}$]$_{SN}$ en què la base del primer nom és un verb no especialitzat i el segon nom no és terminològic, podem concloure que la seqüència és discursiva: *augment del nombre de factors, eliminació de la vegetació,* etc. Les dades, però, demostren que en el textos especialitzats de medicina aquesta combinació no és gaire freqüent.

f) En els textos especialitzats també hi ha moltes seqüències polilèxiques que tenen com a nucli un nom quantitatiu (*majoria de casos, la resta de símptomes, la totalitat dels pacients, el conjunt de característiques*) o un nom que facilita l'estructuració del discurs ((al) *final del capítol, (al) llarg de l'article, (a l)'inici del capítol,* etc.[26]), que sempre són unitats complexes discursives. En el primer cas només provoca soroll el nucli de l'estructura perquè el complement tendeix a ser una UT, en canvi, en el segon cas, sempre s'ha de rebutjar tot el segment.

g) Si el primer nom d'una seqüència [N [de (art) [N]$_{SPrep}$]$_{SN}$ és un paraterme (és a dir una paraula que precedeix un terme i que, implícitament, en dóna informació semàntica), el sintagma nominal del complement d'aquestes seqüències és sempre una USE, però no tota la seqüència[27]:

[26] Observem que aquest tipus de noms molt sovint va precedit de la preposició locativa *a*.
[27] Els paratermes són molt abundants en els textos especialitzats i, per bé que ocasionen soroll, semànticament són interessants perquè donen informació pragmàtica sobre les UT

> *causa del xarampió, presència de taca negra, denominació de rickèttsia, començament de la malaltia, característica de la malaltia, existència d'infeccions subclíniques*, etc.

h) Si en una estructura [N [de (art) $[N]_{SPrep}]_{SN}$ cap dels constituents és un terme, es tracta d'un segment sense valor especialitzat que genera soroll i que pot ser, o bé una unitat discursiva (*mes de maig, darreria dels anys vuitanta*), o bé una unitat lèxica no pertinent en medicina (*bossa de pa, estació de l'any, rellotge de sorra,* etc.).

i) Finalment, si en els textos especialitzats de medicina trobem seqüències amb estructura [N [de (art) $[N]_{SPrep}]_{SN}$ en què el primer nom no és terminològic, però en canvi el segon sí que ho és, i si a més el primer nom no és ni un paraterme ni un quantitatiu ni un organitzador de l'estructura discursiva, aleshores es tracta d'una UT.

Les dades mostren que, en aquests casos, la unitat resultant sol denominar una part molt específica del cos humà. Aquesta especificitat s'aconsegueix a través de la determinació:

> *ala de la ròtula, angle de la costella, apèndix del testicle, base del cor, escorça del cristal·lí, falç del cervell, istme de la pròstata, radi del cristal·lí, tenda del cerebel, vàlvula de l'urèter, vel del paladar,* etc.

5.2.3.2.2 Estratègies d'extracció automàtica

(topogràfica, metalingüística, temporal, relacional, morfològica, etc.) i les relacionen conceptualment.

Ens proposem a continuació de suggerir els mecanismes per extreure les combinacions amb estructura [N [de (art) [N]$_{SPrep}$]$_{SN}$ pertinents, basats en les característiques que acabem de comentar.

Tinguem en compte que hem fet l'opció d'aplicar estratègies diferents segons el tipus d'unitat de què es tracti i també segons el tipus de combinació que aquestes unitats formin. Amb aquesta lògica hem argumentat que és pertinent de distingir el caràcter especialitzat o no especialitzat dels constituents de la unitat, i, en una altra perspectiva, el caràcter eventiu o no del nucli, el qual ens permet distingir entre una UTP i una UFE.

Així, per reconèixer les combinacions presentades als punts *a*, *d*, *e*, *f*, *g*, *h* i *i:*

[N$_{esp}$ [de [N$_{patronímic}$]]$_{SPrep}$]$_{SN}$ = UT
[N$_{deverbal}$ [de (art) [N$_{esp}$]]$_{SPrep}$]$_{SN}$ = UFE
[N$_{deverbal}$ [de (art) [N$_{no\ esp}$]]$_{SPrep}$]$_{SN}$ = UD
[N$_{quant}$ [de (art) [N]]$_{SPrep}$] $_{SN}$ = UD
[N$_{paraterme}$ [de (art) [N]]$_{SPrep}$] $_{SN}$ = UD
[N$_{no\ esp}$ [de (art) [N$_{no\ esp}$]]$_{SPrep}$] $_{SN}$ = UD o UL
[N$_{no\ esp}$ [de (art) [N$_{esp}$]]$_{SPrep}$] $_{SN}$ = UT

serà suficient utilitzar una sèrie d'estratègies basades en la caracterització dels elements formals —lèxics, morfosintàtics, tipogràfics— dels constituents de la unitat; però per a les combinacions *b* i *c:*

[N$_{esp}$ [de (art) [N$_{no\ esp}$]]$_{SPrep}$]$_{SN}$ = UT o UD
[N$_{esp}$ [de (art) [N$_{esp}$]]$_{SPrep}$]$_{SN}$ = UT o combinació recurrent

aquestes característiques no són suficients i, a més, caldrà aplicar-ne d'altres basades en trets morfosemàntics.

El procés que un extractor ha de seguir en tots els casos és el següent: en primer lloc, ha de saber si els constituents de cada unitat polilèxica tenen caràcter especialitzat; a continuació, pel que fa al nucli de les unitats, ha d'analitzar si es tracta d'un nom deverbal, d'un paraterme o d'un especificador de nom quantitat; i, finalment, pel que fa al complement, ha de saber si és un nom comú o propi.

Per determinar el valor especialitzat dels components d'una unitat, el SEACUSE pot utilitzar les mateixes estratègies que usa per detectar les USE monolèxiques (5.2.1 i 5.2.2). Per als noms deverbals, se servirà del programa de relació d'arrels que permet relacionar el nom deverbal amb el verb a partir del qual s'ha format. I per saber que un nom és propi, podrà recórrer a la grafia majúscula de la seva lletra inicial, tenint en compte la seva disposició en el text.

Per reconèixer i descartar les combinacions que ocasionen soroll, caldria que el sistema comptés amb un diccionari integrat per quantificadors, paratermes i organitzadors del discurs[28], que funcionés de filtre i evités que el sistema proposés com a candidats a USE aquest tipus de combinacions.

Descartats tots els segments que provoquen soroll per causa del primer nom de la seqüència, el sistema podrà proposar satisfactòriament la llista de les unitats que responen a les combinacions *a*, *d* i *i* de l'apartat anterior, que poden correspondre a UT:

$[N_{esp}\ [de\ [N_{patronímic}]]_{SPrep}]_{SN}$
$[N_{deverbal}\ [de\ (art)\ [N_{esp}]]_{SPrep}]_{SN}$
$[N_{no\ esp}\ [de\ (art)\ [N_{esp}]]_{SPrep}]_{SN}$

i rebutjar les que responen a les combinacions *e, f, g* i *h*, que no poden ser UT:

$[N_{deverbal}\ [de\ (art)\ [N_{no\ esp}]]_{SPrep}]_{SN}$

$[N_{quant}\ [de\ (art)\ [N]]_{SPrep}]\ _{SN}$

$[N_{paraterme}\ [de\ (art)\ [N]]_{SPrep}]\ _{SN}$

$N_{no\ esp}\ [de\ (art)\ [N_{no\ esp}]]_{SPrep}]\ _{SN}$

Però li mancarà encara desambiguar aquelles combinacions que podem donar lloc a UT i a unitats no terminològiques, és a dir les combinacions *b* i *c*:

$[N_{esp}\ [de\ (art)\ [N_{no\ esp}]]_{SPrep}]_{SN}$

$[N_{esp}\ [de\ (art)\ [N_{esp}]]_{SPrep}]_{SN}$

Per determinar si aquestes dues combinacions són UT, UD o combinacions recurrents, el sistema haurà d'aplicar filtres morfosemàntics de restricció, perquè les estratègies que combinen característiques morfològiques, sintàctiques, tipogràfiques o pragmàtiques no són suficients per discriminar quan una unitat amb estructura $[N_{esp}\ [de\ (art)\ [N_{no\ esp}]]_{SPrep}]_{SN}$ és especialitzada o no, i quan un segment $[N_{esp}\ [de\ (art)\ [N_{esp}]]_{SPrep}]_{SN}$ és terminològic o simplement recurrent. En l'apartat 5.3.1 proposarem alguns patrons semanticoformals pertinents en medicina que actuen de filtres de restricció per a aquestes combinacions.

[28] En l'apartat 4.3.3 del capítol anterior, hem elaborat una llista d'aquest tipus de noms a partir dels textos estudiats, que caldria completar a partir de l'anàlisi d'altres corpus textuals.

5.2.4 Símbols i fórmules

5.2.4.1 Elements de reconeixement

Els símbols que s'usen en els textos mèdics pertanyen, bàsicament, a les nomenclatures de la química, de la genètica i la citogenètica, de la immunologia i de la fisiologia. I també es fan servir alguns símbols de les unitats que integren el *Sistema Internacional d'Unitats*.

Tipogràficament, la majoria de símbols estan compostos o bé d'una lletra que s'escriu en majúscula o bé de dues lletres, i en aquest cas la primera s'escriu en majúscula, però la segona en minúscula: *H, F, P, S, Al, Be, Ca, Cl, Mg,* etc. Molts d'aquests símbols formen part de la *Taula Periòdica*, encara que també es fan servir símbols del *Sistema Internacional d'Unitats* i d'elements genètics, immunològics i fisiològics. Alguns dels símbols van acompanyats per números subindexats: B_1, B_2, B_6, B_{12}, Pg_1, Pg_2, etc.

Aquests trets gràfics poden ajudar un SEACUSE a detectar els símbols i les fórmules pertinents en el domini mèdic, tot i que cal tenir present que aquests tipus de caràcters depenen naturalment del format del text.

5.2.4.2 Estratègies d'extracció

Un extractor que vulgui recuperar automàticament els símbols i les fórmules especialitzades dels textos especialitzats es pot valer d'un diccionari de símbols freqüents en l'àmbit de la medicina amb la finalitat que li serveixi de filtre restrictiu[29]. Aquest filtre perquè pogués reconèixer símbols nous o fórmules podria complementar-se amb un programa molt senzill d'instruccions de reconeixement de símbols i fórmules basat en les

[29] Un filtre lèxic integrat per tres llistats de símbols: les unitats del *Sistema Internacional*, pertinents per a les professions de la salut; els elements de la Taula Periòdica; i els elements genètics, immunològics i fisiològics. L'Organització Mundial de la Salut l'any 1977 va publicar un volum de les unitats del SI relacionades amb les professions de la salut.

característiques tipogràfiques (majúscules, subíndex, combinació de lletres i números, posició discursiva, etc.).

5.2.5 Nomenclatures científiques

5.2.5.1 Elements de reconeixement

En medicina s'utilitzen diverses nomenclatures normalitzades, unes de molt consolidades, aprovades per la comunitat científica corresponent, que segueixen unes normes o pautes que les regulen; i unes altres encara no normalitzades. Al primer grup pertanyen:

- la nomenclatura química
- la nomenclatura botànica
- la nomenclatura zoològica
- la nomenclatura bacteriològica
- la nomenclatura anatòmica.

El grup de nomenclatures menys fixades està integrat per:

- la nomenclatura de la virologia
- la nomenclatura de la genètica
- la nomenclatura d'immunologia
- la nomenclatura de fisiologia[30].

Cada una d'aquestes nomenclatures segueix unes normes internes establertes per les comissions corresponents, que les fan fàcilment identificables i que faciliten la feina de detecció automàtica. Així, les nomenclatures de botànica, zoologia i bacterologia estipulen terminacions normalitzades per a tots els rangs taxonòmics. La base de tota la nomenclatura biològica és un sistema binomial que, amb algunes diferències

[30] Encara que aquestes dues estan revisades només parcialment.

de detall segons la temàtica, utilitza termes en llatí. En general, el nom científic que denomina una planta, un animal o un bacteri està format per dues unitats, la primera correspon al gènere i la segona a l'espècie; el gènere s'expressa en nominatiu singular, seguit de l'espècie, expressada mitjançant un adjectiu, un substantiu en aposició o un substantiu en genitiu. És prescriptiu que el nom científic s'escrigui en una tipografia diferent de la resta de l'escrit (normalment en cursiva i a vegades subratllat). La primera inicial del primer nom s'escriu en majúscula (*Rickettsia sibirica, Dermacentor variabilis*). El segon nom que indica l'espècie no té validesa per ell mateix i, per tant, no s'utilitza sol per referir-se a cap organisme; però, en cas que el context no sigui ambigu, és habitual indicar el nom del gènere amb la lletra inicial: *R. coronii, R. japonica, R. akari*, són espècies del gènere de les *Rickettsia*[31].

La nomenclatura química, a diferència de les altres, no utilitza noms en llatí, sinó arrels, sufixos i prefixos grecollatins i, així, el diccionari de formants que hem proposat haurà d'incloure també els afixos i les arrels pròpies de la nomenclatura química[32].

Finalment, pel que fa a la nomenclatura anatòmica, si bé la llista oficial està escrita en llatí, normalment s'utilitza traduïda a cada llengua, sobretot per a usos pedagògics i divulgatius. Cada unitat de la nomenclatura consta d'un

[31] A vegades, com indica Planas (1985: 137): "*El nombre científico va seguido de la inicial o abreviación del nombre del autor que citó por primera vez aquella especie. Así, los nombres de las especies que van seguidos de la letra L., son especies descritas por Lineo de las 4.236 especies clasificadas hasta 1758. Algunas especies han sido bautizadas por diferentes autores pudiendo variar la nomenclatura. Esto dificultó en principio la labor del sistemático, hasta que no se estableció un cuadro de equivalencias, que se denomina sinonimias, y se tendió a la uniformidad de las denominaciones a fin de no complicar más el difícil problema de la clasificación de ciertas especies.*"

[32] La IUPAC publica normes fixades de combinació dels formants químics que poden facilitar la tasca del reconeixement automàtic; per exemple, López Piñero i Terrada Ferrandis (1990: 67-77) en citen algunes: "*Según la nomenclatura sistemática de la propia IUPAC, todas las combinaciones de los metales y no metales con el oxígeno se nombran con la palabra óxido y el otro elemento en genitivo, indicándose las proporciones de ambos mediante raíces numerales de origen griego: óxido de diyodo, trióxido de diyodo, pentaóxido de diyodo, heptaóxido de diyodo, etc. En química orgánica, de particular importancia para la medicina, la nomenclatura de la IUPAC consiste esencialmente en una raíz que corresponde al esqueleto*

nucli nominal en indicatiu i d'un complement adjectiu que expressa localització o funció. Els adjectius de localització presenten entre ells oposició: *lobus anterior, lobus posterior, fovea coastalis inferior, fovea costalis superior, musculus palmaris longus, musculus palmaris brevis*, etc. Els adjectius de funció es basen en els dotze sistemes funcionals del cos: *nervus lacrimales, nervus auditus*, etc.

5.2.5.2 Estratègies d'extracció

Les característiques formals dels noms que integren les nomenclatures semblen suficients per poder detectar automàticament aquest tipus de noms que no formen part del llenguatge natural[33].

Així, pel que fa a la recuperació dels noms llatins que integren les nomenclatures biològiques, el sistema podria utilitzar un protocol basat fonamentalment en les característiques morfològiques i tipogràfiques d'aquestes unitats. En canvi, per a la detecció de la nomenclatura química, al marge d'un protocol de normes, caldria que l'extractor utilitzés també un diccionari dels formants (arrels i sufixos) propis de la nomenclatura química.

5.2.6 Conclusions

En els apartats anteriors hem proposat que un SEACUSE que tingui com a finalitat recuperar les USE dels textos del domini de les ciències de la salut ha de fer servir estratègies diferents segons els tipus d'unitat a detectar. Aquestes estratègies, que en aquest treball circumscrivim a l'àmbit de la medicina, pressuposen que el mòdul d'extracció de l'extractor utilitzi dos tipus d'eines: un conjunt de **filtres** (lèxics, morfosintàctics i morfosemàntics) i una sèrie de **programes**.

---

carbonado y en sufijos y prefijos que indican los grupos funcionales: metano, etano, 2-aminopropanol, ácido 3-hidroxipropanoico, etc."

[33] En el Corpus textual de l'IULA els noms en llatí queden marcats d'una manera especial en l'etapa del marcatge estructural [Bach i al., 1997].

En el quadre següent sintetitzem les estratègies que un extractor faria servir per detectar cada tipus d'USE i les eines de què se serviria per fer-ho en cada cas:

1. **Detecció d' USE lingüístiques**
    - 1.1 monolèxiques
        - 1.1.1 simples:
            - ☜**diccionari d'USE simples de medicina**
        - 1.1.2 derivades:
            - ☜**diccionari d'USE simples de medicina**
            - ☜**diccionari de formants cultes**
            - ☜**protocol de relació de famílies d'arrels**
        - 1.1.3 compostes patrimonials
            - ☜**diccionari d'USE simples de medicina**
            - ☜**protocol de relació de famílies d'arrels**
        - 1.1.4 compostes cultes
            - ☜**diccionari de formants cultes**
        - 1.1.5 sigles
            - ☜**diccionari de sigles freqüents**
            - ☜**instruccions de detecció de sigles**
    - 1.2 polilèxiques
        - 1.2.1 UTP
            - ☜**diccionari d'USE simples de medicina**
            - ☜**diccionari de formants cultes**
            - ☜**diccionari de quantitatius i organitzadors discursius**
            - ☜**diccionari de paratermes de medicina**
            - ☜**filtre d'estructures**
            - ☜**filtres semanticoformals per a l'estructura NA i N de (art) N**
            - ☜**protocol de relació de famílies d'arrels**
        - 1.2.2 UFE
            - ☜**diccionari d'USE simples de medicina**
            - ☜**diccionari de formants cultes**
            - ☜**diccionari de quantitatius i d'organitzadors discursius**
            - ☜**diccionari de paratermes de medicina**
            - ☜**filtre d'estructures**
            - ☜**filtres semanticoformals per a l'estructura N de (art) N**

☜**eina per extreure freqüències d'ús**

2. **Detecció d'USE no lingüístiques**

2.1 símbols:

☜**diccionari de símbols**

☜**instruccions de detecció de símbols i fórmules**

2.2 noms llatins:

☜**instruccions de detecció de la nomenclatura anatòmica, botànica, zoològica i virològica**

2.6 nomenclatura de la química i fórmules químiques:

☜**diccionari de formants cultes**

☜**diccionari dels símbols**

☜**instruccions de detecció de símbols i fórmules**

5.3 Mòdul de detecció d'un SEACUSE: components

En els apartats anteriors hem establert, primerament, les unitats que un extractor hauria de reconèixer perquè fos suficientment exhaustiu des del punt de vista dels especialistes. Hem analitzat també alguns dels elements formals i semàntics de les USE que podien facilitar el seu reconeixement automàtic a partir dels textos especialitzats, hem proposat estratègies de detecció diferents per a cada un dels tipus d'unitats basades en aquests elements i els hem atribuït recursos i programes perquè ho fessin.

A continuació, intentarem modelitzar el mòdul central d'un SEACUSE tenint en compte les característiques i condicions que li hem atribuït i les estratègies descrites. Aquest mòdul estaria format per dos tipus de components principals:

- un component de filtres, que anomenarem FILTRES
- un component de programes, que anomenarem PROGRAMES.

El component FILTRES tindria com a funció principal la de discriminar entre unitats pertinents i unitats no pertinents a partir de la detecció de característiques diverses. La sortida de cada filtre seria sempre una llista d'unitats considerades pertinents en cada fase de l'aplicació.

El component PROGRAMES es compondria de diversos corpus d'instruccions destinades a associar o detectar directament determinats tipus d'unitats.

5.3.1 Component FILTRES

El component FILTRES estaria integrat per tres subcomponents que operarien sobre nivells d'informació diferents:

a) FILTRE LÈXIC
b) FILTRE MORFOSINTÀCTIC
c) FILTRE MORFOSEMÀNTIC.

a) El submòdul FILTRE LÈXIC operaria sobre les unitats lèxiques i es limitaria a constatar si les unitats del text formen part o no d'un diccionari. Estaria compost per un conjunt de sis diccionaris, quatre dels quals actuarien de filtres positius (servirien per confirmar el caràcter especialitzat d'una unitat) i dos de filtres negatius (es farien servir per saber que una unitat no és especialitzada):

A. Els diccionaris següents actuarien de filtres positius:

1. un diccionari d'USE simples
2. un diccionari de formants cultes
3. un diccionari de sigles freqüents

4. un diccionari de símbols freqüents[34].

B. Els diccionaris que actuarien de filtres negatius serien[35]:

5. un diccionari de paratermes
6. un diccionari de mots que generen soroll (quantitatius i organitzadors de l'estructura).

Una vegada decidit els tipus i nombre de diccionaris que requeriria un SEACUSE, cal preguntar-se com dotar-lo amb aquests filtres. Per fer-ho, tindríem dues possibilitats: fer servir una base lexicosemàntica informàtica ja acabada, com EuroWordNet[36], o crear-los a partir d'altres recursos existents, informatitzats o no.

Per a la llengua catalana, tenint en compte, d'una banda, que encara no disposem de diccionaris o tesaurus automatitzats amb informació sobre la classe conceptual a la qual pertany un mot i, de l'altra, que cap dels diccionaris que proposem seria gaire voluminós[37], una solució a curt termini seria crear-los a partir de recursos ja existents[38].

---

[34] Els quatre primers lèxics haurien de portar informació gramatical i semàntica de la classe conceptual a la qual pertany cada entrada.
[35] Habert i al. (1997) anomenen aquest tipus de filtres lèxics *antidiccionaris*.
[36] EuroWordNet, però, no es pot fer servir immediatament perquè és un projecte en curs d'elaboració i, a més, per utilitzar-lo com a etiquetador semàntic s'hauria d'adaptar.
[37] Dels diccionaris proposats el més extens seria el dels formants cultes que contindria unes 1.100 entrades.
[38] Alguns dels recursos existents en català que podrien facilitar l'elaboració de diccionaris del tipus que necessita un SEACUSE podrien ser els següents:

- diccionari de termes simples ⇒ *Hiperdiccionari* en CD-ROM (1993) i el programa DIGIT de Yzaguirre (1998)
- diccionari de formants cultes ⇒ Bernabeu i al. (1995)
- diccionari de sigles freqüents ⇒ Garrido (1984)
- diccionari de símbols freqüents ⇒ taula periòdica d'elements (IEC), IUPAC (1979), Villarrasa (1984); llistat de les unitats del SI pertinents en medicina OMS (1975)
- diccionari de paratermes ⇒ proposem una llista elaborada a partir del nostre corpus textual que caldria completar

b) El subcomponent FILTRE MORFOSINTÀCTIC serviria perquè un SEACUSE pogués discriminar d'entre les estructures sintagmàtiques, les que podrien correspondre a USE polilèxiques. Estaria compost d'un conjunt de patrons positius i/o un conjunt de patrons negatius, que permetrien fer un primer reconeixement de les unitats sintagmàtiques dels textos especialitzats.

Els patrons morfosintàctics de què disposarà aquest filtre podrien procedir d'estudis lingüístics sistemàtics de la forma dels sintagmes terminològics d'àmbits temàticament restringits[39]. Per a la llengua catalana i en concret per a l'àmbit mèdic, un SEACUSE es pot servir dels patrons proposats a Estopà (1996b).

c) El subcomponent FILTRE MORFOSEMÀNTIC tindria l'objectiu de desambiguar certes combinacions sintagmàtiques que tot i les anàlisis lèxica, morfològica i/o sintàctica poden rebre més d'una interpretació: tant podrien ser UT, com UFE, combinacions especialitzades recurrents o UD. Per assolir aquest objectiu el SEACUSE hauria d'utilitzar els esquemes morfosemàntics de restricció només quan els altres recursos lingüístics fossin insuficients per desambiguar una seqüència polilèxica.

Aquest cas, com ja hem vist, es dóna en dues combinacions: en la combinació $[N[A]_{SAdj}]_{SN}$ en què el nom és especialitzat, però l'adjectiu no ho és (*sonda acanalada* versus *sonda groga*); i en l'estructura $[N\ [de\ (art)\ [N]]_{SPrep}]_{SN}$ en què el primer nom de la seqüència és terminològic (*càncer de mama* versus *radiografia de la mà*; *fòrceps de serra* versus *fòrceps de color platejat*).

- diccionari de mots que sempre generen soroll: quantitatius i organitzadors del discurs ⇒ proposem també una llista elaborada a partir del nostre corpus textual que caldria completar.

[39] Per a algunes llengües (francès, italià, alemany, anglès) aquesta mena d'estudis s'han aplicat a extractors concrets [Bourigault, 1993], [Jacquemin, 1994], [Bordoni i Anzaldi, 1996], [Heid i al., 1996], etc.

En el SEACUSE que proposem, els **esquemes semàntics de combinació de mots** tindrien com a finalitat especificar si certes combinacions de paraules que formalment corresponen a USE, ho són realment; i en cas que ho siguin, aclarir si són UT o col·locacions. Aquesta discriminació reduiria el soroll que genera tot sistema d'extracció automàtica quan només aplica patrons formals o estratègies basades en els aspectes formals rellevants de les unitats en qüestió.

Però per utilitzar automàticament filtres morfosemàntics és necessari que prèviament el corpus textual al qual s'aplica l'extractor estigui etiquetat i desambiguat morfològicament i semànticament. Pel que fa a l'etiquetador i desambiguador morfològic el català compta amb eines competents, però no disposa de sistemes d'etiquetatge ni desambiguació semàntics.

Actualment, les opcions més eficients que podrien resoldre aquesta mancança serien les tres següents:

- utilitzar una classificació conceptual existent sense cap canvi
- adaptar una classificació conceptual existent
- crear un sistema de classes semàntiques nou.

Si l'opció és utilitzar o adaptar un sistema ja dissenyat per ser implementat informàticament, tindríem dues possibilitats: usar un etiquetador general o usar-ne un d'especialitzat. En el primer cas disposem de WordNet [Millner i al., 1990] i d'EuroWordNet [Vossen i al., 1997][40]. En el segon cas de l'Unified Medical Language System (UMLS) [Humphrey i Lindberg, 1989].

[40] De fet, WordNet és, des de principis dels anys noranta, un dels tesaurus electrònics més utilitzats, per bé que EuroWordNet té l'avantatge que serà una base de dades **multilingüe** (en una primera fase per al holandès, italià, castellà i grec, i, posteriorment, per al català, basc, gallec, francès, polonès, entre d'altres llengües) amb xarxes de paraules per diverses llengües. EuroWordNet s'està creant a partir de recursos existents en cada llengua, i aquesta característica, que en un principi és positiva perquè suposa racionalitzar els recursos existents, ocasiona desequilibris lèxics importants entre les àrees de coneixement d'una llengua a una altra.

Aquestes tres bases lèxiques, però, presenten algunes mancances per al nostre objectiu:

- Són parcials pel que fa al tractament de les categories gramaticals: els verbs i els adjectius no hi són representats o hi estan molt pobrament representats[41].
- Els dominis especialitzats no estan desenvolupats en la mateixa profunditat, hi ha unes branques molt més precises i especificades que unes altres.
- WordNet és una base concebuda per a la llengua anglesa[42]; en canvi, UMLS i EuroWordNet estan pensats per a diverses llengües de famílies lingüístiques diferents.
- WordNet i UMLS són projectes, en principi, acabats, que es poden utilitzar. En canvi, EuroWordNet és un tesaurus multilingüe, encara en curs d'elaboració[43].

Presentarem a continuació, al marge de decidir quina estratègia es faria servir per etiquetar semànticament els textos, una primera aproximació de com hauria de ser el subcomponent FILTRE MORFOSEMÀNTIC del component FILTRES del SEACUSE. Aquest subcomponent només

[41] Habert i al. (1997: 115) comenten que els resultats de l'aplicació de tesaurus lèxics informatitzats són parcials perquè "*Souvent, seuls les noms sont pris en compte. Il y a plusieurs raisons à cela. La finalité des analyseurs en permet pas toujours d'exploiter les contextes verbaux. La description lexicale des noms dans un réseau comme WordNet est plus riche et plus structurée —donc plus exploitable— que pour les autres catégories*."

[42] Com remarca Habert i al. (1997:91) per al francès, per bé que les experiències anglosaxones en aquest terreny són força aprofitables sempre s'han d'adaptar perquè existeixen desajustaments culturals importants: "*Si la recherche sur les corpus en français peut sans doute tirer profit de l'expérience anglo-saxonne pour éviter certains tâtonnements, des problèmes spécifiques se posent pour chaque langue, qui imposent certains ajustements, voire la mise au point de méthodes particulières ou le développement d'outils spécifiques.*"

[43] Actualment, la Universitat Politècnica de Catalunya és responsable de la versió castellana i catalana d'EuroWordNet, amb la col·laboració de la Universitat de Barcelona. Paral·lelament, s'ha de tenir en compte que diferents experiències han mostrat el fracàs de reutilitzar la base de coneixements UMLS per etiquetar semànticament un corpus textual [Charlet i al., 1996], [Habert i al., 1997]. Aquests autors creuen que "*l'ontologie d'un domaine dépend d'un point de vue sur ce domaine et de la tache pour laquelle a été conçue: elle n'est donc réutilisable que dans la mesure où la tâche demeure la même, ce qui est exceptionnel*". UMLS es va concebre per necessitats documentals, per indexar textos, i aquesta característica limita en certa manera les possibilitats de reutilitzar-la per a altres finalitats.

s'aplicaria per analitzar els constituents de les tres combinacions lèxiques següents, que són les que després d'utilitzar estratègies formals encara resulten ambigües:

1. $[N_{term}\ [A_{no\ esp}]_{SAdj}]_{SN}$ = UT o UD
2. $[N_{term}$ [de (art) $[N_{no\ esp}]_{SPrep}]_{SN}$ = UT o UD
3. $[N_{term}$ [de (art) $[N_{esp}]_{SPrep}]_{SN}$ = UT o combinació recurrent

Per a cada estructura el SEACUSE activaria estratègies i condicions específiques.

1. **El filtre morfosemàntic 1 s'aplicaria a l'estructura $[N_{term}\ [A_{no\ esp}]_{SAdj}]_{SN}$.** El problema que aquesta estructura planteja a un SEACUSE és múltiple:

a) establir la classe semàntica a què pertany el N_{term}
b) establir la classe semàntica de l'$A_{no\ esp}$ adjectiu no especialitzat
c) saber si la combinació $[N_{term}\ [A_{no\ esp}]_{SAdj}]_{SN}$ dóna lloc a una UT.

Pel que fa a l'anàlisi dels noms que apareixen en aquesta estructura i a fi d'atribuir-los a una classe semàntica, el filtre parteix de la classificació de noms presentada a 3.2.1.1, que és la següent:

- NOM A: malalties, estats o manifestacions patològiques
- NOM B: parts i components dels cos humà
- NOM C: objectes instrumentals

- NOM D: mètodes, proves i tècniques
- NOM E: accions
- NOM F: operacions.

Per determinar els $A_{\text{no esp}}$ pertinents en medicina, però, cal tenir en compte prèviament dos fets. En primer lloc, que cada domini d'especialitat prioritza un grup restringit d'adjectius que serveixen per classificar els noms només en el domini considerat. I en segon lloc, que, dins d'una parcel·la concreta del coneixement especialitzat, no tots els adjectius poden acompanyar tots els noms. Aquests dos fets impliquen que els filtres s'hagin d'aplicar sempre en el context d'un domini específic[44].

Per establir els filtres morfosemàntics ens basem en la classificació dels adjectius pertinents en medicina que hem presentat a 4.2.2 que parteix de les propietats, intrínseques o extrínseques, que els adjectius manifesten respecte del nom que complementen, que és la següent:

[44] En medicina, el segment *febre groga,* per posar un exemple, és un terme pertinent, però *sonda groga* és una combinació discursiva. En aquest cas, l'adjectiu de color *groc* és terminològicament pertinent quan s'aplica al terme *febre* perquè indica un subtipus de febre (*febre aftosa, febre àlgia, febre efímera, febre espamàstica, febre groga, febre herpètica, febre negra, febre quintana, febre romana,* etc.); però, en canvi, no ho és quan *groc* s'aplica a *sonda* perquè no té el poder suficient per classificar al terme *sonda,* com ho fan altres adjectius que indiquen la forma, la funció, el lloc d'aplicació d'aquest instrument (*sonda acanalada, sonda bicolzada, sonda vertebrada*; *sonda flexible; sonda intestinal, sonda lacrimal, sonda uretral*, etc.) i, per tant, la combinació final és una unitat discursiva. *Sonda groga* presenta les mateixes característiques sintaticofuncionals que *cotxe groc* o *cotxe blau*.

Si analitzem amb més detall aquestes seqüències —*febre groga* i *sonda groga*— i les contrastem amb altres de la mateixa naturalesa, arribem a la conclusió que, si en un corpus textual trobem un adjectiu de color que acompanya un instrument mèdic, es tracta d'una combinació discursiva: *agulla groga, bisturí negre, catèter vermell, sonda blava,* etc.; però, en canvi, si en un text trobem un adjectiu de color que acompanya un nom de la classe semàntica de les malalties, les manifestacions o estats patològics, la seqüència final constitueix una unitat terminològica polilèxica: *diarrea verda, escara blanca, pesta blanca, pesta negra, pesta verda, taca negra,* etc. D'aquesta manera, podem formular filtres morfosemàntics com els següents que cal tenir present que no es poden aplicar en abstracte, sinó sempre en el marc d'un text especialitzat:

$[N_{\text{d'instrument}}\ [A_{\text{de color}}]]_{\text{unitat discursiva}}$

$[N_{\text{de malaltia, manifestació o estat patològic}}\ [A_{\text{de color}}]]_{\text{UTP}}$

- ADJ 1: adjectius que indiquen la localització anatòmica del N[45]
- ADJ 2: adjectius que indiquen la causa que ha provocat el N
- ADJ 3: adjectius que indiquen el període de temps en què es realitza el N
- ADJ 4: adjectius que indiquen la freqüència en què es dóna el N
- ADJ 5: adjectius que indiquen els efectes del N
- ADJ 6: adjectius que indiquen la funció del N
- ADJ 7: adjectius que indiquen el mitjà o font energètica amb què funciona el N
- ADJ 8: adjectius que indiquen el funcionament del N
- ADJ 9: adjectius que indiquen elements constitutius del N
- ADJ 10: adjectius que indiquen la matèria primera del N
- ADJ 11: adjectius que indiquen l'agent que ha ocasionat el N
- ADJ 12: adjectius que indiquen l'experimentador del N[46]
- ADJ 13: adjectius que indiquen la quantitat afectada pel N
- ADJ 14: adjectius que indiquen el nivell de gravetat del N
- ADJ 15: adjectius que indiquen propietats perceptibles sensorialment del N:

  ADJ 15.1: color

  ADJ 15.2: textura

  ADJ 15.3: olor

  ADJ 15.4: forma

  ADJ 15.5: dimensió

  ADJ 15.6: altres
- ADJ 16: adjectius que indiquen la relació respecte d'un estàndard.

Un cop establerts els grups semàntics a què pertanyen el nom i l'adjectiu de l'estructura $[N_{term}\ [A_{no\ esp}]_{SAdj}]_{SN}$, el filtre es proposa de saber si cada combinació dóna lloc o no a una UT. Per fer-ho, segueix les condicions següents, organitzades a partir de les classes semàntiques dels noms:

1. Noms del grup A: malalties, estats o manifestacions patològiques

[45]L'etiqueta de localització cal entendre-la en un sentit ampli ja que la informació que del nom transmet l'adjectiu es troba a mig camí entre el que seria un objecte (la malaltia afecta una part del cos humà) i el que seria un lloc (la malaltia es localitza en una part del cos humà determinat).

[46] L'adjectiu experimentador s'ha d'entendre en un sentit ampli, ja que sovint es produeix una sinècdoque i l'experimentador és un nom de col·lectiu humà a qui pot afectar la malaltia o el nom del lloc.

Una seqüència [NOM A [A $_{\text{no esp}}$]$_{\text{SAdj}}$]$_{\text{SN}}$ on N pertany al grup A és una **unitat terminològica** només si el nom està complementat per un adjectiu que pertany als setze grups següents:

- Adjectius del GRUP 1 que indiquen la localització anatòmica del N que complementen[47]: *anterior, central, unilateral, bilateral, inferior, superior, lateral, col·lateral, posterior*[48].

- Adjectius del GRUP 2 que indiquen la causa o el tipus de causa que ha provocat el N: *accidental; adquirit, -ida; controlat, -ada; endogen, -ògena; espontani, -ània; essencial; específic, -a; exogen, -ògena; inespecífic, -a; induït, -ïda, provocat, -ada; simple, sobtat, -ada* [49].

- adjectius del GRUP 3 que indiquen el període de temps en què es realitza el N: *diürn, -a; estival; estiuautommal; matinal; matutí, -ina; nocturn, -a; premenstrual; -òria; primaveral; puerperal; vernal*[50].

---

[47] La majoria d'adjectius que es refereixen a la part del cos humà on es manifesta prioritàriament una malaltia són de caràcter especialitzat (es formen amb un formant culte que es refereix a un component del cos humà i un sufix relacional, que sol ser *-ic; -al; -ar; -í; -i*) i, per tant, no presenten problemes d'ambigüitat perquè el sistema els pot identificar mitjançant el diccionari de formants cultes. Hi ha, però, un grup d'adjectius no especialitzats que també indiquen el lloc anatòmic o la part del cos en què es produeix la malaltia o que s'emmalalteix. En aquests casos, i des del punt de vista del significat, podem dir que s'ha anaforitzat la part anatòmica afectada; per exemple, *escleritis posterior* s'ha d'entendre com "*una escleritis que afecta el pol posterior de l'escleròtica i la caròtida i la tetina adjuntes*" (DEM).

[48] *estafiloma anterior, estafiloma posterior, estafiloma intercalat, fístula externa, fístula interna, otitis externa, otitis interna, otitis mitjana, conjuntivitis unilateral, conjuntivitis bilateral, estrabisme superior,* etc. A vegades, l'adjectiu també pot expressar simplement si la malaltia està localitzada o si, per contra, la seva localització és difusa o fins i tot si la localització no és fixa: *local, difús, -a; errant; ocult, -a (asfíxia local, epilèpsia oculta, nefritis local, nefritis difusa,* etc.)

[49] *enterocolitis simple, erisipela espontània, exantea sobtat, fagocitosi espontània, fagocitosi induïda, hemoglobinúria accidental, hemorràgia provocada, hipermetropia adquirida, immunodeficiència específica, immunodeficiència inespecífica, micosi endògena, micosi exògena, ossificació accidental, sífilis adquirida,* etc.

[50] *catarro estival, conjuntivitis primaveral, edema premenstrual, encefalitis vernal, epilèpsia matinal, epilèpsia nocturna, hidroa estival,* etc. Un subtipus d'aquest grup estaria constituït pels adjectius que expressen el temps en què es produeix o es desenvolupa una malaltia respecte del temps normal en què hauria d'aparèixer o desenvolupar-se: *precoç; prematur; tardà, -ana: alopècia prematura, ascites precoç, epilèpsia tardana, limfodema precoç, periostitis precoç, raquitis tardana, raquitisme tardà,* etc.

- Adjectius del GRUP 4 que indiquen la freqüència en què es dóna el N: *cíclic, -a; crònic, -a; constant; esporàdic, -a; habitual; incessant; intermitent; oscil·lant, permanent; periòdic, -a; persistent; progressiu, -iva; remitent; únic, -a; temporal; temporani, -ània; transitori, -òria*[51].

- Adjectius del GRUP 5 que indiquen els efectes del N. L'únic cas que hem documentat en què l'arrel lèxica de l'adjectiu no està relacionada ni amb un terme simple ni amb un formant grecollatí és ***silenciós, -osa,*** que assenyala l'absència d'efectes; així, una malaltia o estat patològic és *silenciós* quan no presenta manifestacions ni conseqüències[52].

- Adjectius del GRUP 8 que indiquen el funcionament del N: *actiu, -iva; lent, -a; passiu, -iva*[53].

- Adjectius del GRUP 11 que indiquen l'agent que ha ocasionat el N: *alcohòlic, -a; alimentari, ària; climàtic, -a; químic, -a, solar*[54].

- Adjectius del GRUP 12 que indiquen l'experimentador del N: *adult, -a; familiar; infantil; juvenil; presenil; senil; rural; tropical; urbana; professional*[55]. I al costat d'aquests adjectius, també poden expressar indirectament l'experimentador els adjectius gentilicis: *africà, -ana;*

---

[51] *artritis temporal, bogeria transitòria, bogeria regressiva, còlera esporàdic, eritema persistent, esteonosi permanent, estrabisme cíclic, estrabisme periòdic, exoftalmia temporal, faringitis crònica, exoftalmia temporal, febre efímera, miopia progressiva, miositis progressiva, periodontitis crònica, febre remitent*, necrospèrmia progressiva, etc. A cavall d'aquest tipus d'adjectius i dels següents, situem un subtipus d'adjectius no especialitzats que expressen la fase en què es troba una determinada malaltia: *inicial* i *terminal: endocarditis terminal, hematúria inicial, hematúria terminal, pneumònia terminal,* etc.

[52] Els efectes que provoca la malaltia en la gairebé totalitat de les USE polilèxiques s'expressen a través d'adjectius de base terminològica: *catarral; congestiu, -iva; depressiu, -iva; dolorós, -osa; espàstic, -a; febril; flegmonós, -osa; granulós, -osa; hemorràgic, -a; irritatiu, -iva; mortal; mucós, -a; papulós, -osa; petequial; quístic, -a, supurat, -ada,* etc.

[53] *endocarditis lenta, hemorràgia activa, hemorràgia passiva, trombosi passiva,* etc.

[54] *edema alimentari, icterícia solar, meningitis química, nefritis solar, neuritis alcohòlica,* etc. Tot i que la majoria dels adjectius que expressen l'agent d'una malaltia o manifestació patològica són especialitzats i no presenten problemes d'ambigüitat perquè es detecten mitjançant el diccionari de formants clàssics.

[55] *esclerosi presenil, demència senil, demència juvenil, periodontitis juvenil, dermatitis industrial, malaltia rural, paràlisi infantil, melanoma juvenil, nanisme infantil, nanisme senil, nefrosi familiar, neurosi militar, neurosi professional,* etc.

americà, -ana; asiàtic, -a; brasiler, -a; centreuropeu, -ea; cubà, -ana; egipci, -ípcia; japonès, -esa; rus, -usa; etc[56].

- Adjectius del GRUP 13 que indiquen la quantitat afectada pel N: *absolut, -a; alterna, complet, -a; doble, focal; generalitzat, -ada; multifocal, parcial, total*[57].

- Adjectius del GRUP 14 que indiquen el nivell de gravetat del N: *advers; agut, -uda; ascendent; benigne, -a; descendent; estable; favorable; greu*[58]*; intens, -a; lleuger, -a; lleu; major; maligne, -a; menor; moderat, -ada; nociu, -iva; primari, -ària; profund, -a; progressiu, -iva; pulsatiu, -iva; secundari, -ària; superficial*[59].

- Adjectius del GRUP 15 que indiquen propietats perceptibles sensorialment del N. Hi ha un grup d'adjectius qualificatius de caràcter no especialitzat que expressen propietats perceptibles sensorialment i intrínseques del nom que acompanyen:

    - ADJ15.1: COLOR: *blanc, -a; blau, -ava; bru, -una; gris, -a; groc, -oga; negre, -a;* ***pàl·lid, -a;*** *roig, -oja;* ***rosat, -ada****; verd, -a; vermell, -a*[60].
    - ADJ15.2: TEXTURA: *cremós, -a; gelatinós, -osa*[61].

---

[56] *disenteria japonesa, febre espanyola, febre tropical, hemoptisi oriental, leishmaniosi americana, leishmaniosi brasilera, miïtis tropical, meningitis africana, tripanosomiasi gambiana,* etc.

[57] *ascites parcial, epilèpsia focal, epilèpsia generalitzada, estrabisme absolut, fibromatosi generalitzada, fístula completa, fístula simple, fístula complexa, flegmó total, hipermetropia total, lipodistrofia total, necrospèrmia focal, pneumònia multifocal, pneumònia parcial* , etc.

[58] Hem documentat en el DEM el terme *hemorràgia molt greu* amb la introducció de l'adverbi *molt* entre el nom i l'adjectiu, estructura que és molt poc habitual.

[59] c*òlera fulminant, encefalitis secundària, epilèpsia major, epilèpsia menor, febre secundària, febre primària, faringitis agudaedema maligne, edema benigne, hemorràgia greu, mesenquimoma benigne, mesenquimoma maligne, miastèmia greu,, otitis aguda, periostitis aguda,* etc. Les dades obtingudes mostren que aquests adjectius són adjectius qualificatius en oposició gradual: *maligne/benigne; greu/moderada/lleu, major/menor, primari/secundari.*

[60] *asfíxia blava, asfíxia blanca, atròfia banca, atròfia grisa, atròfia roja, atròfia negra, edema blau, febre negra, febre vermella, granulació grisa, icterícia negra, loqui blanc, loqui vermell,* etc.

- ADJ15.3: OLOR: *fètid, -a; purulent, -a; pútrid, -a*[62].
- ADJ15.4: FORMA: *anular; cilíndric, -a; circular; multiforme; pla, -ana; rodó, -ona; romboïdal* [63].
- ADJ15.6: ALTRES: calent, -a; fred, -a; humit, -ida; irritable; opac, -a; sec, -a; tou, -ova.[64]

- Adjectius del GRUP 16 que indiquen una relació respecte un estàndard: *artificial; anormal; atípic, -a; atòpic, -a; comú, -una; estrany, -a; fals, -a; normal; patològic, -a; típic, -a; tòpic, -a; vulgar* [65].

2. Noms dels grup B: parts i components del cos humà

Un NOM B només dóna lloc a una unitat terminològica si l'adjectiu de caràcter no especialitzat pertany a un dels grups següents[66]:

- Adjectius del GRUP 1 que indiquen la localització anatòmica del N: *anterior; ascendent; bilateral; central; descendent; distal; dorsal; dret, -a; errant; esquerre, -a; extern, -a; frontal; global; inferior; intern, -a; lateral; local; medià, -ana; medial; mitjà, -ana; posterior; primer, -a; proximal; segon, -a; superior; supí, -ina; transvers, -a; transversal; unilateral*[67].

---

[61] *estomatitis cremosa, pus cremós,* etc.
[62] *abscés purulent, bronquitis fètida, meningitis purulenta, miositis purulenta, peritonitis pútrida,* etc.
[63] *eritema multiforme, hemangioma pla, pitiriasi rodona, úlcera circular,* etc.
[64] *enema opac, estenosi irritable, faringitis seca, fibroma tou, hemoaglutinació calenta, hemoaglutinació freda, nòdul calent, nòdul fred, necrospèrmia humida, pàpula seca, pàpula humida, tos seca, tumor fred,* etc.
[65] *dermatitis artificial, dermatitis atòpica, erupció anòmala, estenosi falsa, màcula comuna, melanosi falsa, migranya comú, pitiriasi vulgar,* etc.
[66] En la resta de combinacions el resultat d'unir un NOM B amb un adjectiu no especialitzat és una UD sense interès terminològic.
[67] *fibra endògena, fibra exògena, ventricle lateral, ventricle medià, fibra transversal, fibra intermèdia,* etc. Tot i que, normalment, aquests adjectius es col·loquen en segona posició després dels adjectius de localització anatòmica: *artèria espinal anterior, artèria espinal posterior, artèria pulmonar dreta, artèria pulmonar esquerra, os maxil·lar inferior, os maxil·lar superior, mesoderma lateral. menisc intern, menisc extern,* etc. Els adjectius de posició solen presentar-se en parelles contrastives que ajuden a distingir parts simètriques que només es diferencien pel lloc que ocupen en el cos humà.

- Adjectius del GRUP 6 que indiquen la funció del N: *accelerador, -a; assassí, -ina; defensiu, -iva; depressor, -a; motor, -a, nerviós, -osa; obstructiu, -iva; obturador, -a; regulador, -a; supressor, -a*[68]

- Adjectius del GRUP 10 que indiquen la matèria primera del N: *escumós, -osa; espinós, -osa; gras, -assa; nucleat, -ada*[69].

- Adjectius del GRUP 15.1 que indiquen el color del N: *blanc, -a; bronzejat, -ada; gris, -a; magent, -a; vermell, -a*[70].

- Adjectius del GRUP 15.2 que indiquen la textura del N: *llis, -a*[71].

- Adjectius del GRUP 15.4 que indiquen la forma del N: *acanalat, -ada; anular; ciatiforme; cilíndric, -a; caliciforme; estrellat, -ada; falciforme; foliat, -ada; geminat, -ada; lobulat, -ada; oval; pla -ana; poligonal; quadrat, -ada; rodó, -ona; romboide; solcadat, -ada*[72].

- Adjectius del GRUP 15.5 que indiquen la dimensió del N: *mitjà, -ana; major; menor; curt, -a; llarg, -a; gros, -ossa; petit, -a; prim, -a; gegant, -a; simple; complex, -a*[73].

Cal tenir present, però, que la localització sol indicar-se mitjançant un adjectiu integrat per un NOM A i un sufix relacional (*artèria espinal, artèria pulmonar*, etc.).

[68] *escorça motora, estímul nerviós, fibra inhibidora, fibra depressora, gen regulador, gen supressor, hèrnia obturadora, hidrocefàlia obstructiva, múscul obturador, nervi obturador, nervi motor, neurona motora, proteïna defensiva,* etc. L'examen de les UTP del corpus mostra que aquesta propietat també es pot expressar amb adjectius especialitzats: *olfactori, -òria; òptic, -a; respiratori, -òria; secretor, -ora; sensitiu, -iva; tàctil, visceral,* etc.

[69] *fetge gras, fibra nucleada, os espinós,* etc.

[70] Les dades dels corpus constaten que els colors *negre, verd* i *blau* que podien acompanyar els NOMS A, no són pertinents combinats amb els NOMS B, però aquesta restricció és pragmàtica, de tal manera que si no existeix cap component del cos humà verd, en un text especialitzat no trobarem aquesta combinació lèxica. Alguns exemples d'unitats terminològiques d'aquest grup són: *fibra groga, fibra grisa, fibra fosca, fibra clara, fibra blanca, fibrocartíleg groc, glòbul vermell,* etc.

[71] *llengua llisa, múscul llis*, etc.

[72] *articulació plana, cèl·lula poligonal, gangli estrellat, limbe angulós, lòbul quadrat, múscul quadrat, pelvis oval, pelvis plana, pelvis rodona, pelvis triangular, tòrax pla, úlcera circular,* etc.

[73] *cèl·lula gegant, hipocamp menor, hipocamp major, os ample, pelvis ampla, pelvis gegant, pelvis petita,* etc.

- Adjectius del GRUP 15.6 que indiquen altres propietats del N: *elàstic; esponjós; madur, -a; tou, -ova*[74].

3. Noms dels grup C: objectes i instruments

Un Nom del grup C serà el nucli d'una UTP amb estructura [N[$A_{no\ esp}$]$_{SAdj}$]$_{SN}$[75], si el modifica un adjectiu dels tipus següents:

- Adjectius del GRUP 1 que indiquen la localització anatòmica on s'aplica el N: *alt, anterior, baix, extern, intern, errant, posterior*[76].

- Adjectius del GRUP 3 que indiquen el període de temps en què es fa servir el N: *definitiu, -iva; permanent; provisional; temporal*[77].

- Adjectius del GRUP 6 que indiquen la funció del N: *comparatiu, -iva; enregistrador, -ora; excavador, -a; explorador, -a; diferencial*; *protector, -a; tallant*[78].

- Adjectius del GRUP 7 que indiquen el mitjà o la font energètica amb què funciona el N: *elèctric, -a; electrodinàmic, -a; electrònic, -a; protònic, -a; nuclear, ultrasònic, -a*[79].

---

[74] *fibra elàstica, fibrocartíleg elàstic, fibrocartíleg esponjós,òvul madur, pelvis elàstica, pelvis tova*, etc.

[75] Cal remarcar que en els corpus no hem trobat massa NOMS C amb aquesta estructura perquè la majoria es categoritzen amb un N_{propi} introduït per una preposició *de* que indica el científic que ha inventat l'objecte; per exemple, el DEM (1990) inclou 75 tipus de pinces, 60 de les quals presenten una estructura eponímica.

[76] Són esporàdics els casos en què un ADJ1 de caràcter no especialitzat complementa un NOM C i indica el lloc per on s'aplica o instal·la: *marcapàs extern, marcapàs intern, marcapàs errant, fòrceps alt, fòrceps baix, fòrceps anterior, fòrceps posterior.* Els dos últims exemples s'han de llegir com el "*fòrceps que s'aplica per damunt de l'estret superior*" i el que s'aplica "*sobre el cap fetal situat entre els plans II i IV de Hodge*".

[77] Aquest tipus de combinació es dóna sobretot en instruments que s'instal·len dins del cos humà: *marcapàs permanent, marcapàs provisional, sonda permanent,*etc.

[78] *electroscopi diferencial, microscopi comparatiu, termòmetre enregistrador,* etc. Tot i que en aquests casos se sol utilitzar més un complement preposicional: *d'aplanament, d'adsorbció, de partició*, etc.

[79] *balança electrodinàmica, bisturí elèctric, electroscopi electrònic, oftalmoscopi elèctric, marcapàs nuclear, microscopi electrònic, microscopi ultrasònic,* etc.

- Adjectius del GRUP 8 que indiquen el funcionament del N: *aperiòdic; -a; automàtic, -a; logarítmic, -a; lineal; rotatori, òria*[80].

- Adjectius del GRUP 9 que indiquen elements constitutius del N: *bifocal; binocular; centrífug*[81]*; compost, -a; monoplat; trifocal*[82].

- Adjectius del GRUP 10 que indiquen la matèria primera del N: *metàl·lic, -a; salí, -ina*[83].

- Adjectius del GRUP 15.4 que indiquen la forma del N: *acanalat, -ada; bicolzadat, -ada; canalicular; colzat, -ada; cònic, -a; corb; -a; olivari, -ària; prisat, -ada; recte, -a; triangular*[84].

- Adjectius del GRUP 15.6 que indiquen altres propietats del N: *adhesiu, -iva; elàstic, -a, fix, -a; flexible; opac, -a; vertebrat, -ada*[85].

4. Noms del grup D: mètodes, proves i tècniques

Un nom del grup D amb estructura $[N[A_{no\ esp}]_{SAdj}]_{SN}$ és terminològic si l'adjectiu de caràcter no especialitzat pertany a un del grups següents[86]:

- Adjectius del GRUP 1 que indiquen la localització anatòmica on s'aplica el N: *extern, -a; intern, -a; lateral; superficial*[87].

[80] *balança aperiòdica, balança automàtica, microscopi rotatori, oxigenador rotatori,* etc.
[81] Es refereix a un instrument que conté una centrifugadora.
[82] *balança monoplat,* m*icroscopi centrífug, microscopi binocular, oftalmoscopi binocular*, etc.
[83] *termòmetre salí, termòmetre metàl·lic, termòmetre bimetàl·lic,* etc.
[84] *bena triangular, catèter bicolzat, filtre prisat, pinces triangulars, sonda acanalada, sonda bicolzada, sonda colzada, tisores corbes, tisores rectes,* etc.
[85] *bena adhesiva, bena elàstica, microscopi opac, sonda flexible, sonda vertebrada,* etc.
[86] Els mètodes i tècniques que serveixen per explorar, observar, diagnosticar, tractar, curar i operar, normalment, es denominen amb el nom del mètode, prova o tècnica, complementat amb un sintagma preposicional introduït per la preposició *de* i l'antropònim del científic que ha inventat el mètode en qüestió, com passava amb els noms dels instruments.
[87] *palpació superficial, pelvimetria externa, pelvimetria interna,* etc., encara que el lloc del cos humà on s'aplica el mètode o tècnica s'expressa majoritàriament mitjançant adjectius

- Adjectius del GRUP 4 que indiquen la freqüència en què es realitza el N: *periòdic, -ca; progressiu, -iva;*

- Adjectius del GRUP 5 que indiquen els tipus de resultats en què presenta el N: *bidimensional; encreuat, -ada; lineal; múltiple; panoràmic, -a; seriat, -ada; unidimensional*[88].

- Adjectius del GRUP 6 que indiquen la funció del N: *comparat, -ada; destructiu, -iva; diferencial; electiu, -iva; funcional*[89].

- Adjectius del GRUP 7 que indiquen el mitjà o la font energètica amb què funciona el N: *bimanual; electrònic, -a; factorial; gasós, -a; instrumental; líquid, -a; magnètic, -a; neutrònic, -a; químic, -a; radiològic, -a; tèrmic, -a* [90].

- Adjectius del GRUP 8 que indiquen el funcionament del N: *automàtic, -a; ascendent; descendent; desequilibrat, -ada; directe, -a; doble; equilibrat, -ada; indirecte, -a; mòbil; negatiu, -iva; objectiu, -iva; positiu, -iva*[91].

5. Noms del grup E: accions

especialitzats (Noms A + sufix relacional). Aquesta és la raó per la qual hem documentat pocs adjectius no anatòmics que indiquen el lloc on es practica una determinada tècnica.

[88] *anàlisi encreuada, ecografia bidimensional, radiografia panoràmica,* etc.

[89] *anàlisi funcional, anàlisi comparada, pelvimetria electiva,* etc. Els adjectius documentats d'aquest grup són escassos perquè aquesta característica s'indica majoritàriament amb adjectius especialitzats: *analític, -a; clínic, -a; colorimètric, -a; cromatogràfic, -a; demogràfic, -a; densimètric, -a; gravimètric, -a; mètric, -a; potenciomètric, -a; quirúrgic, -a; topogràfic, -a; volumètric, -a.*

[90] *anàlisi química, anàlisi tèrmica, encefalografia líquida, encefalografia gasosa, micrografia electrònics, pelvimetria radiològica, pelvimetria ultrasònica, pelvimetria instrumental, radiografia neutrònica,* etc.

[91] *anàlisi equilibrada, audiometria objectiva, auscultació directa, auscultació indirecta, oftalmoscòpia directa, oftalmoscòpia indirecta, urografia ascendent, urografia descendent,* etc.

Quan el nucli d'una UTP amb estructura [N$_{deverbal}$[A$_{no\ esp}$]**$_{SAdj}$**]**$_{SN}$** és un nom del grup E (normalment un terme deverbal), l'adjectiu de caràcter no especialitzat pertany als tipus següents:

- Adjectius del GRUP 1 que indiquen la localització anatòmica on es realitza el N: *anterior, col·lateral; encreuat, -ada; extern, -a; extrínsec, -a; frontal; intern, -a; intrínsec, -a; lateral; posterior, superficial*[92].

- Adjectius del GRUP 2 que indiquen la causa o el tipus de causa que ha provocat el N: *accidental; espontani, -ània; controlat, -ada; mixta; oportunista; sobtat, -ada; violent, -a; voluntari, -ària*[93].

- Adjectius del GRUP 6 que indiquen la funció del N: *assimilador, -ora; excitant; expressiu, -iva; funcional; protector, -a; preservatiu, -iva*[94].

- Adjectius del GRUP 7 que indiquen la font energètica amb què funciona el N: *elèctric, -a; electromagnètic,-a; manual; nuclear; químic, -a; tèrmic, -a*[95].

- Adjectius del GRUP 8 que indiquen el funcionament del N: *accelerat, -ada; adequat, -ada; afuncional; anormal; artificial; dinàmic, -a; explosiu, -iva; habitual; natural; passiu, -iva; ràpid, -a*[96].

- Adjectius del GRUP 13 que indiquen la quantitat afectada pel N: *absolut, -a; binari, -ària; doble; múltiple; parcial; relatiu, -iva; segmentari*[97].

---

[92] *fixació interna, producció endògena, producció exògena, oclusió anterior, oclusió posterior, oclusió lateral,* etc.
[93] *atenció voluntària, deshidratació voluntària, producció accidental,* etc.
[94] *coloració protectora, conducció regenerativa, inducció assimiladora, injecció excitant, injecció preservativa, injecció sensibilitzant,* etc.
[95] *excitació nuclear, inducció electromagnètica, inversió elèctrica, inversió química, inversió tèrmica, tracció manual,* etc.
[96] *conducció accelerada, comulsió habitual, hivernació artificial, nutrició adequada, oclusió habitual, oclusió anormal,* etc.
[97] *convulsió parcial, hospitalització parcial, visió doble, visió múltiple,* etc.

- Adjectius del GRUP 15.1 que indiquen el color de la part afectada pel N: *gris, -a; groc, -oga; vermell, -a, roig, -ja; negre, -a, verd, -a*[98].

6. Noms del grup E: accions

Un nom del grup F és una UT si està subcategoritzat pels següents tipus d'adjectius no especialitzats[99]:

- Adjectius del GRUP 1 que indiquen la localització anatòmica on s'aplica el N: *anterior; central; extern, -a; frontal, inferior; intern, -a; lateral, medià, -ana; perifèric, -a; posterior; subcostal; superior*[100].

- Adjectius del GRUP 6 que indiquen la funció del N: *explorador, -a; preparatori, -òria; refractiu, -iva; substitiu, -iva*[101].

- Adjectius del GRUP 7 que indiquen la font energètica amb què es realitza el N: *químic, -a; nuclear*[102].

- Adjectius del GRUP 13 que indiquen la quantitat afectada pel N: *completa; doble; parcial; sectorial; simple; radical; total*[103].

---

[98] *hepatització grisa, hepatització groga, hepatització vermella, inducció gris, inducció roja, inducció negra, visió blava, visió groga, visió verda,* etc.

[99] Normalment, però, les operacions se solen denominar o bé mitjançant compostos cultes (*ectomia, -stomia, -plastia*) o bé amb la nominalització d'un verb al qual s'afegeix el sufix *-ció* o *-ment* (*ablació, amputació, extirpació, implantació, intervenció, operació, secció, trasplantació, trasplantament*).

[100] *amputació transversal, faringotomia externa, faringotomia interna, frageotomia superior, frageotomia inferior, implantació central, indectomia perifèrica, laringotomia inferior, laringotomia superior, litotomia lateral, lobotomia frontal, talamectomia anterior, titotomia bilateral, urotrotomia externa, urotrotemia interna,* etc.

[101] *laparatomia exploradora, iridectomia preparatòria, punció exploradora, queratoplàstia refractiva, transfusió substitutiva,* etc.

[102] *simpactectomia química, trasplantació nuclear,* etc.

[103] *amputació absoluta, iridectomia sectorial, laringotomia completa, laringotomia parcial, laringotomia total, mastectomia simple, mastectomia total, piloplàstica doble, queratoplàstia total, transposició parcial, vagotomia completa,* etc.

- Adjectius del GRUP 15.4 que indiquen la forma en què es realitza el N: *circular; lineal; obliqua; transversal*[104].

Com a síntesi, presentem un quadre resum de les classes de noms terminològics i d'adjectius de caràcter no especialitzat que combinats donen com a resultat una UTP pertinent en ciències de la salut:

[104] *amputació circular, incisió oblíqua, incisió transversal, secció lineal, secció longitudinal, secció sagital,* etc.

QUADRE RESUM DE LES UTP AMB ESTRUCTURA $[N_{term}\ [A_{no\ esp}]SAdj]_{SN=\ UTP}$

	NOMS A	NOMS B	NOMS C	NOMS D	NOMS E	NOMS F
ADJ1	X	X	X	X	X	X
ADJ2	X	--	--	--	X	--
ADJ3	X	--	X	--	--	--
ADJ4	X	--	--	X	--	--
ADJ5	X	--	--	X	--	--
ADJ6	--	X	X	X	X	X
ADJ7	--	--	X	X	X	X
ADJ8	X	--	X	X	X	--
ADJ9	--	--	X	--	--	--
ADJ10	--	X	X	--	--	--
ADJ11	X	--	--	--	--	--
ADJ12	X	--	--	--	--	--
ADJ13	X	--	--	--	X	X
ADJ14	X	--	--	--	--	--
ADJ15.1	X	X	--	--	X	--
ADJ15.2	X	--	--	--	--	--
ADJ15.3	X	--	--	--	--	--
ADJ15.4	X	X	X	--	--	X
ADJ15.5	--	X	--	--	--	--
ADJ15.6	X	--	X	--	--	--
ADJ16	X	--	--	--	--	--

A continuació i a mode de resum, presentem la llista dels filtres morfosemàntics relatius a l'estructura $[N_{term}\ [A_{no\ esp}]_{SAdj}]_{SN}$ que serveixen per restringir les UTP pertinents en medicina:

A. **Noms A:**

$[NOM\ A\ [ADJ1]_{SAdj}]_{SN=UTP}$
$[NOM\ A[ADJ2]_{SAdj}]_{SN\ =\ UTP}$
$[NOM\ A[ADJ3]_{SAdj}]_{\ SN\ =\ UTP}$
$[NOM\ A[ADJ4]_{SAdj}]_{\ SN\ =\ UTP}$
$[NOM\ A[ADJ5]_{SAdj}]_{\ SN\ =\ UTP}$
$[NOM\ A[ADJ8]_{SAdj}]_{\ SN\ =\ UTP}$
$[NOM\ A[ADJ11]_{SAdj}]_{\ SN\ =\ UTP}$

[NOM A[ADJ12]$_{SAdj}$] $_{SN}$ = UTP
[NOM A[ADJ13]$_{SAdj}$] $_{SN}$ = UTP
[NOM A[ADJ14]$_{SAdj}$] $_{SN}$ = UTP
[NOM A[ADJ15.1]$_{SAdj}$] $_{SN}$ = UTP
[NOM A[ADJ15.2]$_{SAdj}$] $_{SN}$ = UTP
[NOM A[ADJ15.3]$_{SAdj}$] $_{SN}$ = UTP
[NOM A[ADJ15.4]$_{SAdj}$] $_{SN}$ = UTP
[NOM A[ADJ15.6]$_{SAdj}$] $_{SN}$ = UTP
[NOM A[ADJ16]$_{SAdj}$] $_{SN}$ = UTP

B. **Noms B:**

[NOM B [ADJ1]$_{SAdj}$] $_{SN}$ = UTP
[NOM B [ADJ6]$_{SAdj}$] $_{SN}$ = UTP
[NOM B [ADJ10]$_{SAdj}$] $_{SN}$ = UTP
[NOM B [ADJ15.1]$_{SAdj}$] $_{SN}$ = UTP
[NOM B [ADJ15.5]$_{SAdj}$] $_{SN}$ = UTP
[NOM B [ADJ15.4]$_{SAdj}$] $_{SN}$ = UTP

C. **Noms C:**

[NOM C [ADJ1]$_{SAdj}$]$_{SN}$ = UTP
[NOM C [ADJ3]$_{SAdj}$] $_{SN}$ = UTP
[NOM C [ADJ6]$_{SAdj}$] $_{SN}$ = UTP
[NOM C [ADJ7]$_{SAdj}$] $_{SN}$ = UTP
[NOM C [ADJ8]$_{SAdj}$] $_{SN}$ = UTP
[NOM C [ADJ9]$_{SAdj}$] $_{SN}$ = UTP
[NOM C [ADJ10]$_{SAdj}$] $_{SN}$ = UTP
[NOM C [ADJ15.4]$_{SAdj}$] $_{SN}$ = UTP
[NOM C [ADJ15.6]$_{SAdj}$] $_{SN}$ = UTP

D. **Noms D:**

[NOM D [ADJ1]$_{SAdj}$] $_{SN}$ = UTP
[NOM D [ADJ4]$_{SAdj}$] $_{SN}$ = UTP
[NOM D [ADJ6]$_{SAdj}$] $_{SN}$ = UTP
[NOM D [ADJ7]$_{SAdj}$] $_{SN}$ = UTP
[NOM D [ADJ8]$_{SAdj}$] $_{SN}$ = UTP

E. **Noms E:**

[NOM E [ADJ1]$_{SAdj}$] $_{SN}$ = UTP
[NOM E [ADJ2]$_{SAdj}$] $_{SN}$ = UTP
[NOM E [ADJ6]$_{SAdj}$] $_{SN}$ = UTP
[NOM E [ADJ7]$_{SAdj}$] $_{SN}$ = UTP

[NOM E [ADJ8]$_{SAdj}$] $_{SN}$ = UTP
[NOM E [ADJ13]$_{SAdj}$] $_{SN}$ = UTP
[NOM E [ADJ15.1]$_{SAdj}$] $_{SN}$ = UTP

F. **Noms F:**

[NOM F [ADJ1]$_{SAdj}$] $_{SN}$ = UTP
[NOM F [ADJ6]$_{SAdj}$] $_{SN}$ = UTP
[NOM F [ADJ7]$_{SAdj}$] $_{SN}$ = UTP
[NOM F [ADJ13]$_{SAdj}$] $_{SN}$ = UTP
[NOM F [ADJ15.4]$_{SAdj}$] $_{SN}$ = UTP

2. **El filtre morfosemàntic 2 s'aplicaria a l'estructura [N1$_{term}$ [de (art) [N2$_{no\ esp}$]$_{SPrep}$]$_{SN}$.** Per resoldre l'ambigüïtat d'aquesta estructura un SEACUSE ha de fer les següents operacions en aquest ordre:

a) atribuir classe semàntica al N1$_{term}$
b) indicar la relació semàntica que el N2 manté amb el N1
c) analitzar si la combinació [N$_{term}$ [de (art) [N$_{no\ esp}$]$_{SPrep}$] $_{SN}$ resulta una UT o una UD.

Per atribuir la classe semàntica el filtre fa servir la classificació següent:

- NOM A: malalties, estats o manifestacions patològiques
- NOM B: parts i components dels cos humà
- NOM C: objectes instrumentals

Per indicar el tipus de relació que el N2 manté amb el N1, el filtre es val de la llista de relacions següents:

- RELACIÓ A: el N2 indica l'experimentador[105] del N1
- RELACIÓ B: el N2 indica l'agent del N1

[105] Aquest experimentador s'ha d'entendre en el sentit ampli que hem comentat anteriorment: el nom d'un lloc o estat serveix també per referir-se a les persones que preferentment poden patir la malaltia.

- RELACIÓ C: el N2 indica la forma metafòrica del N1
- RELACIÓ D: el N2 indica el material amb què està construït N1
- RELACIÓ E: el N2 indica la funció del N1

D'acord amb la classificació de N1 i N2, constaten que l'estructura [N_{term} [de (art) [$N_{no\ term}$] $_{SN}$]$_{SPrep}$]$_{SN}$ només és terminològica quan es donen les condicions morfosemàntiques següents:

a) Quan el N1 pertany al GRUP A i s'estableix la RELACIÓ A entre els dos noms:

> *erisipela de la costa, febre de les trinxeres, eritema dels nadons, escoliosi dels escolars, esterilitat de la parella, esterilitat de l'adolescència, tifus dels botiguers.*

De vegades, la mateixa relació que manté el complement amb el nucli també es pot expressar amb una estructura lleugerament diferent, sense que el complement estigui determinat:

> *anòxia dels nadons, febre de pantans, sordesa de calderer.*

b) Quan el N1 pertany al GRUP A i s'estableix la RELACIÓ B entre els dos noms:

> *tifus de les paparres, tifus de la paparra de Kenya, tifus de les puces, tifus de les rates, tifus dels ratolins.*

c) Quan el N1 pertany al GRUP B i s'estableix la RELACIÓ C entre els dos noms:

ungla de raqueta, crani d'alforja, crani de campanar, crani de trèvol, crani de torre, escrot de xal, eritròcit de diana, fàcies de follet, fàcies d'ocell, fàcies de plat, fetge d'acordió, fetge de sucre de candi, etc.

d) Quan el N1 pertany al GRUP C i s'estableix la RELACIÓ C entre els dos noms:

filtre de plecs, espècul de bec d'ànec.

e) Quan el N1 pertany al GRUP C i s'estableix la RELACIÓ D entre els dos noms:

escut de plom, fèrula de filferro, filtre de vidre.

f) Finalment, quan el N1 pertany al GRUP C i s'estableix la RELACIÓ E entre els dos noms[106]:

embut d'addició, embut de filtració, embut de separació, filtre de pressió, filtre de premsa, filtre de centrifugació, fòrceps de tracció.

A continuació i a mode de resum agrupem de manera esquemàtica els filtres morfosemàntics que restringeixen les seqüències amb estructura [N_{term} [de (art) [$N_{no\ term}$]]$_{SPrep}$]$_{SN}$ que resulten UTP, en la resta de combinacions que presentin aquesta mateixa estructura, la unitat resultant és una UD sense interès especialitzat:

A. [N_{term} [de (art) [$N_{no\ term}$]$_{SN}$]$_{SPrep}$]$_{SN}$ = UTP

[106] En aquests casos hem constatat que el sintagma preposicional sempre és indeterminat i que el nom que expressa la funció de l'objecte que modifica, normalment, és un nom deverbal construït amb el sufix *-ció.*

[N1 GRUP A [de [N2 RELACIÓ A] SN]SPrep]SN = UTP
[N1 GRUP A [de [N2 RELACIÓ B] SN]SPrep]SN = UTP
[N1 GRUP B [de [N2 RELACIÓ C] SN]SPrep]SN = UTP
[N1 GRUP C [de [N2 RELACIÓ C] SN]SPrep]SN = UTP
[N1 GRUP C [de [N2 RELACIÓ D] SN]SPrep]SN = UTP
[N1 GRUP C [(art) [N2RELACIÓ E]] SN]SPrep]SN = UTP

3. **El filtre morfosemàntic 3 s'aplicaria a l'estructura [$N1_{term}$ [de (art) [$N2_{term}$]$_{SPrep}$]$_{SN}$.** Per resoldre l'ambigüïtat d'aquesta estructura, com en el cas del filtre morfosemàntic 2, un SEACUSE ha de fer les següents operacions en aquest ordre:

a) atribuir classe semàntica al $N1_{term}$

b) indicar la relació semàntica que el N2 manté amb el N1

c) saber si la combinació morfosemàntica [N_{term} [de (art) [N_{esp}]$_{SPrep}$] SN dóna lloc a una UT o a una combinació especialitzada recurrent.

Per atribuir la classe semàntica el filtre fa servir la classificació següent:

- NOM A: malalties, estats o manifestacions patològiques
- NOM B: parts i components dels cos humà

Per indicar el tipus de relació que el N2 manté amb el N1, el filtre es val de la llista de relacions següents:

- RELACIÓ D: el N2 indica la matèria de què està formada el N1
- RELACIÓ E: el N2 indica a funció del N1
- RELACIÓ F: el N2 indica la causa que ha provocat el N1
- RELACIÓ G: el N2 indica la part afectada pel N1

L'estructura [N_{term} [de (art) [N_{term}]$_{SN}$]$_{SPrep}$]$_{SN}$ és terminològica quan es donen les combinacions següents[107]:

a) El N1 pertany al GRUP A i s'estableix la RELACIÓ F entre els dos noms:

   sordesa de transmissió, sordesa de conducció, úlcera per distensió, úlcera d'estrès, diprea d'esforç, edema de fam, etc.

b) El N1 pertany al GRUP A i s'estableix la RELACIÓ G entre els dos noms[108]:

   càncer de mama, càncer de colon, espasme de glotis,etc.

c) El N1 pertany al GRUP B i s'estableix la RELACIÓ E entre els dos noms:

   enzim de restricció, fibra d'associació, tromba d'aglutinació, etc.

d) El N1 pertany al GRUP B i s'estableix la RELACIÓ G entre els dos noms:

   rodet del cos callós, septe del bulb, eix del cristal·lí, espina del pubis, escut del cor, falç del cervell, falç del cerebel, etc.

e) Finalment, també dóna com a resultat una UTP, quan N1 pertany al GRUP B i s'estableix la RELACIÓ D. En el corpus només hem documentat una unitat d'aquest tipus: *retícula de fibrina*.

[107] En la resta de possibilitats la unitat resultant sempre és una col·locació especialitzada que, per a determinades tasques professionals, pot tenir interès terminològic, però que d'entrada cal poder discriminar.

A mode de síntesi presentem els filtres morfosemàntics que restringeixen les seqüències amb estructures $[N_{term}$ [de (art) $[N_{term}]]_{SPrep}]_{SN}$ que resulten UTP:

B. $[N_{term}$ [de (art) $[N_{term}]_{SN}]_{SPrep}]_{SN}$ = UTP

[N1 GRUP A [de [N2 RELACIÓ F] SN]SPrep]SN = UTP
[N1 GRUP A [de [N2 RELACIÓ G] SN]SPrep]SN = UTP
[N1 GRUP B [de [N2 RELACIÓ E] SN]SPrep]SN = UTP
[N1 GRUP B [de [N2 RELACIÓ G]SPrep]SN = UTP
[N1 GRUP B [de [art [N2 RELACIÓ D]] SN]SPrep]SN = UTP

5.3.2 Component PROGRAMES

El component PROGRAMES està constituït per una sèrie de conjunts d'ordres que guiïn i facilitin la detecció de les USE. Aquests programes es poden classificar en dos tipus: d'una banda un programa-protocol que es basa principalment en característiques morfològiques de les USE lingüístiques i, de l'altra, un conjunt de programes-instrucció que es fonamenten en condicions tipogràfiques i discursives per a uns tipus d'unitats molt específiques que presenten unes característiques gràfiques peculiars. Així, el component PROGRAMES estaria integrat per dos subcomponents:

1. Subcomponent A:

    - protocol de relació dels mots que comparteixen la mateixa arrel

2. Subcomponent B[109]:

[108] Aquesta combinació no és molt productiva perquè per expressar la mateixa relació, hem vist anteriorment, que es prefereix l'estructura $[N[A]_{SAdj}]_{SN}$.

[109] De fet aquestes instruccions han de ser molt simples perquè si l'extractor s'aplica sobre els textos del Corpus textual de l'IULA, en l'etapa de marcatge estructural i en el preprocés aquestes unitats ja es detecten automàticament, per bé que no es distingeixen les generals de les especialitzades [Bach i al, 1997].

- instruccions de detecció de sigles
- instruccions de detecció de símbols i fórmules
- instruccions de detecció de noms llatins.

El protocol per relacionar els mots que comparteixen la mateixa arrel, tant si és una arrel grecollatina com catalana, és de gran utilitat i es faria servir per reconèixer i desambiguar diferents tipus d'USE[110].

Quant al programa d'instruccions per a sigles, en l'apartat 5.2.4, hem proposat un conjunt d'heurístiques per detectar-les, basades sobretot en els aspectes tipogràfics i distribucionals, que es podria perfilar a través de l'observació del comportament de les sigles en altres corpus textuals sobre altres temes mèdics. CSIC (1987), OMS (1975), Manuila (1975), Matthews (1982), Peterson (1987), Villarrasa (1984) són obres que fan referència a les regles de formació i escriptura dels símbols i les fórmules químiques.

Finalment, les principals fonts per elaborar les instruccions per detectar els noms llatins de les nomenclatures biològiques i de la *Nomina Anatomica* són les següents: Jeffrey (1976), Lapage (1975), Matthews (1982). En aquestes obres, com hem comentat en l'apartat 5.2.5, es donen directrius de formació i d'utilització dels noms llatins que integren les diverses nomenclatures de la biologia.

[110] Existeixen altres sistemes amb objectius molt diferents que utilitzen també un programa d'aquesta mena. Per exemple, SEXTANT [Grefenstette, 1994] és un programari que ajuda a construir automàticament entrades de diccionari a través d'esquemes sintàctics i lèxics. Per a cada entrada, aquest programa facilita dades quantitatives, la llista de noms veïns, la llista de verbs operadors, la llista d'expressions (els contextos) i la llista de variants ortogràfiques i morfològiques. Totes les relacions paradigmàtiques es calculen mitjançant el grau de similitud entre dos mots. SEXTANT és un programa que s'aplica a la llengua anglesa, però que podria servir de model per construir un programa que agrupés les famílies de paraules per al català perquè és de base estadística. En canvi, un programa de base lingüística es podria crear a partir d'un diccionari derivacional basat en regles de formació de paraules, com el que dirigeix per al francès Corbin (1997).

A més del component FILTRES i del component PROGRAMES, el mòdul de detecció d'un SEACUSE podria ser més eficient si comptés amb un component EINES integrat per una eina per extreure concordances i freqüències d'ús de les estructures pertinents.

5.4 Conclusions

La finalitat d'aquest capítol ha estat proposar elements de descripció i estratègies per tal que un SEACAT augmentés el grau de precisió i d'exhaustivitat. Aquest objectiu implicava bàsicament quatre fases:

1. Establir les unitats dels textos d'especialitat que són pertinents i que han de ser objecte d'un extractor: les USE.

2. Descriure els elements més significatius de cada USE per a la seva extracció automàtica.

3. Proposar les estratègies que hauria d'utilitzar un extractor per extreure només les USE pertinents.

4. Plantejar els components, basats en les estratègies proposades, que haurien d'integrar el mòdul de detecció d'un SEACUSE.

Quant al primer objectiu, hem arribat a la conclusió que calia eixamplar l'objecte d'extracció dels SEACAT clàssics perquè, en el textos especialitzats, hi ha moltes altres unitats a més de les UTP i de les UT que també són pertinents des del punt de vista especialitzat. La descripció de les unitats de significació especialitzada (USE) que hem dut a terme no tenia la finalitat de ser exhaustiva ni completa, sinó només de posar en relleu els elements d'aquestes unitats que podien servir per fer-ne l'extracció automàtica.

Pel que fa a la proposta d'estratègies d'extracció, hem assumit la idea que un SEACUSE no pot servir-se d'una única estratègia per detectar les USE pertinents dels textos (recursos lèxics, morfològics, morfosintàctics, morfosemàntics, tipogràfics, distribucionals i estadístics), sinó que s'han d'aplicar diferents estratègies segons les característiques internes de cada tipus d'unitat.

En el mateix sentit, hem defensat que un SEACUSE tampoc no pot reduir-se a utilitzar estratègies que es basin només en la **forma** de les USE, sobretot pel que fa a la recuperació de les USE polilèxiques. Per això, per a algunes seqüències polilèxiques que ocasionen ambigüitat hem proposat **filtres morfosemàntics** basats en esquemes que tenen en compte tant l'estructura d'una seqüència com la classe semàntica dels constituents que integren aquesta seqüència.

La finalitat principal dels filtres morfosemàntics és desfer l'ambigüitat dels segments ambigus que, estructuralment, donen lloc a dos tipus d'unitats diferents. Així doncs, com més perfilats i complets siguin els filtres morfosemàntics, menys ambigüitat hi haurà i el SEACUSE generarà menys soroll. Hem proposat aplicar filtres morfosemàntics en tres ocasions:

a) en l'estructura $[N_{term}\ [A_{no\ esp}]_{SAdj}]_{SN}$, en què els filtres morfosemàntics ajuden a decidir si és una UT o una UD
b) en l'estructura $[N_{term}$ [de (art) $[N_{term}]_{SPrep}]_{SN}$, per tal com les restriccions de semàntica lèxica desambigüen si la unitat és terminològica o recursiva
c) en l'estructura $[N_{term}$ [de (art) $[N_{no\ term}]_{SPrep}]_{SN}$, en què els filtres morfosemàntics permeten saber si la unitat resultant és terminològica o discursiva.

Per poder utilitzar els filtres morfosemàntics, però, cal etiquetar semànticament els corpus textuals i hem vist que no hi havia etiquetaris

semàntics per al català ni per al castellà. Un tesaurus lèxic com EuroWordNet o com UML podria ser útil, però per elaborar un SEACUSE caldria adaptar-los. Tot i que en el domini de les ciències de la salut, els diccionaris que hem proposat d'utilitzar, sobretot el de formants cultes, serien suficients per construir una aplicació a curt termini.

Conseqüentment, hem establert que el mòdul de detecció d'un SEACUSE que assolís un nivell de satisfacció òptim i detectés les USE pertinents hauria d'estar integrat, bàsicament, per dos components: un component FILTRES i un component PROGRAMES. El primer estaria format per filtres lèxics, morfosintàctics i morfosemàntics; i el segon, inclouria un protocol de detecció d'unitats que comparteixen la mateixa base i un conjunt d'instruccions centrades en aspectes de disposició discursiva, de tipografia i de freqüència d'ús de les unitats.

Finalment, al costat dels components FILTRES i PROGRAMES hem proposat una eina complementària que, opcionalment, facilitaria les freqüències d'ús de les USE i les seves concordances.

5.5 Recapitulació

En aquest capítol primer hem analitzat l'objecte d'extracció d'un buidatge automàtic, és a dir, els tipus d'unitats de significació especialitzada dels textos especialitzats en medicina que són pertinents i, per tant, desitjables de reconèixer de manera automàtica. En aquest sentit, hem mostrat la diferència que existeix entre l'objecte d'extracció dels SEACAT clàssics i el del sistema que proposem i d'aquesta manera hem passat del concepte de SEACAT al de SEACUSE.

A continuació, hem proposat els elements i les estratègies que hauria d'utilitzar un SEACUSE per detectar les USE pertinents i provocar el mínim de soroll i de silenci.

Finalment, hem presentat els components que hauria d'integrar el mòdul de detecció d'un extractor, posant èmfasi en la diversitat d'eines i la discriminació dels seus usos.

En el proper capítol, ens proposem analitzar la utilitat real d'un SEACUSE que treballi en abstracte, sense tenir en compte les finalitats del buidatge automàtic ni l'aplicació dels seus resultats. Ens preguntarem si les USE són les mateixes per a tots els usuaris o si, contràriament, segons les finalitats professionals seran pertinents determinats tipus d'unitats i no d'altres.

6. ACTIVITATS PROFESSIONALS I UNITATS DE SIGNIFICACIÓ ESPECIALITZADA

> Il y a une idée trop répandue selon laquelle la terminologie ne serait qu'une lexicographie des domaines spéciaux, notamment technoscientifiques. Cette idée est fondée sur une situation historique particulière, non sur une réalité générale.
>
> [Rey, 1992: 52]

> Sortosament, la terminologia ha començat a ser percebuda avui en dia com una matèria multipolar, polièdrica i multifuncional, i, doncs, com un camp d'estudi d'un interès inusitat fins fa poc.
>
> [Cabré, 1998b: 11]

En el capítol anterior hem posat en qüestió que l'única unitat d'interès d'un text especialitzat fos la unitat terminològica polilèxica (UTP) i, per això, en contrast amb els sistemes clàssics d'extracció de termes, hem ampliat substancialment l'objecte d'extracció.

En aquest capítol volem mostrar com les unitats de significació especialitzada (USE) varien qualitativament i quantitativament en relació amb les necessitats professionals. Partim, doncs, dels supòsits que segons els **interessos dels diferents professionals**:

a) **no totes les USE** que comprèn un text **són pertinents**.
b) **els tipus d'USE pertinents canvien.**

Com a conseqüència d'aquests dos supòsits, les USE (lingüístiques i no lingüístiques) d'un text que interessen un especialista, un traductor, un terminògraf o un documentalista poden no coincidir totalment.

Aquesta idea ha estat plantejada des del punt de vista del tractament del llenguatge natural per Naulleau (1998) en introduir els conceptes de

sintagma nominal pertinent i *sintagma nominal no pertinent*. De manera indirecta, Bourigault (1994, 95) suggereix la mateixa idea quan afirma que el soroll que genera el seu sistema d'extracció de termes LEXTER s'ha de valorar de manera positiva perquè d'aquesta manera és l'usuari qui escull el conjunt de termes que més l'interessa.

La diferència entre aquests dos autors és, però, rellevant. Mentre que el primer concep la noció de pertinència d'un segment com un element previ a l'aplicació del sistema (és a dir, el sistema no pot aplicar-se si l'usuari no facilita el conjunt de sintagmes pertinents i no pertinents); el segon, en canvi, introdueix aquesta noció a posteriori, com a argument per justificar el soroll que produeix el seu sistema (és a dir, després que el sistema ha generat una llista de candidats a terme molt àmplia i heterogènia, de manera que cada usuari ha d'escollir manualment els candidats més pertinents segons les seves necessitats). Però malgrat que els fonaments i les estratègies d'aquests dos sistemes siguin diferents, els punts febles principals que presenten són els mateixos: el temps que comporta realitzar una selecció manual i la subjectivitat amb què es fa.

En efecte, en el sistema de Naulleau, l'usuari esmerça molt temps proporcionant al sistema el conjunt de sintagmes pertinents i no pertinents segons els seus criteris (que poden variar), i aquesta selecció s'ha de fer cada vegada que s'utilitza el sistema. El plantejament de base del sistema que proposa Naulleau, doncs, condiciona subjectivament els resultats finals.

LEXTER també presenta problemes de temps i de subjectivitat, tot i que aquests sorgeixen una vegada el sistema ja ha intervingut. D'una banda, el programa genera una llista de candidats sense cap mena de filtre funcional, cosa que suposa un treball manual de selecció considerable, sobretot si tenim en compte que, normalment, els índexs de soroll es mouen

entre el 40% i el 75% de les unitats proposades pel sistema; de l'altra, encara que davant d'un text determinat el sistema sempre generi la mateixa llista, els usuaris elaboren les seves seleccions a partir dels seus criteris particulars, que també corren el risc de ser subjectius.

Així, un sistema d'extracció automàtica de candidats a unitats de significació especialitzada (SEACUSE) que no tingui en compte el punt de vista de l'usuari —com ho fan la majoria de SEACAT—, aplicat a un text concret produeix sempre la mateixa selecció de termes. Ara bé, si es vol que els sistemes d'extracció puguin realitzar seleccions qualitativament i quantitativament adequades a les necessitats professionals dels diferents col·lectius d'usuaris, s'han de poder perfilar prèviament aquestes necessitats.

Tenint en compte els elements presentats fins ara, els objectius d'aquest capítol són, d'una banda, validar la hipòtesi que les finalitats professionals condicionen la concepció de pertinència d'una USE i, de l'altra, arribar a configurar perfils amplis de necessitats terminològiques, de tal manera que un sistema d'extracció automàtica pugui generar una llista selectiva dels candidats a USE segons finalitats professionals específiques.

6.1 Activitats professionals relacionades amb la terminologia

Per confirmar la hipòtesi que col·lectius professionals diferents quan realitzen una determinada activitat professional s'aproximen als textos especialitzats amb interessos diferents, per tal com tenen punts de vista distints sobre les USE pertinents d'un text especialitzat, hem realitzat la prova experimental següent:

Hem donat el mateix text a diferents tipus de professionals demanant-los que fessin el buidatge de les unitats especialitzades pertinents.

Per fer la prova, hem seleccionat quatre col·lectius professionals relacionats amb els textos especialitzats:

- especialistes
- documentalistes
- traductors especialitzats
- lingüistes/terminògrafs.

El buidatge del text ha estat realitzat per tres especialistes de cada un dels col·lectius professionals proposats; això significa que en total hem analitzat dotze buidatges[1].

D'acord amb la hipòtesi que és la **finalitat professional,** i no el col·lectiu, l'element pertinent que condiciona la selecció de les USE d'un text, hem restringit les activitats professionals a una única finalitat diferent per a cada col·lectiu, i no hem tingut en compte que un mateix professional podia realitzar diverses activitats a partir d'un text d'especialitat. En aquest sentit, ens hem decantat per les activitats professionals més específiques de cada col·lectiu:

- la transmissió del coneixement (metges)
- la indexació de textos (documentalistes)
- la traducció (traductors especialitzats)
- l'elaboració de terminologies (terminògrafs).

[1] En l'annex 3 presentem la llista dels professionals que han col·laborat en aquest experiment.

Per poder perfilar les necessitats terminològiques prototípiques de cada col·lectiu, hem partit d'una qüestió general: ***Quines són les USE d'un text que interessa cada grup d'aquests professionals?*** I l'hem adaptada a cada col·lectiu:

> ***Quines són les USE pertinents d'un text per als especialistes quan transmeten coneixement especialitzat?***
>
> ***Quines són les USE pertinents d'un text per als documentalistes quan indexen un text especialitzat?***
>
> ***Quines són les USE pertinents d'un text per als traductors quan tradueixen un text especialitzat?***
>
> ***Quines són les USE pertinents d'un text per als terminògrafs quan elaboren un vocabulari especialitzat?***

6.1.1 Transmissió del coneixement: metges

És ja un lloc comú afirmar que la terminologia és la base de la comunicació entre especialistes i que els especialistes se serveixen de terminologia per a dues activitats diferents: representar el coneixement especialitzat i transferir-lo. Aquestes activitats recullen —segons Cabré (1992)— les dues grans funcions de la terminologia: la representativa i la comunicativa. Ambdues funcions es donen en els textos especialitzats i es vertebren a través d'USE.

Els especialistes amb una perspectiva més general sobre el tema de l'àmbit del text que analitzem —les malalties infeccioses— són els metges.[2] El text ha estat buidat per tres metges, que presenten les característiques professionals següents:

> **Metge1 (LB):** llicenciat en Medicina i professor en una facultat de medicina.
> **Metge 2 (TV):** llicenciat en Medicina i professor en una facultat de medicina.
> **Metge 3 (PH):** llicenciat en Medicina i metge professional, especialista en medicina interna d'un centre privat.

Aquests tres informants, tot i ser homogenis des del punt de vista de la titulació, presenten algunes diferències relacionades bàsicament amb l'especialitat i amb la pràctica de la medicina[3].

6.1.2 Indexació de textos especialitzats: documentalistes

Els documentalistes, tant per a les tasques d'indexació de textos com per a la recuperació d'informació es valen també d'USE, perquè són les unitats lingüístiques que expressen, de manera més comprimida, el coneixement especialitzat que transmet un text.

---

[2] Hi ha d'altres professionals que també són especialistes en ciències de la salut: infermers, biòlegs, químics, farmacèutics, auxiliars de clínica, etc., que no els hem tingut en compte en aquest treball.

[3] La **formació complementària:** LB és doctor en medicina i es dedica habitualment a la recerca de la història de la medicina. TV està fent la tesi doctoral en el marc de la neurofisiologia i PH es dedica a la pràctica habitual de la medicina.
L'**experiència professional** en la pràctica de la medicina també és un factor que els diferencia: mentre que, actualment, ni LB ni TV es dediquen a la medicina assistencial sinó a la recerca mèdica i a la docència de la medicina, PH treballa de metge especialista en medicina interna en un centre de medicina.

Els especialistes en documentació que indexen un text sintetitzen el seu contingut informatiu a través de mots clau o descriptors, que solen ser termes, normalment controlats per tesaurus.

Els tres documentalistes que han realitzat l'exercici experimental responen als currículums següents:

> **Documentalista 1 (MR):** documentalista d'un centre universitari de recerca i professora en una facultat de biblioteconomia.
> **Documentalista 2 (MC):** documentalista d'un servei de documentació d'un consorci d'hospitals.
> **Documentalista 3 (FV):** documentalista d'un col·legi oficial de veterinaris i postgraduat en *Informació i Documentació.*

Tots tres es caracteritzen pel fet de ser **documentalistes de titulació que han treballat amb documentació mèdica,** i es diferencien per la formació complementària que han rebut i l'experiència en el món professional de la documentació especialitzada[4].

6.1.3 Traducció de textos especialitzats: traductors especialitzats

Els traductors especialitzats que actuen de mediadors entre els especialistes i els destinataris d'un text mantenen, com els especialistes, una doble relació amb les USE. D'una banda, els serveixen per accedir al vessant cognitiu del tema del text que tradueixen; de l'altra, els faciliten l'activitat traductora.

---

[4] La **formació complementària**: a més de documentalistes, MR està cursant la llicenciatura d'informàtica, MC és llicenciada en filologia catalana i FV és llicenciat en veterinària.
L'**experiència laboral**: si bé la primera documentalista només ha treballat amb textos mèdics de manera puntual, els altres dos documentalistes treballen habitualment en l'ordenació, indexació i recuperació de documentació mèdica.

Els tres traductors que han col·laborat en aquest projecte són professionals que habitualment realitzen traduccions especialitzades sobre temes mèdics:

> **Traductor especialitzat 1 (XB):** tècnic lingüístic del Servei Català de la Salut
> **Traductor especialitzat 2 (XM):** metge i traductor especialitzat de textos mèdics, i professor de traducció científica en una facultat de traducció.
> **Traductor especialitzat 3 (JD):** professora universitària de traducció i traductora professional de textos mèdics.

Els tres informants comparteixen el fet **de ser traductors de textos mèdics i de no ser llicenciats en Traducció i Interpretació.** Ara bé, es donen unes variables que fan que cada informant respongui a un subperfil diferent:

- la titulació d'origen
- el textos de traducció
- les llengües de traducció.

La titulació d'origen és la característica que més diferencia aquests tres traductors: XB és llicenciat en filologia catalana i s'ha format en el marc de la **planificació lingüística;** XM és llicenciat en **medicina** (per tant, els aspectes cognitius dels textos especialitzats no li són un obstacle); i JD és llicenciada en **lingüística** i especialista en morfologia lèxica.

Pel que fa al **tipus de textos que tradueixen,** XB és mediador lingüístic i, entre d'altres activitats, tradueix textos relacionats amb el Servei Català de la Salut. Aquests textos s'adrecen majoritàriament a professionals de

l'àmbit mèdic i tenen un format molt fixat: butlletins d'informació farmacològica, protocols, informes, etc. i sobretot textos de caràcter institucional que edita el mateix servei com ara memòries, plans de salut, revistes d'informació adreçades a usuaris i personal de l'àmbit de la sanitat, etc. Els altres dos traductors, en canvi, tradueixen articles científics i manuals especialitzats.

Finalment, les **llengües de traducció** també és un factor de divergència: XB tradueix i corregeix textos del castellà al català i viceversa, XM tradueix de l'anglès al castellà i al català i JD tradueix textos del castellà i del català a l'anglès.

Observem que cap d'aquests traductors és llicenciat en Traducció i Interpretació, però, en canvi, responen als tres tipus de traductor especialitzat més habituals[5]:

- traductor-especialista (XM)
- traductor-lingüista (JD)
- traductor-assessor lingüístic (XB).

Tot i que el fet que cada informant provingui d'una formació diferent pugui diversificar les necessitats professionals, en aquest treball hem optat per no tenir-ho en compte perquè, independentment de la seva formació

[5] En primer lloc, s'ha de tenir en compte que la llicenciatura en Traducció i Interpretació a l'Estat espanyol data de 1989, i, d'altra banda, el professional de la traducció a casa nostra sovint ha estat una figura polivalent, moltes vegades autodidacte, i, en alguns casos, la traducció era considerada una activitat complementària a la seva professió. És un fet que moltes de les traduccions literàries han estat fetes per escriptors cèlebres i que les traduccions cientificotècniques solen estar elaborades per especialistes en el tema que dominen molt bé les llengües estrangeres.
Cal afegir que la situació lingüística del català ha propiciat la figura del tècnic lingüístic que actua de mediador lingüístic i, entre altres tasques, tradueix textos del castellà al català. Pensem en la creació relativament recent dels serveis lingüístics en els diferents centres de l'administració pública, de la sanitat pública, dels jutjats, de les universitats, etc.

d'origen, hem partit de la tasca professional concreta que realitzen: **la traducció de textos especialitzats**.

6.1.4 Elaboració de terminologies: terminògrafs

Els lingüistes que exerceixen de terminògrafs recopilen els termes a partir dels textos especialitzats amb la finalitat d'elaborar un recull lèxic o de crear una base de dades terminològiques que serveix per confeccionar diferents tipus de diccionaris.

Per dur a terme els objectius d'aquest capítol, hem comptat amb tres terminògrafs de les característiques següents:

> **Terminògraf 1 (SM):** responsable de l'àrea de terminologia d'un servei lingüístic universitari.
> **Terminògraf 2 (AE):** responsable de l'àrea de terminologia d'un servei lingüístic universitari.
> **Terminògraf 3 (CG):** professora universitària de lexicografia i terminògrafa.

Dels quatre col·lectius participants en aquest exercici, els dels terminògrafs és el grup més homogeni, atès que comparteixen:

- la titulació: són llicenciats en Filologia Catalana
- la formació complementària: han fet cursos de formació de terminologia
- l'activitat terminogràfica: han elaborat o coordinat diccionaris especialitzats
- el lloc de treball: treballen o han treballat en el marc de la normalització de la llengua catalana.

Tot i així, presenten també algunes diferències, entre les quals podem destacar les dues següents:

1. L'**experiència professional** en relació amb el marc on s'ubica el servei lingüístic en què treballen o han treballat, el temps d'experiència i el tipus de treball terminològic que elaboren.

2. La **formació de postgrau**: només dos dels informants han seguit cursos de tercer cicle en terminologia.

Fins aquí hem presentat les característiques dels professionals que han participat en la prova i hem remarcat que, com a conseqüència de l'assumpció que el criteri que preval en la selecció de les USE d'un text és la finalitat professional, l'element que agrupa els membres de cada col·lectiu és el fet de compartir una mateixa activitat professional[6].

Som conscients que tres informadors no són representatius de tota una col·lectivitat professional. En aquest sentit, les dades quantitatives que extraiem dels seus buidatges difícilment poden ser generalitzables, però sí que són **qualitativament indicatives** del tipus d'USE pertinents per a cada activitat professional i aquest és l'objectiu de l'exercici proposat.

6.2 El text de buidatge

El text que hem utilitzat per al buidatge de les USE és el mateix que vam utilitzar en el capítol segon, tercer i quart d'aquest treball:

[6] No obstant això, és possible que en etapes posteriors sorgeixi la necessitat de desglossar algun d'aquests col·lectius: per exemple, el dels traductors especialitzats ja que, analitzant la realitat socioprofessional, ja s'intueix diversos perfils de traductors.

> *Malalties produï des per Rickettsia* —que forma part del capítol sobre *Malalties infeccioses* que a la vegada forma part del llibre *Medicina interna*[7].

Recordem breument que es tracta d'un document extret d'un manual especialitzat, escrit per metges, adreçat a metges o a estudiants de medicina i integrat per 12.069 ocurrències. Una característica d'aquest text, que no hem assenyalat anteriorment —i que creiem que en aquest moment és significatiu que ens hi referim—, és que reflecteix la **interdisciplinarietat de la medicina.**

Només analitzant el títol del text seleccionat per al nostre estudi —***Malalties*** *produï des per* **Rickketsia**[8]—, observem que no es tracta d'un text que parli de coneixements *exclusius* de medicina (en un sentit molt restrictiu del terme *medicina*). El text transmet coneixements de molts àmbits del saber (biologia, farmacologia, química, zoologia, estadística, etc.) relacionats entre si a través d'un nus de coneixement central: *una malaltia infecciosa*. Ja havíem comentat en el capítol tercer que la medicina, com totes les ciències humanes i socials, és una matèria interdisciplinària i aquest text n'és una mostra. Però aquesta peculiaritat del text, com veurem més endavant, ha ocasionat problemes de selecció de les USE a diferents professionals.

Per realitzar el buidatge terminològic dels textos de la manera més neutra possible, hem donat als professionals la consigna següent:

[7][Farreras i Rozman, 1997].
[8] Els subratllats són nostres.

D'aquest text, marca totes les unitats especialitzades que consideris *significatives i pertinents segons els teus objectius professionals*.

És a dir:

als metges:
"marca totes les unitats especialitzades del text que consideris que transmeten un concepte pertinent en medicina."

als documentalistes:
"marca totes les unitats especialitzades del text que faries servir per indexar-lo."

als traductors especialitzats:
"marca totes les unitats especialitzades del text que destacaries abans d'abordar-ne la traducció."

als terminògrafs:
"marca totes les unitats especialitzades del text que inclouries en un diccionari especialitzat de medicina."

És cert que aquestes instruccions poden semblar poc explícites perquè no s'especifica què s'entén per *unitats especialitzades*, però la nostra voluntat ha estat justament no condicionar el subjecte d'experimentació per tal de no generar-li idees prefixades d'allò que havia d'assenyalar i intentar que hi hagués els mínims biaixos possibles en els resultats.

6.3 Objectius dels buidatges

La prova experimental que hem proposat pretén demostrar que els supòsits següents són adequats:

a) Les USE (i per tant també el termes) es modelen en funció dels textos i de les necessitats funcionals dels seus destinataris.

b) No totes les USE que hi ha en un text són pertinents per a totes les activitats professionals.

c) La noció d'USE pertinent depèn de l'activitat professional que es realitzi, la qual cosa pressuposa que, des d'un punt de vista funcional, les USE d'un text no estan prefixades.

d) En un text especialitzat hi hauria USE:

pertinents **per a tots** els col·lectius
pertinents només **per a alguns** col·lectius
pertinents només **per a un** dels col·lectius.

Així mateix, l'anàlisi de totes les USE permetrà elaborar perfils d'interessos terminològics amplis per a activitats concretes.

6.4 Resultats dels buidatges: anàlisi quantitativa

Per facilitar l'anàlisi dels resultats dels buidatges fets pels diferents professionals, hem agrupat les dades en els quatre conjunts següents:

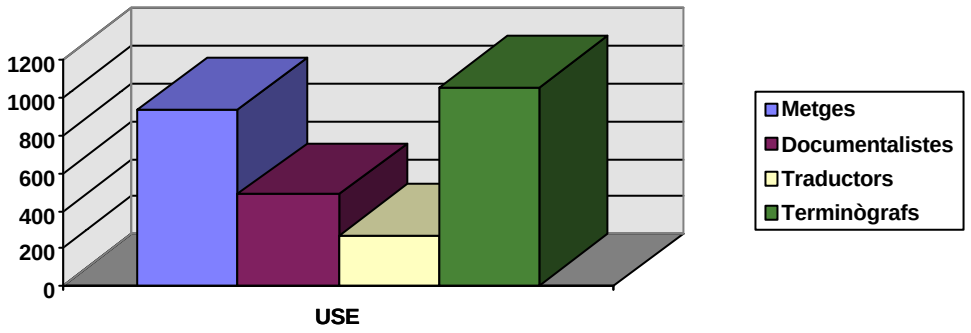
a) USE seleccionades per cada col·lectiu professional
b) USE comunes a tots els col·lectius professionals
c) USE pertinents només per a un col·lectiu professional
d) USE que comparteixen només alguns dels col·lectius professionals.

A continuació presentem els resultats dels buidatges en diversos quadres que mostren quantitativament les tries realitzades pels col·lectius d'informadors. Paral·lelament, en l'annex 2, llistem les ocurrències concretes de les USE seleccionades.

6.4.1 USE marcades per *cada* col·lectiu professional

Per apreciar millor les tendències de cada col·lectiu, hem fusionat les dades dels diferents informants com es pot observar en les taules següents:

	metges		**documentalistes**		**traductors**[9]		**terminògrafs**	
	Nombre	%	Nombre	%	Nombre	%	Nombre	%
noms	824	87,84	426	87,65	211	78,14	900	85,56
verbs	17	1,81	0	0	1	0,37	49	4,65
adjectius	28	2,98	5	1,02	27	10	35	3,32
adverbis	5	0,53	0	0	2	0,74	4	0,37
sigles	13	1,38	12	2,46	5	1,85	12	1,13
símbols	6	0,63	3	0,61	0	0	6	0,56
noms científics	44	4,69	22	4,52	0	0	45	4,27
noms propis	0	0	18	3,70	1	0,37	0	0
frases discursives	1	0,10	0	0	23	8,51	1	0,09
total	**938**	**100**	**486**	**100**	**270**	**100**	**1052**	**100**



Nombre total d'USE seleccionades per cada col·lectiu professional

[9] Un cas especial és el dels traductors perquè els resultats de T1 s'acosten més als dels terminògrafs que als traductors. Aquest fet s'explica per l'entorn i els objectius del treball: T1 treballa, com els terminògrafs seleccionats, en l'àmbit de la normalització de la llengua catalana. Per això, hem cregut oportú no tenir en compte els resultats d'aquest traductor en la valoració general dels buidatges dels traductors especialitzats.

Les dades globals dels quatre col·lectius professionals reforcen la idea que cada col·lectiu té un criteri propi de selecció d'USE pertinents d'un text, i aquesta diversificació de criteris comporta, com es pot veure en la taula, una diversitat en el nombre d'USE seleccionades i en el tipus d'USE prioritzades.

En efecte, els resultats obtinguts constaten que **el nombre d'USE seleccionades no coincideix** ni entre col·lectius (diversitat externa) ni entre els membres d'un mateix col·lectiu (diversitat interna)**.** Aquesta constatació verifica el supòsit que, des del punt vista funcional, les USE (i per tant també els termes) d'un àmbit o d'un objecte temàtic no estan preestablertes, sinó que varien d'acord amb les **necessitats professionals.** Això significa, d'una banda, que en un mateix text hi ha mots que **només són pertinents** per a un col·lectiu d'acord amb l'*activitat professional* que dur a terme.

Una altra característica general que es desprèn de l'anàlisi global dels buidatges és la manca de coincidència entre col·lectius en **els tipus d'USE marcades.** En aquest sentit, constatem una gamma de possibilitats que va des de considerar només les USE nominals (com és el cas d'alguns documentalistes) a seleccionar tots els tipus d'USE presentades en el capítol anterior (les seleccions dels metges). En general, els terminògrafs, seguits dels especialistes, són els que més unitats han marcat i més variades, quant a l'estructura i a la categoria gramatical. Els documentalistes i els traductors, en canvi, són els que n'han marcat menys[10].

Segurament que en un futur s'hauria de considerar autònomament el perfil dels traductors-assessors en contextos de planificació lingüística.

[10] Hem de dir que si bé imaginàvem que els documentalistes serien els que marcarien menys unitats del text, pensàvem que els traductors serien els que en marcarien més. Però no havíem tingut en compte la diferència que el traductor fa entre les USE que conté el text i les USE pertinents perquè presenten problemes de traducció. I aquestes últimes són, sortosament, moltes menys que les primeres.

Pel que fa als aspectes quantitatius dels buidatges dels metges, podem dir que aquests han marcat moltes unitats, encara que no sempre de manera sistemàtica. Sembla que en l'aplicació dels criteris de selecció, tendeixen a ser molt sistemàtics quant a les UT, les USE adverbials, les sigles, els símbols, els noms científics que formen part de nomenclatures consensuades i les UT amb un nucli deverbal; no ho són tant, en canvi, pel que fa a les UFE formades per combinacions molt freqüents de dues unitats terminològiques com *tractament de + una malaltia*, *radiografia de + art + part anatòmica*; i són molt poc sistemàtics quant a les USE adjectives i verbals, tot i que cal que diguem que n'assenyalen algunes. Finalment, cal remarcar que no marquen UFE verbals ni adverbials, ni noms propis aï llats.

Al marge de l'heterogeneï tat en la selecció d'algunes unitats, els tipus d'USE que els especialistes consideren pertinents, perquè són unitats que transmeten coneixement especialitzat, són les següents:

- USE nominals, és a dir termes
- USE verbals
- USE adjectivals
- USE adverbials
- noms científics
- símbols
- sigles
- UFE nominals.

Pel que fa als documentalistes, quantitativament han marcat molt poques USE del text en relació amb altres informadors i els tipus d'unitats seleccionades es redueixen als següents:

- unitats monolèxiques nominals (42,91%)

- unitats polilèxiques nominals (55,85%)
- adjectius (0,41%)
- sigles (2,46%)
- símbols (0,61%)
- noms científics (però només els noms de microorganismes) (4,51%)
- noms propis (3,69%).

No consideren pertinents ni els adverbis ni els verbs ni les UFE. A diferència de la resta d'informadors, dos dels documentalistes han marcat com a unitats vàlides per indexar un text **noms propis**. Aquests noms fan referència a països on es donen les condicions que provoquen una malaltia determinada, a escoles de medicina determinades o a metges cèlebres, i faciliten la delimitació precisa de les cerques potencials.

La freqüència d'ús de les USE pertinents i la seva ubicació en el text són dues dades molt rellevants per als documentalistes. En aquest sentit, si una USE es dóna en una freqüència alta (es va repetint en cada paràgraf o en cada apartat) i/o integra el títol del document, algun dels seus subtítols, esquemes o resums, existeix una probabilitat molt alta que es tracti d'una unitat representativa del text. A més, els documentalistes tendeixen a seleccionar les unitats més especificades (les UTP), perquè permeten precisar més i, consegüentment, reduir el soroll que provoquen les unitats massa genèriques o polisèmiques.

Així, podem dir que el buidatge de les USE pertinents per indexar textos es diferencia d'altres finalitats professionals per:

a) la reducció de categories gramaticals: només noms i algun adjectiu

b) la quantitat d'USE: només aquelles unitats que són representatives del text (en aquest sentit la freqüència i la ubicació de les unitats en el text són dades imprescindibles)
c) la importància que poden tenir certs noms propis per identificar un document
d) el predomini d'unitats polilèxiques.

Els traductors són el col·lectiu que ha seleccionat més poques unitats. De fet, només han seleccionat les unitats que desconeixen o que els poden ocasionar problemes de traducció. Aquesta és la raó per la qual hi ha una reducció considerable dels tipus d'USE seleccionades:

- no seleccionen símbols
- no seleccionen noms científics en llatí
- seleccionen molt poques sigles.

Aquesta restricció és lògica, si tenim en compte que els símbols i els noms en llatí de les nomenclatures científiques, en general, són universals i, per tant, no són objecte de traducció. Les sigles que s'usen en medicina apareixen habitualment en anglès (per la seva condició de llengua internacional), fet que elimina qualsevol problema de traducció. Malgrat això, els traductors han marcat algunes sigles que, tot i no traduir-se, són poc conegudes i les han acompanyades del context en el qual es troba el seu referent: *concentració inhibidora mínima* (CIM).

Una altra peculiaritat del buidatge dels traductors és el fet que hagin seleccionat les USE en context sintàctic perquè, moltes vegades, és el context que els proporciona elements lingüístics i pragmàtics per proposar els equivalents més adequats[11]. També han marcat els referents

[11] Com la categoria gramatical d'una USE: *la IFI;* l'equivalent d'una USE: *immunotransferència de Western (Western Immunoblot), taca negra (tache noire);* els complements preposicionals de les USE: *granulomes centrats per un vacúol lipídic,*

socioculturals que inclou el text, que en la traducció caldrà adaptar o explicar[12].

Finalment, pel que fa al buidatge, els terminògrafs són els més exhaustius en la selecció d'unitats, encara que després no aprofitin totes les USE seleccionades per fer un diccionari.

És interessant observar que els terminògrafs que han marcat verbs (AE i CG) en alguns casos han fet referència també al seu context d'ús. El context pot facilitar exemples, en el cas que el diccionari n'inclogui, però també pot indicar la presència de fraseologia verbal. Aquesta fraseologia ha estat tradicionalment exclosa dels diccionaris especialitzats, però, en els darrers anys, s'ha començat a considerar la pertinència d'introduir-la en aquest tipus d'obra.

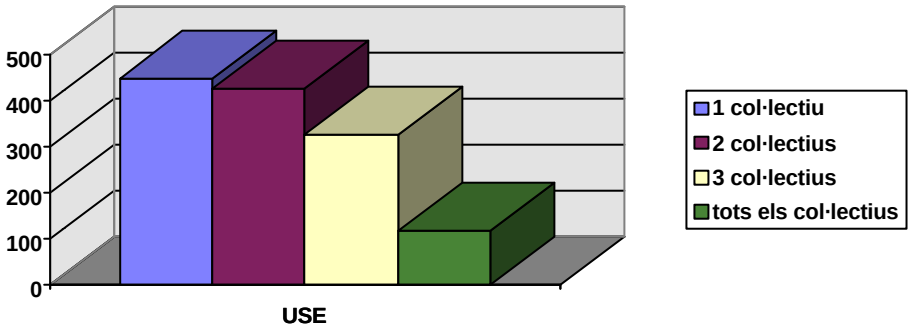
6.4.2 USE comunes a *tots* els col·lectius professionals

Hi ha un conjunt poc nombrós d'USE assenyalades pels quatre col·lectius professionals:

infiltrats radiològics bilaterals en camps mitjans i inferiors, títols a partir d'1/80; marques tipogràfiques: *prova "específica", "vasculitis";* variants d'una mateixa USE: *aglutinació de làtex / aglutinació en làtex, fixació de complement / fixació del complement, tifus murí o endèmic;* les sigles acompanyades del segment al qual substitueixen: concentració *inhibidora mínima (CIM), creatinofosfocinasa (CPK), factor de necrosi tumoral (TNF), immunofluorescènica indirecte (IFI),* etc.

[12] "*La referència més remota que hom té del tifus exantemàtic procedeix* ***del convent de La Cava****, l'any 1083.*"

	metges, documentalistes, traductors, terminògrafs
unitats monolèxiques nominals	59
unitats polilèxiques nominals	57
sigles	3
total	**119**



Nombre total d'USE seleccionades conjuntament per més d'un col·lectiu professional

De 1.268 ocurrències diferents, només 119 unitats són compartides per tots quatre col·lectius. **La xifra de coincidència,** que és realment **baixa**, es pot argumentar pel fet que les necessitats d'USE de cada grup estan molt allunyades: hi ha grups que han seleccionat molt poques unitats (documentalistes i traductors) i grups que n'han marcat moltes i molt variades (especialistes i terminògrafs). Aquesta dada corrobora que un

SEACUSE no pot ser *universal*, sinó que s'ha d'adaptar a cada activitat professional si vol donar compte dels seus requisits específics.

Observem, a més, que la coincidència es dóna només entre les **unitats nominals**, que són les clàssicament considerades terminològiques. Quant a les USE verbals, adjectivals o adverbials no hi ha cap coincidència compartida entre els quatre col·lectius, per bé que, com veurem en la taula de l'apartat 6.4.4, hi ha certs col·lectius que coincideixen en algunes de les seves tries.

6.4.3 USE marcades per *un sol* col·lectiu professional

En contraposició amb el grup anterior, trobem que hi ha força USE que han estat seleccionades només per un col·lectiu professional; la taula següent presenta aquestes peculiaritats:

	metges	**documentalistes**	**traductors**	**terminògrafs**
noms simples	32	11	11	48
noms polilèxics	70	20	14	117
verbs	3	0	1	34
adjectius	5	3	13	10
adverbis	1	0	1	0
sigles	1	0	0	0
frases discursives	1	2	24	1
símbols	0	0	0	0
noms científics	1	3	0	3
noms propis	0	18	1	0
total	**113**	**57**	**65**	**213**

Entre les unitats marcades només per un col·lectiu professional també es poden observar diferències quantitatives importants. Els terminògrafs són

els que han seleccionat més USE en general i també més USE lingüístiques.

Els traductors, en canvi, es diferencien dels altres grups per la quantitat de frases o de sintagmes que han destacat. Els adjectius que han marcat no són compartits pels especialistes ni pels terminògrafs, moltes vegades són adjectius o noms de caràcter no especialitzat que formen part d'una UTP, fet que indica que els interessen els segments desconeguts que els poden generar problemes de significat i/o d'equivalència.

Els documentalistes es caracteritzen pel fet de destacar una quantitat important de noms propis (la majoria, noms dels paï sos on es dóna una malaltia infecciosa). Els adjectius seleccionats pels documentalistes i els traductors no són compartits per cap altre grup. Finalment, els especialistes també han seleccionat algunes USE nominals que no comparteixen amb els altres col·lectius.

6.4.4 USE compartides per *alguns* col·lectius professionals

Hi ha USE que, si bé no les han seleccionat tots els col·lectius, són compartides per dos grups professionals. El quadre següent mostra aquestes coincidències i especifica quins són els grups professionals que coincideixen:

	metges i document.	**metges i traductors**	**metges i lingüistes**	**document. i traductors**	**document. i lingüistes**	**traductors i lingüistes**
noms simples	6	1	100	1	7	5
noms polilèxics	9	4	207	4	8	3
verbs	0	0	13	0	0	0
adjectius	1	2	10	1	0	1
adverbis	0	0	4	0	0	0
sigles	0	0	1	0	0	0
símbols	0	0	3	0	0	0
noms científics	2	0	18	0	0	0
total	**18**	**7**	**356**	**6**	**15**	**9**

D'aquesta taula ressalta la poca rellevància de les coincidències de la majoria de parelles de col·lectius professionals, a excepció de la parella **metges-terminògrafs** que s'explica per l'interès comú a seleccionar les unitats que transmeten coneixement especialitzat, ja sigui a través de conceptes en el cas dels especialistes o a través de significats en el cas dels lingüistes.

També trobem algunes unitats compartides per **tres col·lectius**: metges-documentalistes-terminògrafs o metges-traductors-terminògrafs. No hem documentat cap unitat seleccionada només per metges-documentalistes-traductors o per documentalistes-traductors-terminògrafs. Les dades següents reforcen la similitud, des del punt de vista quantitatiu, dels buidatges metges-terminògrafs, centrada sobretot en les UT:

	metges, documentalistes i lingüistes	**metges, lingüistes i traductors**
noms simples	124	0
noms polilèxics	116	23
verbs	0	0
adjectius	1	2
adverbis	0	0
sigles	9	0
símbols	3	0
noms científics	18	0
total	**271**	**25**

6.5 Principals problemes dels buidatges

Hem cregut interessant recollir i analitzar els problemes que el buidatge del text ha ocasionat als diferents professionals. Així, amb l'objectiu de conèixer les dificultats que la selecció d'USE podia plantejar a cada un dels col·lectius, hem preguntat directament als informadors *quins eren els problemes amb els quals s'havien trobat quan seleccionaven les USE del text*.

De les seves respostes, es desprèn que, d'una banda, hi ha problemes que afecten tots els col·lectius i, de l'altra, hi ha problemes exclusius de cada col·lectiu. Les qüestions comunes a tots col·lectius responen a dos factors:

1. La interdisciplinarietat del lèxic de la medicina
2. El nombre d'USE seleccionades.

Els problemes propis de cada col·lectiu professional fan referència bàsicament a cinc aspectes:

3. El valor conceptual de les USE d'un text (metges)
4. Les variants denominatives i les paraules polisèmiques (documentalistes)
5. La subjectivitat de la selecció d'USE (traductors)
6. El desconeixement del tema (terminògrafs)
7. La indefinició de la finalitat del buidatge (terminògrafs).

Paral·lelament a les dificultats de selecció de les USE pertinents del text a les quals explícitament al·ludeixen els professionals, hem constatat, a través de l'anàlisi dels resultats obtinguts, problemes d'asistematicitat de selecció de certes USE. En aquest sentit, hem observat incoherències en els criteris de selecció d'USE i, més en concret, en els tres aspectes següents[13]:

8. La pertinència de la fraseologia especialitzada
9. La diversitat de categories gramaticals
10. La confusió d'USE amb unitats de coneixement especialitzat més complexes.

A continuació ens ocupem d'analitzar aquests deu problemes, explícits o implícits, derivats dels diversos buidatges.

6.5.1 La interdisciplinarietat del lèxic de la medicina

Una de les característiques de l'àmbit de la medicina és el caràcter interdisciplinari del seu lèxic. A excepció dels documentalistes, els informadors destaquen com un problema que, en el text, a més dels

[13] Aquests tres punts de desequilibri afecten alguns dels aspectes menys estudiats i sobre els quals hi ha menys consens entre els terminòlegs teòrics.

"termes mèdics", hi hagi també "termes" d'altres disciplines, com la biologia, la zoologia, la química o la farmacologia.

Aquesta característica del text els planteja dubtes pel que fa a la selecció de les unitats que consideren que pertanyen a àrees afins a la medicina. Les opinions següents avalen aquesta preocupació:

- *"no sabia si els noms dels animals també els havia de marcar"*
- *"no estic segur què havia de fer amb els noms llatins dels animals"*
- *"crec que has triat un capítol dedicat a unes malalties que en tocar l'àmbit microbiològic (més biològic que mèdic i també més epidemiològic que mèdic) inclou una petita part de vocabulari especialitzat d'aquests camps"*
- *"he tingut molts problemes de selecció de termes a causa de la interdisciplinarietat del text"*
- *"hi havia termes relacionats amb la terminologia mèdica, però no de manera exclusiva"*
- *"hi ha molts termes que es van allunyant del que podríem dir nucli dur de la terminologia mèdica"*
- *"potser caldria fer una distinció organitzativa temàtica dels termes: medicina, zoologia, botànica, química, etc."*

Cal observar que les unitats especialitzades d'àrees no pròpiament mèdiques (*gos, puça, rickèttsia*, Coixella, *cultiu cel·lular, tinció de Giemsa, dooxiciclina, ofloxacina, quinolona*, etc.) usades en el context de la medicina solen adquirir un significat determinat.

Per exemple, les *rickèttsies* són per als metges *agents bacteriològics que causen les rickettsiosis* i, en canvi, per als biòlegs són *"bacteris paràsits intracel·lulars dels vertebrats, que tenen un cicle natural en el qual intervenen artròpodes hematòfags, que pertanyen a la família de les rickettsiàcies i a l'ordre del rickettsials*" (DLC). Un altre exemple: la

tetraciclina és per a un metge "*qualsevol dels antibiòtics d'ample espectre d'acció bacteriostàtica, actius contra una gran varietat de microorganismes entre ells bacteris grampositius i gramnegatius, rickèttsies, micoplasmes, clamícide, certs virus i actinomicets*" (DEM); a més, els metges saben que la tetraciclina "*s'absorbeix bé per via oral i té afinitat electiva per les cèl·lules tumorals, el teixit ossi i els teixits amb inflamació crònica necrotitzant*" *(DEM).* En canvi, per a un farmacèutic, *la tetraciclina és* "*una substància que es presenta en forma de pólvores grogues, inodores, estables a l'aire, però sensibles a la llum*" i, finalment, per a un químic la tetraciclina és "*una substància de fórmula* $C_{22}H_{24}N_{20}O_8$"[14][15].

L'explicació del fet que els professionals se sentin incòmodes davant de textos que presenten lèxic interdisciplinari, però, l'hem de buscar en els tres supòsits següents:

1. La idea clàssica que els termes **pertanyen a una àrea temàtica** determinada:

> *Los términos son elementos léxicos que* ***pertenecen a***[16] *áreas especializadas de uso en una o más lenguas.*

[14] Rondeau (1984), en l'annex del primer capítol, es refereix a una situació paral·lela quan es planteja el problema del mots que s'usen tant en el discursos generals com en els discursos especialitzats:

> *What happens if we find that words like sugar, spirits, soap, rubber, plastics, and so on are words which can exist both in the everyday and the specialised chemical language? To answer this question we must compare the definitions in the chemical dictionary with those in the general one. It will be interesting to note that the way they are understood by a layman is quite different from the way they are used in chemistry or chemical industry.*
>
> [Rondeau, 1984:190]

[15] Com assenyala Gutiérrez (1998: 23) no tots els científics s'aproximen a la realitat de la mateixa manera, sinó que cada branca té els seus objectius i les seves preferències que varien la perspectiva des de la qual es contemplen les coses: "*Un médico puede pensar en medicamentos antihipertensivos, antitusígenos o antiácidos apoyándose en su función, y para esas realidades, un químico hablará de diferentes composiciones químicas; donde éste ve sulfuro de hierro o fluoruro de calcio, un especialista en minerales verá pirita o fluorita, porque piensa en su aspecto externo, en su forma de cristalizar, etc., además de en su composición química.*"

[16] El subratllat és nostre.

[Sager,1993: 21]

2. La creença consegüent, repetida per la Teoria General de la Terminologia, que el terme és **unívoc** i **monoreferencial**:

> *S'ha mantingut que tots els termes formen part d'una disciplina i que és en el si de la qual que adquireixen el seu significat, de forma que una mateixa unitat usada en una disciplina diferent es considera un terme diferent. Aquest plantejament deixa sense cap explicació la mobilitat de conceptes i termes dins de la ciència i les especialitats.*
>
> [Cabré, 1998b: 14]

3. La influència que exerceix la **divisió compartimentada del coneixement** en disciplines; sobretot en les disciplines que solen comptar amb una gran tradició universitària i professional (com és el cas de la biologia, la botànica, la química, la física, les matemàtiques, etc.):

> *Una subdivisió rigorosa de la realitat pressuposa un desconeixement sorprenent de la realitat científica i professional. En una mateixa disciplina conviuen escoles de pensament diferents, corrents organitzats, treballs de recerca que es duen en paral·lel i que no són necessàriament coincidents. La teoria terminològica pressuposa que tota la realitat especialitzada està regulada per consens entre els especialistes, però aquesta suposició és d'un idealisme absolut. En qualsevol esfera d'activitat humana hi ha constantment conceptes en circulació i els contextos socials, culturals i lingüístics on es desenvolupen activitats especialitzades són molt diferents.*
>
> [Tebé, 1996:153]

En aquesta mateixa línia es pronuncia Cabré (1998b:14) quan afirma que una mateixa realitat pot ser percebuda de maneres diferents i, consegüentment, conceptualitzada diversament per disciplines diferents.

6.5.2 El nombre d'USE

Les dades obtingudes corroboren que el nombre d'USE detectades en un text està en relació amb la finalitat professional per la qual es requereixen. En aquest sentit, molts informadors mostren inseguretat sobre el nombre d'USE que cal seleccionar i davant dels dubtes, tots diuen que han tendit a marcar-ne amb escreix, de manera que tant afirmen que han tendit a marcar moltes unitats els que només n'han assenyalat poques (60 unitats diferents) com els que n'han seleccionat moltes (800 unitats diferents):

- *"en els casos que em resultava dubtós considerar-lo com a terme en sentit estricte, els he marcat amb un criteri ampli"*
- *"més aviat he tendit a incloure tota la terminologia encara que fossin termes bàsics perquè llavors, depenent de les possibilitats d'extensió del treball, sempre es pot fer una selecció més específica"*
- *"he marcat més per excés que no pas per defecte"*
- *"he fet una selecció molt àmplia".*

A parer nostre, el nombre d'USE pertinent depèn de la funció de la selecció i els dubtes pel que fa al nombre d'USE pressuposen o que les funcions no són prou explícites o que certes unitats són tipològicament i funcionalment ambigües.

6.5.3 El valor conceptual de les USE

Segons els comentaris dels especialistes i dels terminògrafs, no totes les USE especialitzades del text tenen la mateixa importància en la medicina. Hi ha un grup d'USE relatives a les patologies i l'anatomia humana que constitueixen el nucli conceptual de la medicina, i d'altres, en canvi, serveixen per denominar coneixements relacionats amb aquests dos centres de coneixement, són les referides a coneixements biològics,

bioquímics, físics, farmacològics, químics, sociològics, jurídics, econòmics, etc.

Un exemple que permet il·lustrar la característica de nuclear o complementari de les USE és el comentari d'un dels metges (PH) sobre el segment *augment de la permeabilitat vascular;* PH diu que, estrictament, només s'hauria de considerar com a terme mèdic *vascular* perquè *permeabilitat* és un mot manllevat a la física i *augment* forma part del lèxic general; però tot i així —raona l'especialista —, els metges fan servir el segment sencer *augment de la permeabilitat vascular* amb un significat concret, ja que implica unes característiques determinades i ocasiona unes conseqüències específiques. Però malgrat que s'usen segments d'aquest tipus amb un significat especialitzat concret, els especialistes consideren que la unitat *augment de la permeabilitat vascular* i, per exemple, les unitats *hepatitis*, *rickettsiosi* o *febre tacada*, tenen una importància diferent en tant que USE mèdiques.

El mateix passa entre *rickettsia* i *rickettsiosi*. Segons els metges, *rickettsiosi* és un terme exclusivament mèdic: és el nom d'una malaltia; en canvi, *rickettsia* és un terme de la microbiologia, però alhora és l'agent de la *rickettsiosi* i, per tant, integra també el coneixement especialitzat d'un metge.

6.5.4 Les variants denominatives i les paraules polisèmiques

Els documentalistes seleccionen les unitats que representen el contingut global del text. Els principals problemes que es donen en aquest procés són el tractament de les variants denominatives i el de les unitats polisèmiques

(polisèmiques en el sentit que són unitats que en altres àmbits del coneixement tenen significats diferents)[17].

Si s'elimina la variació, es limita la recuperació dels textos i la cerca de documents és molt rígida. El mateix passa amb les unitats polisèmiques: si se seleccionen, les cerques s'amplien, però al mateix temps aquestes es diversifiquen molt i hi ha el perill que es facin cerques inoperants per excés d'informació.

Habert i al. (1997:106-107) comenten amb un exemple molt il·lustratiu aquest problema:

> *Considérons une requête d'un étudiant en médecine: problème de circulation dans les artères. Un système fondé sur les mots clés indexe cette requête comme CIRCULATION et ARTÈRE qui comporte aussi bien des textes sur la circulation sanguine que des textes sur la circulation automobile. En réponse à sa requête, l'automobiliste va donc trouver beaucoup de textes médicaux non pertinents pour lui (faible précision) tandis que des textes qui l'auraient intéressé en sont pas sélectionnés parce qu'ils parlent de trafic et non de circulation (faible rappel). Prendre en compte les relations de synonymie permettrait de gagner respectivement en rappel et en précision*.

Aquesta manca de precisió també es dóna dintre d'un mateix domini d'especialitat, quan s'escullen mots clau poc representatius del contingut global del text; Així, per exemple, si s'utilitza el mot *cervell* (que en aquest text és una informació poc rellevant que només surt una vegada) com un mot clau per indexar el text sobre malalties infeccioses, quan se cerca informació sobre el funcionament del cervell, entre d'altres molts textos, també es recuperarà aquest text sobre malalties infeccioses, que serà un text molt perifèric. I el mateix succeeix quan s'usen com a descriptius mots tan genèrics com *manifestació, afectació, afecció, prova*, etc. que, si no van

[17] Un dels primers sistemes d'indexació de documentació mèdica ja feia referència a aquests dos aspectes [Graitson, 1975].

subespecificats per un complement que restringeixi la informació, generen una gran quantitat de soroll informatiu[18].

6.5.5 La subjectivitat de selecció d'USE

La tria d'USE pot estar condicionada també per factors molts subjectius com és el cas de la selecció d'USE que els traductors fan abans d'iniciar una traducció. En efecte, com ells mateixos afirmen, la seva selecció és subjectiva perquè depèn del grau de coneixement del tema. Com més coneixement d'un àmbit temàtic, més reduï t serà el *buidatge* inicial. Les dues opinions següents reflecteixen aquesta idea:

- *"Els termes del text són gairebé tots els noms del text, però els termes que interessen un traductor són només els més estranys, aquells menys habituals i que plantegen algun problema de traducció."*
- *"Els termes que ha de recollir un traductor abans de començar una traducció són només els que li poden plantejar algun problema lingüístic, pragmàtic o cognitiu."*

6.5.6 El desconeixement del tema

El desconeixement del tema del text de buidatge és un altre problema que dificulta la selecció d'USE d'un text i, d'una banda, condueix a no assenyalar certes unitats del text que les dades analitzades mostren que només els especialistes reconeixen, i, de l'altra, a delimitar inadequadament una unitat. Els terminògrafs i els traductors són els grups que han explicitat tenir més problemes per aquest motiu:

[18] Sobre aquest tema vegeu la tesi de Romana (1997).

- *"El desconeixement de l'àrea m'ha fet dubtar si alguns termes eren més aviat d'àmbit general o si eren pertinents"*

Les USE que només detecten els especialistes són paraules del lèxic general que dins d'un àmbit especialitzat com la medicina adquireixen un sentit específic:

> *adult, especificitat, criteri, efectiu, eficaç, inhibir, activitat, ben tolerat, aglutinar, proliferar, eficàcia, incidència, importància epidemiològica, cas esporàdic, distribució mundial, exposició, protecció individual, secundàriament, predomini, inespecífic,* etc.

Aquestes unitats, des del punt de vista formal, no presenten elements d'especificitat respecte de les unitats generals (ni en els formants, ni en els afixos, ni en les estructures); en canvi, des del punt de vista conceptual són unitats portadores de significat especialitzat[19]. Aquestes unitats són les

[19] Reproduï m alguns comentaris dels especialistes en relació amb aquesta qüestió, en els quals hem remarcat en negreta les informacions més rellevants:

> **adult**
> *Depèn com es miri no formaria part del llenguatge especialitzat, encara que té molta utilitat com a* ***element classificador*** *de la població en grans grups d'edat en referència a la possibilitat de patir malalties. Penso que en el llenguatge comú adult s'utilitza com a sinònim de madur psicològicament i adaptat al rol social que li correspon mentre que a nivell biomèdic indica el grau de desenvolupament fisiològic i morfològic i permet classificar i preveure el risc de certes malalties.* (TV)
> **zona coberta**
> *Com a paraules aïllades zona i coberta formen part del llenguatge comú,* ***en microbiologia i en malalties infeccioses es refereix a*** *aquelles parts del cos que estan cobertes normalment per la roba i que per tant o bé són tan susceptibles a l'acció d'altres agents externs com als raigs del sol.* ***Les afectacions de les zones cobertes indiquen*** *parts tapades, humides amb tendència a suar i patir infeccions fúngiques o a la proliferació de paràsits. Les afectacions de les* ***zones descobertes*** *són aquelles que actuen sobre parts desprotegides dels agents externs. Per tant crec que té* ***un significat especialitzat o almenys unes projeccions de significat que un parlant no especialitzat no arriba mai a utilitzar.*** (TV)
> **inhibit, -ida; inhibir**
> *No sé si podria dir que és una paraula exclusivament biomèdica. En biomedicina bàsica (fisiologia, bioquímica, endocrinologia i farmacologia)* ***el concepte d'inhibició és bàsic per explicar el funcionament del sistema nerviós i el metabolisme****. De fet, ambdós sistemes són grans grups de senyals que s'activen els uns als altres ja sigui per mitjà d'estímuls elèctrics o neuroquímics o hormonals. En aquest context parlem d'una neurona inhibida com aquella no que no fa res, sinó que rep senyal explícit de no transmetre senyal o parlem que la producció d'una hormona és inhibida per l'acció d'una altra hormona. La inhibició, així com el seu* ***concepte contrari, l'estimulació o inducció*** *són els noms que reben les accions actives i no les*

que causen més problemes als traductors perquè pel seu caràcter polisèmic poden donar lloc a falsos equivalents.

Al costat de les paraules que la resta d'especialistes difícilment reconeixeria com a especialitzades perquè tenen un sentit general, el desconeixement del tema també genera problemes de delimitació[20], sobretot en segments formats per un nom i un o més adjectius modificadors sense caràcter especialitzat *per se*:

> *cèl·lules* ***gegants****, insuficiència cardíaca* ***greu****, febre* ***alta,*** *tractament* ***clàssic,*** *adenopatia* ***regional sensible****, anèmia normocròmica* ***moderada****, afectació* ***general moderada****,* etc.

---

no accions que experimenten els diferents elements d'alguns sistemes corporals per funcionar tal i com ho han de fer. A nivell més mèdic, ***la gran majoria dels medicaments*** *que s'utilitzen tenen un mecanisme d'acció basat en la inhibició o l'estimulació de la producció d'alguna substància o senyal neuroquímic i això fa que aquest mot també sigui molt utilitzat pels metges per explicar els mecanismes d'acció dels medicaments.* (TV)

incidència

El mot incidència és d'ús general però ***en utilitzar-la el parlant no coneix ni utilitza els matisos que un metge o un epidemiòleg faria servir****.* ***El terme incidència i el de prevalença es contraposen i es complementen****. La incidència i la prevalença són indicadors de freqüència d'aparició d'un problema de salut utilitzat en epidemiologia per estudiar les necessitats sanitàries de la població. La incidència és el nombre de casos d'una malaltia presents en un moment donat de temps en una població. Hi ha malalties que tenen una baixa incidència i una elevada prevalença, per exemple, l'esclerosi múltiple donat que són trastorns crònics i duren moltíssim i malalties que tenen una elevada incidència i una baxíssima prevalença per la seva curta durada per exemple les càries. Per tant,* ***em sembla que és propi del llenguatge mèdic perquè el metge en utilitzar-la ha de conèixer aquests matisos mentre que el parlant no els coneix ni els utilitza.*** (TV)

insuficiència cardíaca o insuficiència cardíaca greu

Quan en biomèdica parlem d'insuficiència cardíaca greu jo no penso en una insuficiència cardíaca que és a més a més greu*, sinó en una entitat molt ben definida i aclarida que em ve al cap com un únic concepte.* ***Una insuficiència cardíaca greu ve definida per una sèrie de paràmetres clínics (signes i símptomes), biològics i electrofisiològics quantitatius que la distingeixen de la insuficiència cardíaca moderada o lleu i que de manera molt ràpida permeten al metge fer-se una idea de quins valors tenen tota una constel·lació de paràmetres que s'inclouen en aquest concepte i sobretot de la gravetat del cas.*** *De ben segur que inicialment el greu era un adjectiu aïllat del nom, però amb el temps, l'aparició de classificacions de síndromes i la necessitat de dir, suggerir o transmetre més informació amb la mínima inversió lingüística ha fet que sigui tractada pel cervell del metge experienciat com tot un conjunt. D'exemples amb l'adjectiu greu te'n podria donar molts més, només cal consultar les malalties que són sotmeses a classificacions. en tot cas em sembla que és la manera com emmagatzemem la informació els metges, per tal de dir molt amb les mínimes paraules.* ***És com una persiana de coneixements sota d'un nom on a mida que l'anem baixant apareixen coses i més coses que no es diuen directament però que se suposen.*** (TV)

Vistos els resultats es podria dir que per un especialista una unitat especialitzada no es limita a representar un sol concepte, sinó que inclou una sèrie de coneixements i connexions (formen part de taxonomies més o menys desenvolupades) als quals només tenen accés els que han estudiat específicament aquella parcel·la del coneixement.

[20] "*També he tingut en alguna ocasió dificultats de delimitació dels termes, com a conseqüència de la meva ignorància sobre el tema*".

6.5.7 La indefinició de la finalitat del buidatge

La indefinició del treball terminològic per al qual es realitza el buidatge comporta que no se sàpiga si una USE és pertinent o no. Aquest és un problema al qual s'han referit els terminògrafs que han participat de l'experiment.

Si tenim en compte que, en la pràctica, els seus buidatges sempre estan condicionats per un producte terminogràfic concret, la inseguretat a l'hora de decidir la pertinència d'un terme a què al·ludeixen és lògica per tal com se'ls havia fet volgudament una demanda massa general: fer el buidatge terminològic per elaborar un diccionari sense precisar-ne la funció. En aquesta línia, els comentaris dels terminògrafs palesen la dificultat de buidar un text en abstracte:

- *"La principal dificultat que he tingut no ha estat tant la identificació de termes, com la dificultat de saber si eren pertinents o no. I per saber si són pertinents, cal saber (al meu entendre) amb quin objectiu estem fent el buidatge? per a qui? per què? El problema de la delimitació del tema ha estat constant, necessitava un arbre de camp que no havíem fet perquè no teníem el treball definit"*
- *"La principal dificultat ha estat la selecció terminològica perquè no coneixia les característiques del lèxic que havíem de fer, ni la funció ni el tipus d'usuari que necessités aquest material, ni el grau d'especificitat del material, ni la quantitat de termes que havia d'incloure."*
- *"He tingut problemes per saber quin tipus de termes havia d'incloure (de medicina o d'altres disciplines, genèrics o només específics) ni quines categories gramaticals havia d'incloure perquè no tenia definit el producte que havia de fer".*
- *"Hi ha molts termes que es van allunyant del que en podríem dir el nucli dur terminològic del tema del text. I és un allunyament progressiu. Per exemple, entre els símptomes propis de les rickettsiosis, n'hi ha que són clarament terminologia específica com taca negra, però n'hi ha que no és gens específica com febre, diarrea i vòmit."*

6.5.8 La selecció de fraseologia especialitzada

Cal matisar d'entrada que la noció de fraseologia referida a un text d'especialitat no és fàcil d'establir, raó per la qual la selecció que els professionals han fet sol ser intuï tiva i, almenys aparentment, força arbitrària.

D'una banda, no tots els professionals marquen UFE; per exemple, cap documentalista troba pertinent la fraseologia. De l'altra, no totes les categories d'UFE són assenyalades com a pertinents; per exemple, els metges només assenyalen fraseologia nominal; els traductors només algunes ocurrències difícils de traduir que moltes vegades donen lloc a falsos amics; els terminògrafs són els únics que, a més de fraseologia nominal, també marquen fraseologia verbal, però no de manera explícita[21]. Així, al costat d'alguns verbs han seleccionat entre parèntesis l'argument que funciona com a tema i que denota la presència de combinacions fraseològiques:

> *administrar (un medicament), aparèixer (una malaltia), baixar (la febre), detectar (una malaltia), instaurar (un tractament),* etc.

L'asistematicitat de les UFE, per tant, sembla constant en tots els buidatges:

- se seleccionen asistemàticament els segments —que corresponen, en general, a UFE— que tenen com a nucli un nom deverbal

[21] Ja hem comentat que el buidatge dels traductors és molt subjectiu, aquesta pot ser la raó per la qual no han assenyalat cap UFE verbal.

- es delimiten asistemàticament les USE que integren una UFE
- s'assenyalen frases llargues que des de l'òptica lingüística són conjunts de termes lligats sintàcticament, però que no tenen ni caràcter terminològic ni fraseològic.

En aquest darrer cas, hem preguntat als especialistes en medicina si pensaven que determinats segments que, des del punt de vista lingüístic semblaven UFE (nucli eventiu de caràcter predicatiu, complement terminològic, absència de referencialitat, semifixació), representaven un concepte i, en la majoria de casos, han respost que es tractava de dues USE sovint usades conjuntament. En aquesta línia, hem reproduït els raonaments d'un metge (MF) a propòsit d'uns segments concrets considerats USE:

afectació del SNC

"*Crec que hi ha dos conceptes: afectació i SNC, i les dues paraules formen part per separat del lèxic mèdic. Afectació pot referir-se a diferents entitats nosològiques, per exemple afectació del SNC, afectació dermatològica, afectació hepàtica, etc.*" (MF)[22]

tractament de les rickettsiosis

"*Per a mi, hi ha dos conceptes independents: tractament i rickettsiosi. Tots dos pertanyen al lèxic mèdic. El tractament pot aplicar-se a qualsevol patologia i la rickettsiosi és una malaltia produïda per una rickèttsia.*" (MF)[23]

formació de microtrombes

"*Per a mi, hi ha tres conceptes: formació, micro i tromba. Només forma part del lèxic mèdic el mot compost microtrombes i el mot simple trombes.*" (MF)

presència de taca negra

[22] Fa referència al caràcter no autònom d'algunes unitats i a la recursivitat de la combinació.

[23] Aquest comentari reforça la idea de la recursivitat de certs patrons morfosemàntics que donen lloc, en aquest cas, a unitats fraseològiques especialitzades.

> "*Crec que hi ha dos conceptes: presència i taca negra, encara que només forma part del lèxic mèdic taca negra. Taca negra fa referència a una afecció concreta i clarament identificable per qualsevol professional mèdic.*" (MF)[24]

La conclusió que es desprèn d'aquestes constatacions és que la frontera entre fraseologia i lèxic no és sempre clara. Per als especialistes, les UFE i les UTP, seguint un criteri de selecció cognitiu, són unitats que els serveixen per transmetre els conceptes especialitzats. La distinció entre fraseologia (*tractament de la rickettsiosi*) i terminologia (*hepatitis bacteriana*) és, bàsicament, lingüística (i a vegades, fins i tot depèn de criteris pragmàtics i extralingüístics). Això explica que els terminògrafs hagin estat els professionals més sistemàtics quant al marcatge de les UFE i els únics que explícitament han fet referència a la diferència de comportament d'aquestes unitats.

6.5.9 La selecció de categories gramaticals

Moltes obres de terminologia teòrica han defensat que la categoria gramatical predominant i exclusiva dels termes és el nom (Rey, 1992) (Sager, 1998):

> *Como signo lingüístico, el término es una variedad funcional del sustantivo. Los términos son unidades léxicas que difieren de las palabras y los nombres propios. Las palabras son unidades lexicales que se dividen en clases como sustantivo, adjetivo, verbo, pronombre, preposición, etc. Los términos y los nombres propios sólo pueden tener la función sintáctica de los sustantivos.*
>
> [Sager, 1998: 13]

En una visió comunicativa de la terminologia, però, al costat dels termes que segueixen sent noms també tenen cabuda les USE d'altres categories gramaticals sense caràcter referencial: adjectius, verbs i adverbis. En

[24] Es distingeix les unitats amb caràcter referencial de les unitats sense valor referencial, que és un dels elements de distinció entre les UFE i les UT.

aquesta última línia, tot i que els buidatges realitzats constaten que el nombre d'USE nominals en els textos especialitzats és molt elevat, és un fet acceptar que en la comunicació especialitzada també s'usen verbs, adjectius i fins i tot adverbis amb significats específics.

Sovint es pot establir una cadena morfològica d'USE (*immunològicament-immunològica-immunologia*) que no se sol recollir completa en molts diccionaris perquè es considera que és una informació redundant ja que les regles de formació de paraules són en aquests casos transparents[25]. Però, que se'n pugui prescindir en un tipus de treball terminològic (com és l'elaboració de certs diccionaris especialitzats), no vol dir que no aparegui en els textos, i que per a determinades activitats professionals sigui necessari retenir-la.

La inclusió asistemàtica d'altres categories gramaticals, al marge dels noms, és un criteri diferencial entre els diversos professionals. Per posar un exemple podem citar el cas dels verbs:

	M1	M2	M3	D1	D2	D3	TR1	TR2	TR3	T1	T2	T3
verbs	1	16	3	0	0	0	13	1	0	0	20	41

Les dades obtingudes mostren que en la selecció de les USE verbals, els documentalistes són els més sistemàtics perquè no en seleccionen cap. En canvi, els buidatges de la resta de col·lectius són molt heterogenis: mentre que un dels terminòlegs en marca 40, n'hi ha un que no en selecciona cap. El mateix passa amb els metges: un n'assenyala 16 i, en canvi, un altre només un. Els traductors, en general, en marquen molt pocs, a excepció de XB, per la seva condició de planificador lingüístic més que no pas de

[25] El buidatge de la terminògrafa SM n'és un exemple: no ha marcat cap verb, un sol adverbi i molt pocs adjectius.

traductor. Cap d'ells, però, no ha assenyalat la totalitat dels verbs especialitzats del text, que són 52.

6.5.10 La confusió d'USE amb unitats de coneixement especialitzat més complexes

Per als metges, les USE es conceben des de la perspectiva d'unitats que transmeten coneixement especialitzat i, per aquest motiu, moltes vegades, els és difícil distingir entre les USE i altres unitats més complexes que també transmeten coneixement especialitzat. Així, al costat d'USE com:

> *ronyó, hepatitis, febre de les Muntanyes Rocalloses, afecció del SNC, SNC, pH*

també han assenyalat segments que traspassen la frontera de les USE, com:

> *polls infectats excreten rickèttsies per la femta, infiltració epidèrmica per cèl·lules mononucleades, pneumonitis amb infiltrats alveolars, proves analítiques sol haver-hi anèmia, incidència menor d'efectes secundaris, curs de l'afectació hepàtica és subclínica,* etc.

6.6 Resultats dels buidatges: anàlisi qualitativa

L'anàlisi dels resultats dels buidatges dels diferents col·lectius permet validar la hipòtesi formulada a l'inici del capítol que defensava la **manca de correspondència en la selecció de les USE d'un text entre col·lectius professionals,** en relació amb dos paràmetres:

1. El nombre d'unitats.

2. El tipus d'unitats.

En efecte, les dades mostren que les diferències dels buidatges entre els professionals són molt significatives tant en el **nombre** d'unitats que seleccionen com en els **tipus** d'USE seleccionades. Aquesta discrepància s'explica pel fet que **cada activitat professional requereix unes USE precises** per realitzar les seves funcions i alhora prescindeix d'unes altres que poden ser pertinents per a alguna altra activitat.

Ara bé, que la tria d'USE sigui específica per a cada treball no pressuposa que el concepte d'USE sigui plural. A parer nostre, la noció d'USE és única i el que canvia és el concepte d'USE ***pertinent,*** de manera que una USE, sense deixar de ser-ho, pot no ser pertinent per a una activitat concreta. Les anàlisis dels resultats obtinguts confirmen i validen la hipòtesi que **la pertinència d'una USE està condicionada per la finalitat professional.**

Així, per als especialistes en medicina, les USE pertinents són totes les unitats que transmeten coneixement especialitzat i aquesta condició la compleixen, per definició, **totes les USE d'un text.** Per als metges les USE són un subconjunt de les unitats de cognició especialitzada. Aquesta és la raó per la qual el buidatge dels metges és molt exhaustiu tant pel que fa al nombre com a la varietat de tipus.

Per als terminògrafs, en una primera fase també són pertinents totes les USE del text, fet que explica la coincidència de seleccions entre els metges i els terminògrafs. Però, en una segona fase, el terminògraf tria d'entre totes les USE seleccionades només les que són adequades als objectius, destinataris i funcions de l'obra terminogràfica.

Pel que fa als documentalistes, l'esquema cognitiu de pertinència que tenen respecte de les USE d'un text és més restrictiu que el dels especialistes i dels terminògrafs; per això, també la seva selecció d'USE és més pobra quantitativament i tipològicament.

De fet, un documentalista només està interessat per aquelles USE que representen més precisament el contingut general del text en què apareixen. Així, després d'una anàlisi conceptual del document, selecciona les USE que funcionen d'identificadors del seu contingut informatiu. Les USE pertinents són, doncs, mots d'identificació que permeten descriure, indexar, ordenar i recuperar un document. Aquesta afirmació explica que pràcticament només assenyalin noms ja que són les USE que descriuen més sintèticament el contingut d'un text, i dins dels noms, les unitats polilèxiques, perquè són les que permeten restringir més les cerques posteriors.

Finalment, en un nivell de restricció més fort des del punt de vista quantitatiu, però mitjà des del punt de vista tipològic, se situen els traductors. Les USE que assenyalen abans d'abordar una traducció són només les que poden ocasionar problemes de traducció. Les dificultats que les USE poden plantejar poden ser cognitives, lingüístiques o socioculturals i l'anàlisi dels seus buidatges confirma que només els interessen les unitats de les quals desconeixen el significat o bé les que intueixen que els ocasionaran problemes d'equivalència. Aquesta és la raó per la qual entre els resultats dels seus buidatges trobem també segments d'USE (i no unitats senceres), sobretot adjectius i noms de caràcter no especialitzat que solen integrar unitats polilèxiques. D'altra banda i d'acord amb aquesta lògica, totes les USE que no es tradueixen no les

prenen en consideració (com els símbols, els noms que formen part de nomenclatures, algunes sigles)[26].

Com a conseqüència de l'anàlisi de les dades també podem concloure que la selecció d'USE per part dels diferents col·lectius implica un cert grau de subjectivitat, en alguns casos més que en d'altres. Per ordre ascendent de subjectivitat, hem de situar en primer lloc el buidatge dels especialistes, seguit del dels documentalistes, del dels terminògrafs i sobretot del dels traductors que són el col·lectiu professional que actua més subjectivament.

6.7 Perfils de necessitats diferents

De les anàlisis quantitatives i qualitatives dels buidatges, podem concloure que les **necessitats professionals** condicionen els tipus i el nombre d'USE pertinents d'un text. Per tal de constituir els perfils partim d'una banda dels tipus d'unitats següents:

1. USE nominals (UT)
2. USE verbals
3. USE adjectives
4. USE adverbials
5. Col·locacions nominals
6. UFE nominals
7. UFE verbals
8. Sigles
9. Símbols
10. Noms llatins
11. Noms propis.

[26] Recordem que les unitats que generen problemes d'equivalència solen ser: les UFE, les sigles no internacionals, els epònims, les UTP amb constituents no especialitzats i els

I de l'altra d'un conjunt d'informacions sobre les unitats:

1. Freqüència d'ús
2. Context d'ús
3. Ubicació en el text
4. Família de paraules amb una mateixa base i/o formant.

D'acord amb aquestes dues variables establirem els quatre perfils que fan referència a les activitats professionals següents:

- la transmissió del coneixement especialitzat (especialistes)
- la indexació d'un text (documentalistes)
- la traducció especialitzada (traductors especialitzats)
- la pràctica terminogràfica (terminògrafs).

6.7.1 Les USE pertinents per a la transmissió del coneixement especialitzat (especialistes)

Les USE pertinents d'un text per als metges són **totes** aquelles unitats que transmeten coneixement especialitzat sobre el tema del text, per tant, interessa que un SEACUSE recuperi d'un text:

- USE nominals (és a dir UT)
- USE lèxiques verbals
- USE lèxiques adjectives
- USE lèxiques adverbials
- símbols
- sigles
- noms llatins
- UFE nominals amb un nucli deverbal

neologismes.

- combinacions recurrents nominals amb un nucli terminològic
- USE relacionades que comparteixen una mateixa base i/o formant culte.

6.7.2 Les USE pertinents per a la indexació d'un text (documentalistes)

Per als especialistes en documentació les USE pertinents d'un text d'especialitat són aquelles que funcionen d'identificadors del contingut informatiu del text i que els permeten descriure, indexar, ordenar i recuperar un text especialitzat determinat**.**

Consegüentment, seria útil que, quan un SEACUSE s'utilitzés per indexar un text, proporcionés informació **sobre la freqüència i sobre la disposició discursiva** de les USE en el corpus textual, de manera que **només** mostrés aquelles USE que superessin una freqüència determinada (com a mínim superior a cinc) acompanyades d'informació situacional i que responguessin als tipus següents, preferentment unitats nominals polilèxiques:

- USE nominals (preferentment UTP)
- USE adjectives
- símbols
- sigles
- noms llatins
- USE relacionades que comparteixen una mateixa base i/o un formant culte.

I a més, els noms propis en context d'ús.

6.7.3 Les USE pertinents per a la traducció especialitzada (traductors especialitzats)

Les USE que interessen els traductors són només aquelles que els podrien plantejar certa dificultat a l'hora de traduir-les: unitats de les quals desconeixen el significat o unitats que intueixen que ocasionaran problemes de traducció. Per això, a vegades només seleccionen segments d'UTP (i no la unitat sencera), sobretot les unitats (nominals o adjectives) de caràcter no especialitzat que integren alguna unitat polilèxica.

El fet que cada traductor tingui necessitats cognitives, lingüístiques i sociolingüístiquess diferents que depenen del seu nivell de coneixement del tema en ambdues llengües comporta que no hi hagi uns tipus d'unitats que interessin més que unes altres, tot depèn de l'experiència professional. En aquest sentit només sabem que:

- hi ha USE que no interessen un traductor perquè no es tradueixen
- les USE pertinents són les que poden generar problemes de traducció
- tots els elements textuals que facilitin la cerca d'un equivalent d'una USE són pertinents per a un traductor.

Paral·lelament, també hem arribat a la conclusió que les USE que ocasionen més problemes de traducció són:

- UFE
- sigles no internacionalitzades
- epònims[27]
- USE o segments d'USE sense caràcter especialitzat

[27] [Hoof van, 1986].

- neologismes.

Consegüentment, si un extractor ha de servir a les necessitats terminològiques de la traducció, és interessant que pugui recuperar les unitats següents, **totes dins del seu context d'ús,** perquè moltes vegades el context ofereix dades que faciliten la comprensió de la unitat o la cerca del seu equivalent:

- UT
- USE lèxiques verbals
- USE lèxiques adjectives
- USE lèxiques adverbials
- sigles
- UFE nominals amb un nucli deverbal
- combinacions recurrents nominals amb un nucli terminològic
- UFE verbals
- les USE relacionades que comparteixen una mateixa base i/o formant culte.

6.7.4 Les USE pertinents per a la pràctica terminogràfica (terminògrafs)

Per al terminògraf seria important que un SEACUSE recuperés les següents USE del text acompanyades, obligatòriament o opcionalment, **del context[28] i la freqüència d'ús,** i relacionés les USE lingüístiques amb les no lingüístiques:

[28] Els contextos d'ús de les unitats poden ajudar al terminògraf a construir la definició d'una unitat i també són una font per trobar exemples d'ús.

- UT
- USE lèxiques verbals
- USE lèxiques adjectives
- USE lèxiques adverbials
- sigles[29]
- símbols[30]
- noms llatins
- UFE nominals amb un nucli deverbal
- combinacions recurrents nominals amb un nucli terminològic
- UFE verbals
- totes les anteriors USE relacionades perquè comparteixen una mateixa base i/o formant culte.

6.8 Conclusions

El primer objectiu d'aquest capítol ha estat confirmar i validar la idea que la pertinència d'una USE depèn de les **necessitats professionals que genera una determinada activitat**. Després de verificar aquesta hipòtesi, el segon objectiu que ens hem fixat ha estat establir les necessitats terminològiques de quatre tasques que impliquen l'ús de textos especialitzats: la transmissió del coneixement, la indexació, la traducció i l'elaboració de diccionaris.

[29] Les sigles han d'estar relacionades amb el segment al qual substitueixen, per això és necessari tenir el context en què apareix una sigla perquè és possible que trobem una referència al segment a partir del qual s'ha format.

[30] Els símbols i els noms llatins pertanyen a un altre codi i, per tant, no solen formar part de la nomenclatura dels diccionaris. Se solen incloure com a informació complementària a l'interior de cada entrada, per això és important que el terminògraf tingui el context d'ús d'un símbol o d'un nom científic per tal de poder-les relacionar amb el segment desenvolupat al qual substitueixen.

Les anàlisis dels resultats del buidatge d'un text que han dut a terme tres persones de quatre col·lectius professionals diferents (metges, documentalistes, traductors especialitzats i terminògrafs) han servit per confirmar les hipòtesis inicials expressades a 6.3. En efecte, d'una banda, hem comprovat que les necessitats terminològiques que generen certes activitats professionals van més enllà de les UT pròpiament dites. De l'altra, hem confirmat que no totes les USE que conté un text són pertinents per a totes les activitats. Hem comprovat també que cada activitat professional requereix uns tipus d'USE específiques i un nombre d'USE determinades. En aquest sentit, hem vist que les diferències entre les necessitats que origina una activitat professional són tant de tipus quantitatiu com qualitatiu.

Hem arribat a la conclusió que la noció d'**USE** és la mateixa per a tots els col·lectius professionals, però que, en canvi, varia la noció d'**USE pertinent**. Entre aquests dos conceptes hi ha una diferència de restrictivitat condicionada per la funcionalitat. Així, hem mostrat que, en el marc d'una activitat professional, la funcionalitat condiciona la selecció dels tipus d'USE:

- per als metges, amb la finalitat de transmetre els conceptes de la matèria, les USE pertinents són unitats que transmeten coneixement especialitzat
- per als documentalistes, que indexen textos, les USE pertinents són unitats representatives d'identificació de documents
- per als traductors especialitzats, les USE pertinents són bàsicament unitats que presenten problemes de coneixement o d'equivalència
- i, finalment, per als terminògrafs les USE pertinents són unitats lingüístiques especialitzades pertinents per a un àmbit especialitzat determinat.

Consegüentment, a partir dels resultats dels dotze buidatges, hem establert les necessitats d'USE que genera cada una de les quatre

activitats professionals analitzades i hem vist que algunes d'aquestes activitats, amb la finalitat de facilitar la selecció definitiva, requereixen que les USE vagin acompanyades d'informació complementària relativa al context d'ús immediat, la freqüència d'ús i/o la ubicació en determinades parts del text.

Més concretament, hem constatat que als especialistes els interessen les unitats sense cap informació addicional; els documentalistes, en canvi, només consideren pertinents les unitats nominals més desenvolupades que tenen una freqüència d'ús alta i que apareixen als títols, resums o esquemes; als traductors, els interessen bàsicament les USE amb informació sobre el context; i als terminògrafs, malgrat que aï llen les USE del text, els interessa poder accedir en tot moment al context i a la freqüència d'ús d'una determinada ocurrència per decidir si una unitat és pertinent per a un treball concret. Hem vist també que hi ha tipus d'unitats que no són pertinents per a una determinada activitat com, per exemple, la majoria dels símbols, els noms científics de les nomenclatures o les fórmules, en el cas dels traductors; o les USE de baixa freqüència en el cas dels documentalistes.

En resum, podem confirmar un cop més que els resultats i els comentaris dels professionals validen la hipòtesi que les activitats professionals condicionen les necessitats terminològiques respecte d'un text especialitzat i que aquestes necessitats es poden perfilar. En aquest sentit, hem proposat quatre perfils de necessitats terminològiques que generen els textos d'especialitat. Perfils que responen a les necessitats més generalitzables de cada activitat que hem estudiat. A mesura que s'aprofundeixi en l'anàlisi de cada activitat es podrà, d'una banda, construir nous perfils, i, de l'altra, proposar elements que afinin aquests perfils.

6.9 Recapitulació

En aquest capítol, ens hem plantejat inicialment si les USE d'un text tenen interès i pertinència general i hem verificat que no, que cada activitat professional selecciona unes USE determinades sempre en el marc d'un objectiu funcional.

A continuació, a partir dels resultats dels buidatges d'un text per part de dotze professionals pertanyents a quatre col·lectius, hem analitzat quines són les necessitats especialitzades de quatre tasques concretes que tenen com a objecte de treball els textos especialitzats —la transmissió del coneixement especialitzat, la indexació de textos, la traducció especialitzada, l'elaboració de diccionaris especialitzats. Les USE pertinents per a cada una de les quatre activitats presenten diferències tant de tipus quantitatiu com qualitatiu. I, finalment, al costat dels problemes que ha generat el buidatge del text, hem perfilat la noció d'USE pertinent per a cada activitat professional tractada.

En el proper capítol proposarem un model de SEACUSE que en la generació i presentació dels resultats tingui en compte la diversitat d'aquestes necessitats professionals.

7. DISSENY D'UN SEACUSE ADEQUAT A LES NECESSITATS PROFESSIONALS

> Decir que una cosa vale es que la consideramos buena para cierto uso. Así, pues, el valor de las cosas se funda en su utilidad o, lo que viene a ser lo mismo, en el uso que podemos hacer de ellas,
>
> [Condillac, 1789, t4: 10][1]

> A mediados de los ochenta se produce la siguiente contradicción: por un lado se hace evidente que las expectativas creadas por la lingüística computacional no pueden verse colmadas a corto plazo con aplicaciones complejas, mientras que, por otro lado, los sistemas informáticos son cada vez más potentes y cada vez son más numerosos los dominios profesionales a los cuales la informática aporta soluciones que revolucionan los métodos de trabajo y que reducen astronómicamente los costes humanos de muchas prácticas. En este contexto y desde hace diez años aproximadamente, empieza a consolidarse la ingeniería lingüística que permite la aplicación de los conocimientos lingüísticos a la industria, las comunicaciones etc.
>
> [de Yzaguirre, 1997: 69]

La majoria de sistemes d'extracció de terminologia s'han dissenyat al marge de les activitats concretes per a les quals havien de ser utilitzats; així, encara que, per exemple, TERMINO va ser creat amb finalitats terminogràfiques, LEXTER i FASTR, per a la recuperació d'informació i HEID per a la traducció, segons els seus autors, són sistemes polivalents aptes per a més d'una finalitat. Tot i que de fet, en aquests extractors la finalitat no és un paràmetre que condicioni ni les fases del procés d'extracció ni els resultats perquè, sigui quina sigui la seva utilitat, extreuen una única llista d'unitats. Això fa que la selecció final manual que l'usuari ha de fer de les unitats generades pel sistema sigui molt feixuga, perquè es produeix molt de soroll i de silenci, no només pel que fa

[1] Cito a través de Foucault (1968).

als tipus d'USE que proporciona el sistema, sinó sobretot per la falta de refinament en **la pertinència professional de les unitats**[2].

En contraposició amb els sistemes tradicionals, en aquest treball hem partit del principi que un sistema d'extracció automàtica de candidats a unitats de significació especialitzada (SEACUSE) no pot abstreure's de les necessitats de les activitats professionals per als quals s'utilitza; dit altrament, no pot ometre la pertinència o no pertinència de les USE per a una activitat concreta. En aquesta lògica, l'objectiu d'aquest capítol és dissenyar un SEACUSE que s'adeqüi a les necessitats que requereix cada tasca professional diferent i que generi uns resultats ajustats a aquesta finalitat.

A continuació presentem les característiques del SEACUSE que proposem, estructurades de la manera següent:

1. Els fonaments teòrics en els quals es basa el sistema d'extracció
2. Les condicions que ha de complir
3. Les fases que ha de seguir
4. Els mòduls que actuen en cada fase i els components que integren cada mòdul
5. El disseny de la maqueta del sistema.

7.1 Fonaments

És sabut que tota aplicació està basada en una sèrie de fonaments teòrics que la validen. El SEACUSE que hem dissenyat es fonamenta en un conjunt de supòsits sobre les unitats d'extracció i sobre les estratègies d'extracció automàtica, que hem anat comentant, analitzant i validant al

[2] De manera que, fins i tot en certes ocasions, és més eficient realitzar el procés de

llarg de tot el treball i que, bàsicament, es resumeixen en els punts següents:

a) L'objecte d'extracció és la unitat de significació especialitzada (USE), que no es redueix a la UT.

b) Totes les USE poden ser objecte d'extracció d'un SEACUSE, però no tots els professionals tenen les mateixes necessitats lingüístiques davant d'un text especialitzat. Així, l'usuari que fa servir un SEACUSE no està interessat en totes les USE del text, sinó només en les **USE pertinents** d'acord amb la seva activitat professional. I en aquest sentit cal tenir present que les USE del text i les USE pertinents del text per a una activitat no solen coincidir.

c) La detecció i extracció de les USE d'un text no pot limitar-se a un sol criteri, si es proposa de reduir el silenci i el soroll; ni tampoc no es pot reduir a criteris exclusivament formals. Les estratègies d'extracció han de basar-se en criteris diversos: gramaticals (morfològics i sintàctics), semàntics, contextuals, pragmàtics i tipogràfics. I si els resultats han de servir per incrementar els coneixements que tenim sobre el llenguatge, els criteris de detecció han de ser bàsicament de naturalesa lingüística.

d) Cada tipus d'USE, per la seva naturalesa, requereix un conjunt d'estratègies de reconeixement i d'extracció determinat.

7.2 Condicions d'un SEACUSE

Per tal, doncs, que un SEACUSE sigui útil a les necessitats professionals per a les quals s'utilitza ha de complir les condicions següents:

---

buidatge manualment des del començament, que utilitzar un sistema automàtic.

a) exhaustivitat
b) precisió
c) adequació
d) polivalència
e) rapidesa
f) simplicitat
g) interacció
h) integració.

En primer lloc, el SEACUSE ha de ser al màxim d'**exhaustiu** quant a l'extracció de les USE, en el sentit que ha de recollir totes les USE del text i no limitar-se només a les UTP, com fan la majoria de sistemes clàssics.

En segon lloc, ha de ser **precís** pel que fa a l'extracció de les USE, de manera que generi el mínim soroll. D'acord amb aquesta condició, no pot centrar-se només en un sol aspecte ni tampoc en aspectes exclusivament formals, com solen fer la majoria de sistemes.

En tercer lloc i des del punt de vista pragmàtic, la selecció d'USE ha de ser tan **adequada** com sigui possible a les necessitats professionals. Per donar raó de les diferents funcions, el sistema ha de ser un sistema **multifuncional,** de forma que serveixi a les diferents finalitats professionals.

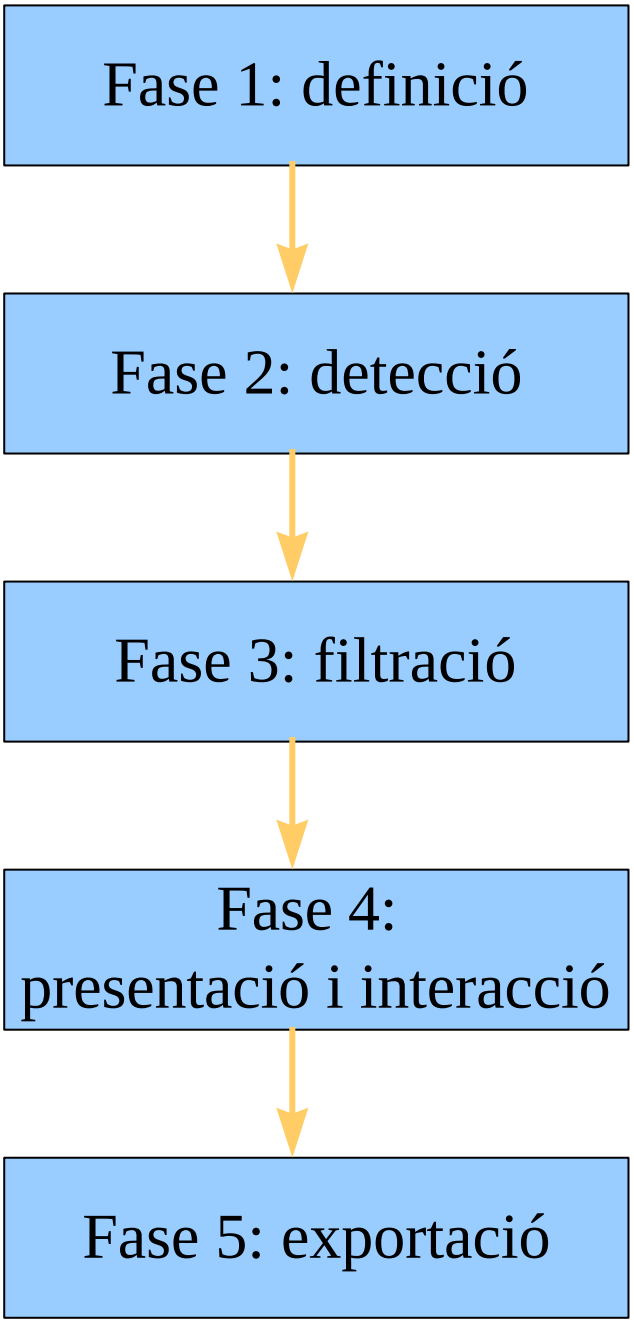
Des del punt de vista informàtic, el sistema ha de ser **ràpid** en l'execució del procés de detecció i **simple** d'utilitzar. El SEACUSE també ha de permetre la **interacció** amb l'usuari per facilitar una selecció d'unitats al més pertinent possible. En aquest sentit, l'usuari hauria de poder manipular els resultats i demanar informacions complementàries sobre les

USE generades pel sistema, i, fins i tot, accedir a les USE que s'han silenciat per a la tasca a la qual s'aplica l'extractor.

Finalment, aquest sistema s'ha de poder **integrar** a altres recursos, com diccionaris, enciclopèdies, bancs terminològics, bases de dades, etc., que facilitin la tria definitiva d'unitats; ha de permetre també l'exportació dels resultats a altres aplicacions (sistemes de traducció automàtica, bases de coneixement especialitzat, bases de dades especialitzades, sistemes d'elaboració de definicions, sistemes d'anàlisi del discurs especialitzat, etc.); i també ha de poder utilitzar els resultats per refinar els seus filtres lèxics.

7.3 Fases metodològiques

El procés d'extracció i selecció de les USE pertinents per part d'un SEACUSE pot seguir les cinc fases següents:



7.3.1 Fase 1: definició

La primera fase consisteix en una presa de decisions que l'usuari ha de poder precisar en relació amb tres paràmetres:

- la font de buidatge, és a dir el corpus textual que es vol analitzar
- el context, és a dir el domini especialitzat en el qual s'emmarca el text de buidatge
- la finalitat, és a dir l'activitat professional per a la qual es requereix la selecció de les USE.

En aquesta fase de definició l'usuari selecciona un text del qual defineix l'àmbit temàtic i la llengua en la qual està escrit[3].

Els textos a què s'aplica aquest SEACUSE han d'estar en format electrònic i han d'haver estat tractats estructuralment i morfològicament. El marcatge estructural consisteix a establir les divisions que componen el document, els títols, els subtítols, les característiques tipogràfiques, les notes a peu de pàgina o a final de text, les llistes, les taules, les figures, les fórmules i els fragments redactats en una altra llengua. El processament lingüístic del text consta de tres etapes: el preprocés, l'anàlisi morfològica i la desambiguació.

En el preprocés es tracten les unitats lèxiques que, per les seves característiques lingüístiques, admeten una detecció automàtica prèvia a l'anàlisi morfològica, com els noms propis, les abreviacions, els identificadors, etc. Totes les unitats que no han estat tractades en la fase del preprocés passen per l'analitzador morfològic i queden etiquetades i lematitzades gramaticalment. Finalment, en la fase de desambiguació dels mots sobreetiquetats se selecciona una etiqueta gramatical entre les proposades, a partir de criteris lingüístics o estadístics.

[3] En aquest estudi, hem treballat sobre textos de medicina escrits en català; encara que pensem que, amb els ajustaments oportuns, el model seria vàlid per a altres llengües romàniques i per a altres àmbits especialitzats; i, en el cas de poder-se utilitzar per a més d'una llengua, caldria també en aquesta primera fase seleccionar-la.

Al costat del text i del domini a què pertany, l'usuari també estableix l'activitat professional per a la qual vol fer el buidatge, concretament selecciona una de les quatre possibilitats següents, ja presentades en el capítol anterior (6.7):

- la transmissió de coneixement
- la indexació de textos
- la traducció
- l'elaboració de diccionaris.

7.3.2 Fase 2: detecció

En la segona fase, el SEACUSE realitza una detecció i extracció exhaustives de totes les USE que conté el text, tant de les USE lingüístiques com de les no lingüístiques. Al final de l'aplicació d'aquesta fase, el sistema hauria d'arribar a una llista d'USE, classificades per tipus. En aquesta etapa el sistema encara no té en compte els requisits de l'activitat per a la qual es realitza la selecció i, per això els resultats són interns al sistema.

7.3.3 Fase 3: filtració

En la tercera fase el sistema filtra les USE seleccionades en la fase anterior segons el perfil escollit en la primera fase. Així, al final d'aquesta fase el sistema activa UNA de les quatre seleccions possibles d'USE, cada una corresponent a UN dels quatre perfils professionals establerts i s'obté un conjunt d'USE pertinents per al perfil seleccionat i un conjunt d'USE no pertinents.

7.3.4 Fase 4: presentació i interacció

En aquesta fase, el sistema presenta, per defecte a l'usuari, la selecció d'USE segons el perfil seleccionat i li ofereix la possibilitat d'accedir a les USE no seleccionades. El sistema també li permet accedir als contextos, freqüència d'ús, concordances i família de paraules de qualsevol de les USE detectades. En aquesta fase, l'usuari ha de poder interaccionar amb el programa de manera que, d'una banda, pugui manipular activament la selecció inicial per eliminar qualsevol unitat que no consideri pertinent, completar-la a través de la consulta d'altres unitats en un principi no pertinents i demanar al sistema informacions complementàries de cada una de les USE per acabar d'ajustar la tria a les seves necessitats professionals.

També ha de poder consultar altres recursos connectats al sistema (diccionaris, tesaurus, bases de dades terminològiques, etc.) que li permetin refinar la selecció.

7.3.5 Fase 5: exportació

Finalment, en l'última fase, el sistema permet exportar els resultats obtinguts a d'altres aplicacions, refinar els diccionaris del propi sistema i/o, simplement, imprimir els resultats amb informació complementària.

7.4 Mòduls

Per executar els processos que acabem de descriure, concebem un sistema integrat per cinc mòduls, cada un corresponent a una de les cinc fases establertes:

1. Mòdul de definició → fase 1 de definició
2. Mòdul de detecció → fase 2 de detecció
3. Mòdul de restricció → fase 3 de filtració
4. Mòdul de presentació i interacció → fase 4 de presentació i interacció
5. Mòdul d'exportació → fase 5 d'exportació.

7.4.1 Mòdul de definició

El mòdul de definició, format pel component DECISICIÓ, permet a través d'un conjunt de menús definir el perfil del buidatge a partir de tres elements:

- el corpus textual concret al qual s'aplicarà el sistema d'extracció
- l'àmbit especialitzat a què pertany el corpus
- l'activitat professional per al qual es necessita la selecció[4].

7.4.2 Mòdul de detecció

El mòdul de detecció està integrat per tres components principals[5]:

a) component FILTRES
b) component PROGRAMES
c) component EINES.

[4] I, en el cas que calgués, també s'hauria de definir la llengua del corpus de buidatge.

Aquest mòdul és el més complex des del punt de vista lingüístic perquè és el que permet la detecció i extracció de totes les USE del corpus:

a) El component FILTRES està constituï t per tres mecanismes de filtratge:

    a.1) un filtre lèxic
    a.2) un filtre morfosintàctic
    a.3) un filtre morfosemàntic.

    Aquests filtres serveixen per detectar les USE lingüístiques i s'apliquen de manera discriminada segons cada tipus d'unitat, com hem proposat en el capítol cinquè (5.3).

    a.1) El submòdul FILTRE LÈXIC està constituï t per un conjunt de sis diccionaris, quatre dels quals actuen de filtres positius (serveixen per confirmar el caràcter especialitzat d'una unitat) i els altres dos de filtres negatius (es fan servir per saber que una unitat no és especialitzada):

        A. Diccionaris que actuen de filtres positius:

            1. Diccionari de termes simples
            2. Diccionari de formants cultes
            3. Diccionari de sigles freqüents
            4. Diccionari de símbols freqüents

        B. Diccionaris que actuen de filtres negatius:

            1. Diccionari de paratermes

[5] Per a informació més detallada d'aquests components del SEACUSE vegeu l'apartat 5.3 del capítol cinquè.

2. Diccionari de mots que sempre generen soroll: quantitatius, organitzadors del discurs.

Els diccionaris de termes simples i formants cultes serveixen per detectar les USE monolèxiques i polilèxiques; el diccionari de sigles per filtrar les sigles; el diccionari de símbols per extreure símbols i fórmules, i els dos diccionaris que actuen de filtres negatius per complementar l'extracció de les USE polilèxiques.

a.2) El subcomponent FILTRE MORFOSINTÀCTIC està compost d'un conjunt de patrons positius i negatius que permeten confirmar o rebutjar el caràcter especialitzat de les unitats sintagmàtiques. La finalitat d'aquest filtre és la detecció de les USE polilèxiques. Com que el SEACUSE utilitza filtres negatius és interessant que compti, a més, amb una eina de segmentació de sintagmes per poder examinar cada unitat del corpus i d'aquesta manera detectar totes les USE del text, també les superposades.

a.3) Finalment, el subcomponent FILTRE MORFOSEMÀNTIC permet desambiguar certes combinacions sintagmàtiques que si només es tenen en compte aspectes lèxics, morfològics i sintàctics són ambigües ja que no es pot saber si es tracten d'USE pertinents, d'unitats discursives o d'UFE. Per tant, en un principi, proposem que el SEACUSE utilitzi els esquemes morfosemàntics de restricció només quan els altres recursos lingüístics són insuficients per desambiguar una seqüència polilèxica. Aquesta situació, com hem vist al capítol cinquè, es dóna en dues combinacions: d'una banda, en els segments d'estructura $[N[A]_{SAdj}]_{SN}$ en què el nom és terminològic, però l'adjectiu no; i, de l'altra, en les seqüències amb estructura $[N\ [de\ (art)\ [N]]_{SPrep}]_{SN}$ en què el primer nom de la seqüència és terminològic.

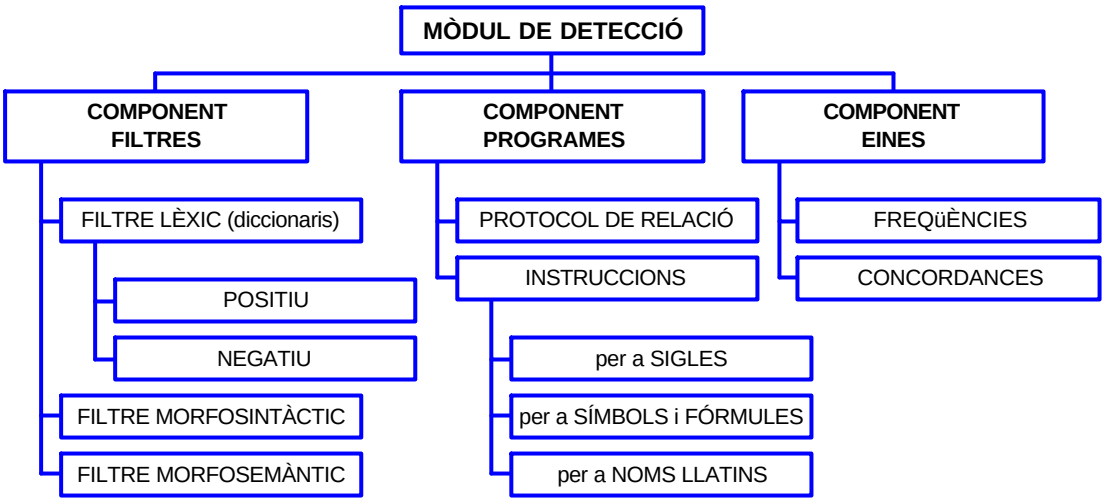
b) El component PROGRAMES està integrat per un protocol basat, principalment, en característiques morfològiques de les USE lingüístiques, que relaciona les USE amb una mateixa base; i, per un conjunt d'instruccions molt simples fonamentades en condicions tipogràfiques i discursives per a uns tipus d'unitats molt específiques (sigles, símbols, fórmules i noms científics en llatí).

c) Finalment, el component EINES està format per una eina d'extracció de freqüències i de concordances d'ús de les USE. Aquestes informacions serveixen per completar la informació i prendre una decisió en la selecció definitiva de les unitats.

Com hem assumit en el capítol cinquè, cada tipus d'USE requereix unes estratègies diferents de reconeixement; així, per exemple, per extreure les USE simples només cal aplicar el FILTRE LÈXIC; però, en canvi, per detectar les USE derivades cal aplicar el PROTOCOL de RELACIONS i el FILTRE LÈXIC; i, per extreure les unitats polilèxiques cal fer servir primer el FILTRE MORFOSINTÀCTIC, després el PROTOCOL de RELACIONS i el FILTRE LÈXIC i, en alguns casos, fins i tot el FILTRE MORFOSEMÀNTIC. En el capítol cinquè, hem establert més precisament les estratègies més convenients per a l'extracció de cada tipus d'unitat.

Al final de l'aplicació d'aquest mòdul, el sistema hauria de proporcionar totes les USE del text, classificades per tipus, sense cap mena de restricció funcional. Sobre aquesta selecció exhaustiva i interna s'aplicaria en la fase següent el mòdul RESTRICCIÓ en què les USE quedarien filtrades pel perfil d'activitat professional seleccionat per l'usuari.

La figura següent esquematitza els components que integren el mòdul més complex de l'extractor:



7.4.3 Mòdul de restricció

Una vegada detectades totes les USE que conté el text i després de classificar-les segons els tipus, el sistema aplica, com avançàvem, el component PERFILS per restringir la selecció inicial. Aquesta selecció es realitza d'acord amb els perfils de necessitats professionals que hem establert en l'apartat 6.7 del capítol anterior.

7.4.4 Mòdul de presentació i interacció

El mòdul de presentació del sistema consta d'un component GENERACIÓ, que proporciona la selecció de les unitats corresponents al perfil professional i facilita l'accés a la selecció de totes les unitats en un principi descartades, per si l'usuari vol rectificar la selecció; i d'un component d'INTERACCIÓ que permet a l'usuari elaborar la selecció d'USE definitiva.

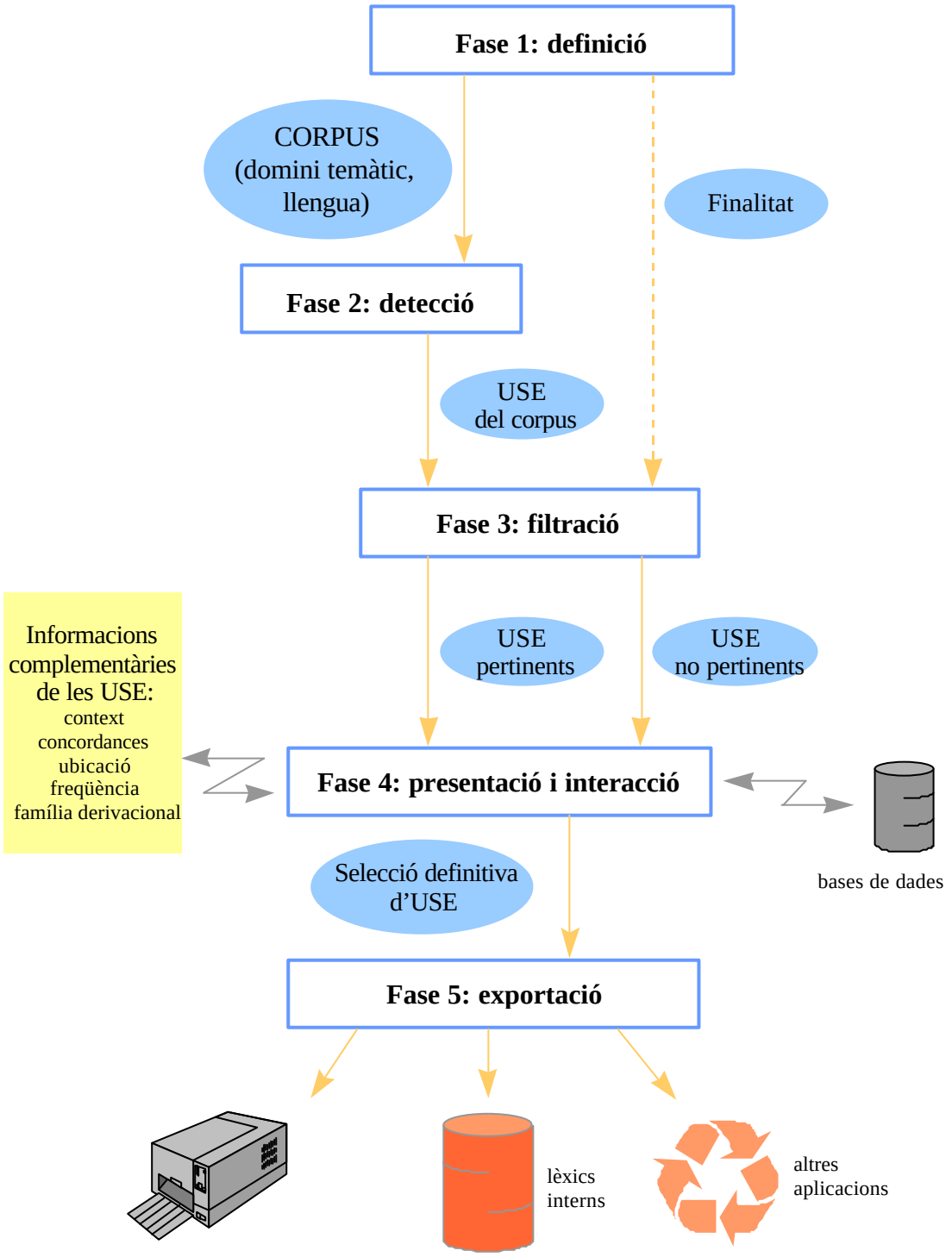
En aquest mòdul el sistema permet a l'usuari l'accés, d'una banda, a informacions complementàries de les USE seleccionades del text, i, de l'altra, a altres eines informàtiques com, per exemple, bases de dades terminològiques, diccionaris, bases de coneixement especialitzat, etc.

7.4.5 Mòdul d'exportació

Finalment, el mòdul d'exportació consta d'un component EXPORTACIÓ que ha de permetre exportar les dades per reutilitzar-les en altres aplicacions informàtiques i per autoalimentar els seus propis diccionaris.

7.5 Maqueta

A continuació mostrem una maqueta del SEACUSE que proposem:


Fase 1: definició
CORPUS
(domini temàtic,
llengua)
Finalitat
Fase 2: detecció
USE
del corpus
Fase 3: filtració
Informacions
complementàries
de les USE:
context
concordances
ubicació
freqüència
família derivacional
USE
pertinents
USE
no pertinents
Fase 4: presentació i interacció
bases de dades
Selecció definitiva
d'USE
Fase 5: exportació
lèxics
interns
altres
aplicacions

Tot seguit hem esquematitzat els prototipus de la selecció d'USE corresponents a cada un dels quatre perfils establerts que el SEACUSE extrauria després d'aplicar el mòdul de restricció al final de la fase 3 i que presentaria a l'usuari en la fase 4. A través d'un sistema de possibilitats activades i desactivades es permet l'accés, d'una banda, a la selecció d'USE pertinents per a un perfil determinat i a les informacions complementàries sobre aquestes unitats; i, de l'altra, a les USE i a les informacions complementàries en un principi no pertinents per aquest perfil:

a) perfil 1: transmissió del coneixement
b) perfil 2: indexació de textos especialitzats[6]
c) perfil 3: preparació de traduccions especialitzades
d) perfil 4: elaboració de terminografies especialitzades.

[6] Prèviament a la generació de la selecció d'USE corresponent al perfil 2, el programa pregunta a l'usuari la freqüència a partir de la qual vol que es filtrin les unitats. Aquesta pot ser una freqüència per defecte (en relació amb l'extensió del corpus) o bé una freqüència escollida per l'usuari.

7.6 Validació

Per tal de validar a tall d'experiment l'extractor que hem proposat, hem aplicat **manualment** el **mòdul de detecció** —que correspon a la fase central del sistema— a un petit corpus textual. Es tracta d'un fragment breu de 510 ocurrències sobre la hipertensió arterial pulmonar, extret de l'*Enciclopèdia de Medicina i Salut* (1990) que reproduï m a continuació:

Hipertensió arterial pulmonar

Definició

S'anomena hipertensió arterial pulmonar una elevació persistent de la pressió interna de les artèries pulmonars, que pot ésser causada a diverses malalties del cor, dels pulmons o de les mateixes artèries pulmonars. Aquest trastorn provoca un excés d'esforç ventricular dret, capaç d'originar una insuficiència cardíaca amb el pas del temps.
Quan la hipertensió arterial pulmonar és deguda a malalties pulmonars provoca alteracions en el ventricle dret, especialment hipertròfia, i dóna pas a l'anomenat cor pulmonar.

Causes i tipus

Quan la hipertensió arterial és de causa desconeguda s'anomena hipetensió pulmonar primitiva o idiopàtica, i quan és provocada per altres cirncumstàncies o malalties rep el nom d'hipertensió pulmonar secundària.

La hipertensió pulmonar primitiva o idiopàtica és un trastorn poc freqüent, d'una incidència aproximadament quatre vegades superior en les dones, i que es presenta en general entre 30 anys i 50. Per bé que hom ignora la causa d'aquest tipus d'hipertensió arterial pulmonar, es considera que el mecanisme que la provoca és l'existència de moviments de contracció o espasmes repetits i prolongats de les artèries pulmonars.

> D'altra banda, la hipertensió pulmonar secundària pot ésser deguda a malalties cardíaques o pulmonars. Entre les primeres, les més freqüents són afeccions de les vàlvules de les cavitats cardíaques esquerres, com ara estenosi mitral i aòrtica, miocardiopatia hipertròfica i pericarditis contrictiva. En tots aquests casos hi ha un increment del volum de sang i la pressió interna de les cavitats cardíaques esquerres, que es transmet retrògradament a les venes i els capil·lars pulmonars. A conseqüència d'això, un volum variable de sang tendeix a acumular-se en el teixit pulmonar i ocasiona una congestió pulmonar. En aquest cas, cal incrementar la pressió de les artèries pulmonars per tal de mantenir una perfusió sanguínia adequada. D'altra banda, hi ha algunes cardiopaties congènites (comunicació interauricular, comunicació interventricular o persitència del conducte arterial) que es caracteritzen per un pas anormal de la sang des de les cavitats cardíaques esquerres cap a les cavitats cardíaques dretes, és a dir que s'estableix un curt circuit d'esquerra a dreta. En aquests casos, el ventricle dret s'ha de contreure més intensament del que és habitual per a fer arribar un volum de sang superior al normal a les artèries pulmonars i també s'esdevé un increment de llur pressió interna.
>
> Entre les malalties pulmonars que poden provocar hipertensió arterial pulmonar, destaquen les anomenades malalties pulmonars obstructives cròniques com l'asma, la bronquitis crònica i l'emfisema pulmonar. En aquests casos hi ha un increment difús del teixit pulmonar que ofereix una resitència superior al pas de la sang per l'interior dels capil·lars. Així, la pressió interna de les artèries pulmonars s'ha d'incrementar per tal de mantenir la circulació. Igualment, hi ha trastorns pulmonars aguts, com ara embòlia o processos infecciosos, que provoquen un augment sobtat de la resistència circulatòria pulmonar i, per tant, una hipertensió arterial pulmonar.
>
> [Fundació Enciclopèdia Catalana, 1990: 134]

Es tracta d'un text molt breu, les USE del qual són de naturalesa lingüística, no hi ha sigles i, des del punt de vista gramatical, no hi ha USE adverbials.

El mòdul de detecció extreu d'aquest text d'una banda les USE monolèxiques i de l'altra les polilèxiques, classificades en els tipus i categories següents:

1. USE monolèxiques

1.1 USE monolèxiques simples. Per detectar-les, el SEACUSE utilitza el FILTRE LÈXIC d'USE simples de medicina:

> noms: *pressió, artèria, malaltia, cor, pulmó, trastorn, ventricle, espasme, vàvula, sang, capil·lar, asma, vena*
> verbs: *contreure*

1.2 USE monolèxiques derivades. En aquest cas, el SEACUSE utilitza el protocol de relació de bases lèxiques i els filtres lèxics d'USE simples i de formants cultes:

> noms: *alteració, contracció, afecció, congestió, perfusió, circulació*
> adjectius: *arterial, pulmonar, aòrtic, hipertròfic, circulatori*

1.3 USE monolèxiques compostes. El SEACUSE utilitza el filtre LÈXIC de formants cultes per detectar aquest tipus d'unitats:

> noms: *hipertensió, hipertròfia, estenosi, miocardipatia, pericarditis, bronquitis, emfisema, embòlia*
> adjectius: *cardíac, idiopàtic*

2. USE polilèxiques

Per detectar les **USE polilèxiques nominals** primer aplica el FILTRE MORFOSINTÀCTIC i troba les unitats següents:

hipertensió arterial pulmonar; elevació persitent de la pressió interna de les artèries pulmonars; malalties del cor; artèries pulmonars; excés d'esforç ventricular dret; insuficiència cardíaca; pas del temps; malalties pulmonars; ventricle dret; cor pulmonar; hipertensió arterial; causa desconeguda; hipertensió pulmonar primitiva; nom d'hipetensió pulmonar secundària; vegades superior; tipus d'hipertensió arterial pulmonar; existència de moviments de contracció; espasmes repetits; hipertensió pulmonar secundària; malalties cardíaques; afeccions de les vàlvules de les cavitats cardíaques esquerres; estenosi mitral; miocardipoatia hipertròfica; pericarditis constrictiva; increment del volum de sang; pressió interna de les cavitats cardíaques esquerres; capil·lars pulmonars; volum variable de sang; teixit pulmonar; congestió pulmonar; pressió de les artèries pulmonars; perfusió sanguínia adequada; cardiopaties congènites; comunicació interauricular; comunicació interauricular; persitència del conducte arterial; pas anormal de la sang; cavitats cardíaques esquerres; cavitats cardíaques dretes; curt circuit; ventricle dret; volum de sang; pressió interna; malalties pulmonars obstructiva cròniques; bronquitis crònica; emfisema pulmonar; engruiximent difús del teixit pulmonar; pas de la sang; interior dels capil·lars; pressió interna de les artèries pulmonars; trastorns pulmonars aguts; processos infeciosos; resistència circulatòria pulmonar

A continuació, el SEACUSE classifica aquestes seqüències es dos tipus:

- seqüències d'estructura $[N[SAdj]]_{SN}$
- seqüències d'estructura $[N[Sprep]]_{SN}$

Seguidament, el sistema aplica les condicions establertes en el capítol cinquè que determinen si una combinació sintagmàtica que correspon a un patró estructural propi d'una USE, ho és segons les característiques dels seus components.

Sobre l'estructura $[N[SAdj]]_{SN}$

a) Si el nucli i el complement de la seqüència són USE, es tracta d'una UTP. En el text, compleixen aquesta condició les unitats següents:

hipertensió arterial pulmonar, artèries pulmonars, malalties pulmonars, congestió pulmonar, cor pulmonar; hipertensió arterial, malalties cardíaques, estenosi mitral; miocardipoatia hipertròfica, capil·lars pulmonars, teixit pulmonar, cardiopaties congènites, emfisema pulmonar

b) Si el nucli de la seqüència no és una USE i el complement sí que ho és, es tracta d'una UTP, com exemplifiquen els casos següents:

insuficiència cardíaca, comunicació interauricular, processos infeciosos, resistència circulatòria pulmonar

c) Si el nucli i el complement de la seqüència no són USE, es tracta d'una UD o d'una UL, que en tots dos casos causen soroll i el sistema les ha d'eliminar. Compleixen aquesta premissa:

causa desconeguda, vegades superior, curt circuit

d) Si el nucli de la seqüència és una USE i el complement no ho és, el sistema ha d'aplicar filtres morfosemàntics. Les unitats del text que responen a aquesta condició són:

ventricle dret, hipertensió pulmonar primitiva, espasmes repetits, hipertensió pulmonar secundària, pericarditis constrictiva, perfusió sanguínia adequada, cavitats cardíaques esquerres, cavitats cardíaques dretes, pressió interna, malalties pulmonars obstructiva cròniques, bronquitis crònica, trastorns pulmonars aguts

I el filtres morfosemàntics que el sistema aplica per a aquestes unitats de d) són:

[NOM A[ADJ6]$_{Sadj}$] $_{SN}$ = UTP

espasmes repetits, bronquitis crònica, malalties pulmonars obstructiva cròniques

[NOM A[ADJ7]$_{Sadj}$] $_{SN}$ = UTP

hipertensió pulmonar primitiva, hipertensió pulmonar secundària, trastorns pulmonars aguts

[NOM A[ADJ8]$_{Sadj}$] SN = UTP

pericarditis constrictiva, malalties pulmonars obstructiva cròniques

[NOM B [ADJ1]$_{Sadj}$] SN = UTP

ventricle dret, cavitats cardíaques esquerres, cavitats cardíaques dretes, pressió interna

[NOM A[ADJ16]$_{Sadj}$] SN = UTP

perfusió sanguínia adequada

Sobre l'estructura [N[Sprep]]$_{SN}$

a) Si el nucli de la seqüència és un nom deverbal complementat per un sintagma preposicional el nucli del qual és una USE, es tracta d'una UFE[7]. Les unitats del text que responen a aquesta condició són:

elevació persitent de la pressió interna de les artèries pulmonars, increment del volum de sang, persitència del conducte arterial, pas anormal de la sang, engruiximent difús del teixit pulmonar; pas de la sang existència de moviments de contracció, afeccions de les vàlvules de les cavitats cardíaques esquerres; pressió interna de les cavitats cardíaques esquerres, pressió de les artèries pulmonars, pressió interna de les artèries pulmonars

b) Si el nucli de la seqüència és un paraterme o un nom quantitatiu, cal eliminar el nucli i analitzar el complement perquè la majoria de vegades es tractarà d'una USE. En el text hi ha tres unitats que compleixen aquesta condició:

[7] En aquests casos el sistema també analitzarà el o els SN que conté el sintagma preposicional per proposar-los com a USE pertinents o no. Els únics SN que no havien estat documentats aï lladament són els dos següents: *conducte arterial, moviments de contracció.*

excés *d'esforç ventricular dret;* ***nom*** *d'hipertensió pulmonar secundària;* ***tipus*** *d'hipertensió arterial pulmonar*[8]

c) Si cap dels dos noms de la seqüència són USE, es tracta d'una UD o UL sense interès especialitzat que el sistema ha d'eliminar. Una ocurrència del text respon a aquesta condició:

pas del temps

d) Si el primer nom de la seqüència no és un terme, però en canvi el segon sí i, a més, el primer nom tampoc no és un paraterme ni un quantitatiu ni un organitzador de l'estructura discursiva, es tracta d'una UT, com exemplifiquen les tres unitats següents:

interior dels capil·lars, volum de la sang, volum variable de sang[9]

e) Si els dos noms de la seqüència són USE, cal aplicar filtres morfosemàntics. En el text només hi ha una unitat a la qual cal aplicar un filtres morfosintàctic per desambiguar-la: *malalties del cor* . En aquest cas l'extractor aplica el filtre morfosemàntic següent:

[NOM A1 [de [N2 part del cos humà afectada per N1] SN]SPrep]SN = UTP

A continuació, el SEACUSE genera la llista d'USE pertinents, que per aquest text és la següent:

UT: *pressió, artèria, malaltia, cor, pulmó, trastorn, ventricle, espasme, vàlvula, sang, capil·lar, asma, vena; alteració, contracció, afecció, congestió, perfusió, circulació; hipertensió, hipertròfia, estenosi, miocardipatia, pericarditis, bronquitis, emfisema, embòlia*

UTP: *hipertensió arterial pulmonar, artèries pulmonars, malalties pulmonars, congestió pulmonar, cor pulmonar; hipertensió arterial, malalties cardíaques, estenosi mitral; miocardipoatia hipertròfica, capil·lars pulmonars, teixit*

[8] En aquest cas, el sistema incorpora l'USE *esforç ventricular dret*, perquè tant *hipertensió pulmonar secundària* com *hipertensió pulmonar* ja han estat documentades.

[9] Recordem que, com es compleixen en aquests casos, les dades mostren que la unitat resultant sol denominar parts del cos humà o parts de parts del cos humà en què el complement està determinat, ja que la determinació remarca el caràcter de part única de la unitat resultant.

pulmonar, cardiopaties congènites, emfisema pulmonar; insuficiència cardíaca, comunicació interauricular, processos infecciosos, resistència circulatòria pulmonar; ventricle dret, hipertensió pulmonar primitiva, espasmes repetits, hipertensió pulmonar secundària, pericarditis constrictiva, perfusió sanguínia adequada, cavitats cardíaques esquerres, cavitats cardíaques dretes, pressió interna, malalties pulmonars obstructiva cròniques, bronquitis crònica, trastorns pulmonars aguts; conducte arterial, moviments de contracció; esforç ventricular dret; interior dels capil·lars, volum de la sang, volum variable de sang; malalties del cor

USE verbals: *contreure*

USE adjectives: *arterial, pulmonar,aòrtic, hipertròfic, circulatori; cardíac, idiopàtic*

UFE nominals: *elevació persistent de la pressió interna de les artèries pulmonars, increment del volum de sang, persitència del conducte arterial, pas anormal de la sang, engruiximent difús del teixit pulmonar, pas de la sang, existència de moviments de contracció, afeccions de les vàlvules de les cavitats cardíaques esquerres, pressió interna de les cavitats cardíaques esquerres, pressió de les artèries pulmonars, pressió interna de les artèries pulmonars.*

A aquesta llista el SEACUSE aplicaria el component de restricció, que adequaria la selecció de les unitats retingudes a les necessitats professionals d'una tasca concreta.

Silenci i soroll

Després d'aquesta simulació podem observar que, tot i que el sistema ha detectat gairebé totes les USE que contenia el text roman una petita quantitat de silenci.

Així, l'extractor ha deixat sense reconèixer algunes USE amagades per anàfora discursiva per tal com no disposa dels mecanismes adequats per fer-ho:

malalties del cor, dels pulmons
hipertensió pulmonar primitiva o idiopàtica
malalties cardíaques o pulmonars

estenosi mitral i aòrtica

Per detectar aquest tipus d'unitats cal, com hem comentat en capítols anteriors, treballar amb corpus molt més grans o habilitar estratègies alternatives de desambiguació, com per exemple l'ús d'un analitzador sintàctic, que no hem plantejat en aquest treball.

El sistema tampoc no reconeix les UFE verbals tipus *incrementar la pressió de les artèries pulmonars* o les UFE adverbials, tot i que d'aquestes últimes no n'apareixen en aquest text tan breu.

Pel que fa al soroll, és cert que per a aquest text el sistema no ha produït soroll, però de fet el soroll real que el sistema genera depèn de l'abast dels filtres lèxics, morfosintàctics i morfosemàntics que utilitzi.

Així doncs, només podrem valorar realment el silenci i sobretot el soroll del sistema quan aquest estigui implementat i es pugui aplicar a corpus molt més grans.

7.7 Conclusió

L'objectiu d'aquest capítol ha estat proposar un model de SEACUSE basat en els principis i fonaments que hem establert al llarg del treball. Hem partit de la base que aquest SEACUSE havia de complir les característiques següents: havia de ser **exhaustiu** i **precís** quant a l'extracció de les USE, **adequat** a les necessitats professionals per a les quals s'usa, **multifuncional,** perquè servís per a més d'una activitat professional, **ràpid i simple** d'executar, **interactiu** per facilitar una

selecció més pertinent i **integrable** en altres eines informàtiques que permetin la importació i l'exportació de dades.

Després d'establir les condicions del SEACUSE, hem definit les cinc fases del procés de detecció d'USE i en cadascuna un mòdul específic. Hem establert els components i subcomponents que integrarien cada un d'aquests cinc mòduls, que permetrien a l'usuari disposar de la selecció d'USE més adequada a les seves necessitats. Finalment, hem recollit aquestes propostes en una maqueta que il·lustra el funcionament del SEACUSE.

Seguidament, hem simulat com actuaria l'extractor en la **fase de detecció d'USE** aplicant **manualment** el mòdul de detecció a un text. La conclusió principal que hem extret d'aquesta prova experimental és que el SEACUSE proposat dóna resultats més exhaustius i precisos que els extractors existents, i l'aplicació posterior del mòdul de filtres permet que els resultats siguin més adequats a les necessitats professionals. Tot i això, caldrà validar aquests resultats en un corpus més representatiu i de manera automàtica, un cop s'hagi construï t l'aplicació informàtica[10].

Els punts més febles del sistema són segurament la detecció de les USE amagades per anàfora discursiva i la detecció de les UFE verbals i adverbials i també la selecció d'UTP que contenen un adjectiu no especialitzat amb el sufix *-ble* o *-ada* . Per als dos primers casos caldrà pensar en la incorporació d'un analitzador sintàctic al sistema d'extracció. En el cas dels adjectius debervals caldrà afinar els filtres morfosemàntics.

[10] L'elaboració informàtica d'aquest SEACUSE i la seva aplicació a un corpus gran constitueix la tesi de J. Vivaldi en curs d'elaboració, tesi codirigida per M. Teresa Cabré i Horacio Rodríguez.

7.8 Recapitulació

En aquest capítol hem dissenyat un sistema d'extracció automàtica de candidats a unitats de significació especialitzada que hem anomenat SEACUSE. Per fer-ho, hem establert els principals fonaments en els quals es basa i les característiques internes del sistema, hem definit les fases d'actuació i els mòduls per executar-les i, finalment, hem establert els components que integren cadascun d'aquests mòduls. Per mostrar l'eficàcia del SEACUSE proposat, hem simulat manualment com actuaria el seu mòdul de detecció aplicat a un text molt breu.

A continuació presentem les conclusions generals d'aquesta tesi, així com també les seves aplicacions i limitacions.

8. CONCLUSIONS

"No existeix res bo ni dolent, el pensament humà és el que ho fa semblar així."

[Shakespeare]

"La mente humana no puede, al mismo tiempo, asumir un fenómeno en su totalidad y con un alto grado de detalle."

[Codina, 1996: 37]

El propòsit final d'aquest treball ha estat dissenyar un model d'un sistema d'extracció automàtica de candidats a unitats de significació especialitzada (SEACUSE), basat en fonaments lingüístics de diferents tipus i adequat a les necessitats d'unes activitats professionals concretes. Per aconseguir aquesta finalitat vam establir uns objectius que hem anat desenvolupant en els diversos capítols d'aquest treball i que se sintetitzen en els tres punts següents:

1. **Analitzar i avaluar el funcionament dels actuals sistemes d'extracció automàtica de candidats a terme (SEACAT)** amb la finalitat de tenir elements per dissenyar un extractor per al català. Aquest objectiu implicava elaborar un estat de la qüestió dels SEACAT i comprovar que les estructures morfosintàctiques de les UTP proposades en el treball de recerca a partir d'un corpus lexicogràfic, són les mateixes si s'apliquen a un corpus textual.

2. **Proposar criteris de reconeixement de les Unitats de Significació Especialitzada (USE)** pertinents per dissenyar un sistema d'extracció automàtica. Aquest objectiu suposava delimitar els diferents tipus d'USE que s'usen en l'àmbit de les

ciències de la salut, proposar elements i estratègies per reconèixer-los i detectar-los, i establir estratègies per eliminar al màxim les unitats no pertinents que solen generar els sistemes d'extracció automàtica.

3. **Dissenyar un model de Sistema d'Extracció Automàtica de Candidats a Unitats de Significació Especialitzada (SEACUSE) adequat a les necessitats professionals de diferents col·lectius d'usuaris**, que implicava, prèviament, mostrar que cada activitat professional requereix unes necessitats diferents quant a les unitats especialitzades d'un text, establir els criteris per determinar la pertinència d'una USE a l'interior de cada activitat professional i proposar els perfils d'USE d'algunes finalitats professionals específiques.

Els resultats de l'anàlisi de les dades en relació amb els tres grans objectius que ens havíem proposat i amb els supòsits dels quals partíem a l'inici del treball, permet establir tres conclusions generals de les quals se'n desprenen d'altres de més específiques:

1. Els SEACATS que existeixen actualment faciliten considerablement el reconeixement de les unitats terminològiques, però encara no són del tot satisfactoris perquè generen una quantitat important de silenci i de soroll; problemes causats, respectivament perquè el seu objecte d'extracció es limita a les UTP, i perquè les estratègies d'extracció que utilitzen no són prou discriminatòries.

2. Per extreure les USE d'un text s'ha de recórrer a estratègies diferents segons els tipus d'unitats, cada una de les quals es basen en diferents aspectes de les unitats (lèxics, morfològics, sintàctics,

semàntics i pragmàtics). Aquestes estratègies s'han d'aplicar discriminadament segons els tipus d'unitat.

3. Les unitats terminològiques (UT) no són les úniques unitats amb significat especialitzat que poden interessar els especialistes i, per tant, l'objecte d'un extractor que pugui satisfer les diferents necessitats professionals ha d'abraçar, en la mida que pugui, totes les USE del text. Ara bé, no totes les USE d'un text són pertinents per a totes les tasques professionals, sinó que la pertinència d'una USE depèn de l'activitat professional. Per això, un extractor ha de ser flexible i proposar per a cada tasca una selecció d'USE adequada.

Al costat d'aquestes conclusions generals, hem arribat a uns resultats més concrets que responen als objectius que ens havíem proposat en cada capítol i que sintetitzem a continuació en set apartats que fan referència :

- als sistemes d'extracció automàtica de terminologia
- a les unitats terminològiques en els textos
- al silenci
- al soroll
- a la proposta de millora d'un SEACAT clàssic
- a les necessitats professionals i a les unitats de significació especialitzada
- al disseny d'un SEACUSE adequat a les necessitats professionals.

1. Els sistemes d'extracció automàtica de terminologia

L'anàlisi de divuit extractors presentada en el primer capítol ens ha servit per evidenciar les seves característiques, però també algunes de les seves limitacions més rellevants[1]:

1. Cap dels SEACAT analitzats resulta completament satisfactori perquè tots generen una quantitat de silenci i soroll que considerem excessiva.
2. L'objecte de la majoria de SEACAT és la unitat terminològica polilèxica (UTP).
3. La majoria dels SEACAT es basen exclusivament en la forma del terme i concretament en patrons morfosintàctics; només un dels sistemes analitzats utilitza informació semàntica.
4. La intervenció de l'usuari al final del procés de detecció és imprescindible per seleccionar les unitats presentades com a candidates.

2. Les unitats terminològiques en els textos

En el capítol segon hem comprovat que els resultats del treball de recerca de 1996, en què havíem establert els patrons estructurals de les UTP del domini de la medicina a través de corpus lexicogràfics, eren també vàlids per als corpus textuals. I amb el contrast dels resultats obtinguts en els buidatges hem constatat també que les estructures morfosintàctiques, d'una banda, són insuficients per detectar totes les unitats que, segons l'especialista, són especialitzadament pertinents, i aquesta és la raó que explica que es generi silenci; i, de l'altra, aquestes estructures no són

[1] Per a conclusions més detallades sobre aquest tema vegeu l'apartat de conclusions 1.7 del primer capítol.

exclusives de les UTP, i aquesta és la raó per la qual es generen falsos candidats a terme.

Així, hem comprovat que els patrons de les UTP són vàlids en el sentit que les estructures de les USE polilèxiques hi són previstes, però si aquests patrons són l'única base d'un extractor, els resultats obtinguts dels buidatges han demostrat que no són ni exhaustius ni satisfactoris. Al soroll i al silenci produï ts per un SEACAT, s'hi afegeix la delimitació incorrecta de certes unitats i la no detecció de les relacions semàntiques entre les unitats del text que permetrien una extracció més ajustada[2].

3. Les limitacions dels SEACAT: silenci

Hem arribat a la conclusió, en el capítol tercer, que un extractor que volgués satisfer les necessitats de l'especialista hauria de tenir en compte totes les USE d'un text especialitzat, independentment de les seves característiques formals i semàntiques, i no només les UTP.

A través del contrast dels buidatges manual i automàtic d'un mateix text, hem pogut comprovar que hi ha dos tipus de silenci en relació amb l'objecte d'extracció dels sistemes clàssics: el silenci intrínsec, és a dir les UTP del text que no es detecten, i el silenci extrínsec, és a dir totes les altres unitats de significació especialitzada del text que no són UTP i que tampoc no es reconeixen, perquè un extractor no es proposa de fer-ho.

Des de la perspectiva de l'usuari, hem vist que cal trobar elements i estratègies per reduir sobretot el silenci extrínsec, ja que el silenci intrínsec no és percebut per a totes les finalitats com un problema i, en algunes situacions, es pot reduir si s'augmenta i es diversifica el corpus.

[2] En l'apartat 2.5 i 2.7 del capítol segon s'expliciten aquestes conclusions amb més detall.

Pel que fa a les causes del silenci, hem observat que l'intrínsec està causat per:

1. Errors en la fase de desambiguació morfològica, prèvia a l'aplicació de l'extractor.
2. Termes superposats.
3. Termes discursivament amagats.

En canvi, el silenci extrínsec està originat per la definició del mateix objecte de detecció, ja que els sistemes clàssics d'extracció se solen limitar a la detecció de les UTP i ometen la resta d'USE, ja siguin termes monolèxics o altres unitats especialitzades.

Finalment, hem vist que el silenci que actualment ocasiona aquesta diversitat d'unitats amb interès especialitzat es pot reduir si es recorre, a més dels aspectes estructurals de les unitats, als formants grecollatins, a l'establiment de mecanismes que relacionin les paraules que comparteixen una mateixa base, a les regles internes de certes nomenclatures científiques, a determinats afixos i a la consideració de mecanismes que tinguin en compte la semàntica lèxica[3].

4. Les limitacions dels SEACAT: el soroll

Al costat del silenci, l'altra limitació important dels SEACAT clàssics és el soroll que produeixen, que consisteix en el conjunt d'unitats proposades pel sistema que no responen al seu objecte d'extracció.

[3] En l'apartat 3.2.1.5 i 3.3 del tercer capítol es detallen més aquests resultats.

La conclusió principal que es desprèn de l'estudi del soroll que hem realitzat en el capítol quart és que ni els patrons morfosintàctics ni la freqüència d'ús són elements suficients per discriminar les UT dels textos especialitzats. Així, hem constatat que si usen patrons basats en la forma de les UT els extractors proposen com a candidats tant delimitacions errònies com segments sense interès terminològic. A més, els sistemes que funcionen gairebé exclusivament amb patrons no distingeixen les UTP, de les UFE i de les combinacions especialitzades recurrents.

Després de demostrar que el soroll està originat pel fet que les estructures morfosintàctiques de les UT no són exclusives d'aquest tipus d'unitat, hem proposat altres elements lingüístics complementaris que permetin a un sistema afinar més la selecció d'unitats. Hem centrat l'estudi en les dues estructures més freqüents de les UTP: $[N\ SAdj]_{SN}$ i $[N\ SPrep]_{SN,}$ i hem analitzat els seus nuclis i complements per tal de trobar elements que permetin identificar-les. Hem vist també quines unitats responen a una mateixa estructura i hem establert que per identificar les USE és important conèixer el caràcter especialitzat o no especialitzat dels seus constituents i, també el caràcter eventiu o no eventiu del nucli de la unitat i la condició de nom comú o de nom propi del nucli del complement. Finalment, per a determinades combinacions hem proposat recórrer a la semàntica lèxica, en concret a la relació semàntica que en el context d'una estructura mantenen el nucli i el complement d'una unitat[4].

Paral·lelament, hem establert quines són les unitats que sempre produeixen soroll i hem vist que n'hi ha unes que són vàlides per a tots els dominis d'especialitat i d'altres que només causen soroll dins d'un domini d'especialització concret.

5. Proposta de millora d'un SEACAT clàssic

Una vegada analitzats els SEACAT clàssics i estudiades les seves principals limitacions, hem abordat la qüestió de quin ha de ser l'objecte d'un nou model d'extractor que tingui en compte les limitacions dels sistemes existents i les necessitats reals d'utilització.

Així, després de valorar els buidatges que els especialistes han fet d'uns corpus especialitzats, hem arribat a la conclusió que cal eixamplar l'objecte d'extracció dels SEACAT clàssics perquè en els textos especialitzats, efectivament, hi ha moltes altres unitats, a més de les UTP i de les UT, que també són especialitzadament interessants.

Hem partit d'una unitat de significació especialitzada àmplia (USE) que abraça tant unitats lingüístiques com unitats no lingüístiques, i, dins de les primeres, tant unitats lèxiques com fraseològiques, i, encara dins de les unitats lèxiques inclou noms, verbs, adjectius i adverbis.

A continuació hem proposat elements lingüístics de les USE que podien ser útils per a la seva extracció automàtica. A partir d'aquests elements, hem suggerit estratègies d'extracció assumint la idea que un SEACUSE no pot servir-se d'una única estratègia per detectar les USE pertinents dels textos. Així, hem proposat estratègies basades en diversos elements (lèxics, morfològics, morfosintàctics, morfosemàntics, tipogràfics, distribucionals i estadístics) que s'han d'aplicar discriminadament segons les característiques de cada tipus d'unitat.

En el mateix sentit, hem mostrat que un SEACUSE tampoc no pot reduir-se a utilitzar estratègies que es basin només en la **forma** de les USE,

[4] En l'apartat 4.6 del capítol quart es troben sintetitzades les conclusions que fan

sobretot pel que fa a la recuperació de les USE polilèxiques. Per això, per a algunes seqüències polilèxiques que ocasionen ambigüitat, hem proposat **filtres morfosemàntics** basats en esquemes, que tenen en compte l'estructura d'una seqüència, la classe semàntica de tota la seqüència i la dels constituents que la integren.

Els filtres morfosemàntics que hem suggerit són combinacions lexicosemàntiques d'un nom i un adjectiu o d'un nom i un sintagma preposicional que condueixen a una UTP pertinent o a una combinació recurrent en un àmbit especialitzat determinat. La seva finalitat principal és desfer l'ambigüitat dels segments que, estructuralment, poden donar lloc a tipus d'unitats diferents.

Per poder utilitzar els filtres semàntics, hem vist que cal etiquetar semànticament els corpus textuals i hem evidenciat que encara no disposem d'etiquetaris semàntics prou precisos ni específics de la medicina per al català ni per al castellà, però que en el domini de les ciències de la salut, podíem utilitzar recursos alternatius, com ara un diccionari de formants cultes, per construir una aplicació que funcionés a curt termini.

6. Necessitats professionals i unitats de significació especialitzada

Una vegada establert l'objecte d'extracció, en el capítol sisè hem introduï t un nou paràmetre per dissenyar un extractor: el punt de vista de la pertinència funcional d'una unitat. Així després de qüestionar-nos la validesa de totes les USE del text per a qualsevol activitat professional, a través de l'anàlisi dels buidatges que d'un corpus han fet diferents tipus de professionals, hem confirmat i ratificat la hipòtesi inicial que la

---

referència a l'estudi detallat de les combinacions més freqüents de les UTP.

pertinència d'una USE depèn de les necessitats professionals que genera una activitat concreta.

Verificada aquesta idea prèvia, el segon objectiu que ens hem fixat ha estat establir les necessitats terminològiques de quatre tasques que impliquen l'ús de textos especialitzats (la transmissió del coneixement, la indexació de textos especialitzats, la traducció especialitzada i l'elaboració de diccionaris), tenint en compte els resultats del buidatge d'un text que han realitzat diferents professionals (metges, documentalistes, traductors especialitzats i terminògrafs).

D'una banda, hem comprovat que les necessitats terminològiques que generen aquestes activitats professionals van més enllà de les UT pròpiament dites, tant des del punt de vista de la naturalesa com de la funció o de la categoria gramatical. De l'altra, hem confirmat que no totes les USE que conté un text són pertinents per realitzar una tasca determinada. Hem provat també que cada activitat professional requereix uns tipus d'USE específiques i un nombre d'USE determinades que varien segons els criteris de pertinència. En aquest sentit, hem vist que les diferències entre les necessitats que origina una activitat professional són tant de tipus quantitatiu com qualitatiu[5].

Hem arribat a la conclusió que la noció d'USE és vàlida per a tots els col·lectius professionals i, en canvi, el que varia és la noció d'USE pertinent. Entre aquests dos conceptes hi ha una diferència de restrictivitat condicionada per la funcionalitat. Així, hem mostrat que la funcionalitat de les USE en el si d'una activitat professional condiciona la llista d'unitats que s'han de seleccionar.

[5] En els apartats 6.4 i 6.6 es desenvolupen amb més detall aquests resultats.

Finalment, a partir dels resultats dels buidatges dels professionals, hem establert els perfils de necessitats d'USE que genera cada una de les quatre activitats professionals analitzades i hem vist que determinades activitats requereixen que les USE vagin acompanyades d'informació complementària relativa al context d'ús immediat, la freqüència d'ús i/o la ubicació en el text, amb la finalitat de facilitar la selecció definitiva.

Amb aquest experiment hem mostrat que un SEACUSE no pot funcionar independentment de les necessitats que generen les activitats per a les quals s'utilitza.

7. Disseny d'un SEACUSE adequat a les necessitats professionals

El model de SEACUSE que proposem, basat en els principis i fonaments que hem establert al llarg de la tesi, intenta ser **exhaustiu** i **precís** quant a l'extracció de les USE, **adequat** a les necessitats professionals per a les quals s'usa, **multifuncional**, perquè serveixi per a més d'una activitat professional, **ràpid i simple** d'executar, **interactiu** per facilitar una selecció més pertinent i **integrable** en altres recursos.

Per aconseguir l'objectiu final (reconèixer les USE pertinents i adequades a cada perfil professional), aquest SEACUSE ha de processar la informació en cinc fases i en cada fase, modularment, dur a terme una operació específica:

1. Fase 1: definició ⇒ Mòdul de definició
2. Fase 2: detecció ⇒ Mòdul de detecció
3. Fase 3: restricció ⇒ Mòdul de filtració
4. Fase 4: presentació i interacció ⇒ Mòdul de presentació i interacció
5. Fase 5 exportació ⇒ Mòdul d'exportació.

Cada mòdul està integrat per un conjunt de components i de subcomponents[6]. Al final de l'aplicació del sistema l'usuari disposaria de la selecció d'USE més adequada a les necessitats professionals de l'activitat per a la qual l'ha fet servir.

8.1 Aportacions i aplicacions

La principal contribució aplicada d'aquest treball és l'elaboració del model de SEACUSE. Aquest model millora els SEACAT bàsicament en cinc aspectes: augmenta el nombre de les unitats terminològiques reconegudes, obre el concepte d'objecte d'extracció, precisa més el reconeixement de les USE, adapta el buidatge als perfils d'activitats professionals i integra el reconeixement i extracció d'USE en altres sistemes.

El sistema **augmenta el nombre i els tipus d'unitats terminològiques detectades** perquè reconeix no només les unitats terminològiques polilèxiques, sinó també les monolèxiques: simples, derivades, compostes i siglades. Però, a més, no se centra només en les unitats terminològiques, com ho feien els SEACAT, sinó que **detecta totes les unitats del text que tenen un significat especialitzat**, tant les lingüístiques com les no lingüístiques, tant les lèxiques com les sintàctiques, tant les monolèxiques com les polilèxiques. I en la detecció d'aquestes unitats el sistema **aconsegueix una precisió més alta en el reconeixement de les USE** a través de la combinació d'estratègies de diferent naturalesa (lèxiques, morfològiques, morfosintàctiques, morfosemàntiques, tipogràfiques, estadístiques) adaptades a cada tipus d'USE.

Considerem que aquest sistema també supera els SEACAT existents perquè, partint de la base que cada activitat, prioritza un determinat tipus i nombre d'USE, **adapta les seves seleccions finals als perfils de necessitats que requereix una activitat professional**.

Finalment, considerem un fet positiu que el SEACUSE permeti **connectar-se a altres recursos** i **integrar-se en altres sistemes** de tractament del llenguatge més complexos.

Totes aquestes aportacions contribueixen a dissenyar un extractor o millorar-ne el funcionament d'un d'existent. Així, el treball que presentem, iniciat amb el treball de recerca, ha de continuar amb la implementació d'aquest model de SEACUSE[7] i, posteriorment, amb la comprovació i valoració del sistema, per acabar amb la incorporació de les millores que es desprenguin de l'avaluació de la seva aplicació i de l'estudi més aprofundit de determinades unitats.

Al marge de la proposta de disseny d'un SEACUSE, també cal subratllar com a aportació del treball la descripció lingüística de les unitats de significació especialitzada (i per tant també de les UT) dels textos especialitzats d'un domini específic, sobretot la distinció entre UT i USE, i la delimitació entre USE pertinent i no pertinent segons les necessitats professionals establertes empíricament.

Tota tesi té necessàriament un final, però és cert que tot final obre la porta a temes de recerca inexplorats. Amb aquesta tesi s'obren, en la nostra opinió, sis eixos de recerca que permeten continuar aprofundint en el tema:

---

[6] Per a informació detallada de les característiques de les fases, del tipus de components i subcomponents que integren cada mòdul vegeu l'apartat 5.3 del capítol cinquè i tot el capítol setè.

[7] Jordi Vivaldi està realitzant una tesi doctoral sobre la implementació d'un SEACUSE dirigida per la Dra. M. Teresa Cabré i pel Dr. Horacio Rodríguez partint dels materials d'aquesta tesi.

a) Un primer eix, relacionat amb **el model de SEACUSE**, perquè és obvi que una vegada implementat el model d'extractor que hem proposat caldrà avaluar-lo i, naturalment, millorar-lo.

b) Un segon eix de treball sobre **la descripció de les USE**, perquè caldrà aprofundir, d'una banda, en la descripció de les unitats fraseològiques o combinacions especialitzades recurrents i, de l'altra, en l'anàlisi i detecció d'unitats discursivament amagades.

c) Un tercer eix d'investigació, centrat en **l'extracció de determinats tipus d'USE.** Una vegada aplicat el sistema amb les estratègies d'extracció proposades, caldrà afinar els resultats obtinguts i explotar les possibilitats que ens ofereix la semàntica lèxica combinatòria i la lingüística de context.

d) Un quart eix d'estudi, relacionat amb **les activitats professionals,** de les quals caldrà refinar els perfils d'USE pertinents que hem proposat i ampliar-los.

e) Finalment, un cinquè eix de treball a l'entorn de **l'ampliació dels dominis especialitzats d'aplicació,** ja que s'haurà de comprovar fins a quin punt els resultats obtinguts es poden aplicar a altres camps temàtics.

Lentament arriben hores
esperades, ports i costes
que van ser durant temps
somni obsessiu de punt final.

Però passen dos, tres dies,
quatre mesos, mitja vida:
el descans es fa impossible;
esdevé, tot ídol, fals.

Cal refer les naus cremades
i tornar a l'exili ample
de la mar, cada vegada
carregat amb més dens llast.

Final [Jou, D., 1998: 68]

9. BIBLIOGRAFIA

(1996) ACTAS DEL 4º SIMPOSIO IBEROAMERICANO DE TERMINOLOGÍA (1996, Buenos Aires). Buenos Aires: Secretaría de Ciencia y tecnología de la Nación.

(1998) ACTAS DEL 3R SIMPOSIO IBEROAMERICANO DE TERMINOLOGÍA (1994, San Millán de la Cogolla). Barcelona: CINDOC, IULA, SLC.

(1976) *Código Internacional de Nomenclatura Zoológica*. Madrid: Universidad Complutense de Madrid. Traduït per R. Alvarado.

(1969) *ISO/R 1087 Vocabulaire de la Terminologie*. Ginebra: Organisation Internacionale de Normalisation.

(1996) *Revista española de Cardiología.*, Publicación oficial de la Sociedad Española de Cardiología. Barcelona: Doyma, 49/2.

ACADÈMIA DE CIÈNCIES MÈDIQUES DE CATALUNYA I DE BALEARS I ENCICLOPÈDIA CATALANA (1990) *Diccionari Enciclopèdic de Medicina*. Barcelona: Acadèmia de Ciències Mèdiques de Catalunya i de Balears i Enciclopèdia Catalana.

ABEILLE, A.; BLACHE, P. (1997) "État de l'art: la syntaxe". *TAL*, 38/2, 69-90.

ABREU, J.M. (1998) “Las siglas y los acrónimos en el lenguaje técnico”. *Actas del 3r Simposio Iberoamericano de Terminología* (1994, San Millán de la Cogolla). Barcelona: CINDOC, IULA, SLC, 19-28.

ADAMS, R.; VICTOR, M. (1984) *Principios de neurología*. Barcelona: Editorial Reverté S. A.

ADELSTEIN, A. (1996) "Banalización de términos con formantes de origen grecolatino". *Actas del V Simposio Iberoamericano de Terminología, 12-17*. Ciudad de México: Unión Latina, El Colegio de México, ENEP Acatlán e Instituto de Ingeniería de la UNAM, Organización Mexicana de Traductores, Asociación Mexicana de Lingüística Aplicada.

---. (1998 (en premsa)) "Condiciones de reductibilidad léxica de los sintagmas terminológicos". *Actas del VI Simposio Iberorománico de Terminología, Cuba.*

AGÜERO, O. (1987) "Las abreviaturas en las historias y escritos médicos". *Gaceta Médica de Caracas*, 95/1-3, 13-15.

AGUSTÍ, E. (1971) "La terminología médica en documentación clínica". *Medicina Española,* 66, 235-240.

AHMAD, K.; I AL. (1996) "Engineering terminology. A case for a linguistically-informed terminology database". *TKE'96: Terminology and knowledge Engineering. Berlín: Index Verlag*, 166-178.

AITCHISON, J. (1987) *Words in the mind*. Oxford: Basil Blackwell.

ATELIER/FX (1997) *http://www.ling.uqam.ca/ato/FX/AtelierFx.html,* 10 de juliol.

ALES, J. M. (1988) "Uso correcto de nuestro idioma en microbiología". *Enfermedades Infecciosas y Microbiología Clínica,* 6/1, 6-8.

ALSINA, V.; ESTOPÀ, R. (1996) "Las profesiones y los usuarios de la terminología". *Terminómetro,* 1997, número especial/2, 85-87.

ALVAR, M. (1993) *La formación de palabras en español*. Madrid: ARCO/Libros.

ALVARADO, R. (1983) "Los nombres de los taxones y su españolización: estudio del problema sobre un caso práctico". *BRAE*, 63, 227-239.

ANJEWIERDEN A. (1992) "Shelley, computer-aided knowledge engineering". *Knowledge Acquisition*, 4

ANSCOMBRE, J-CL. (1990) "Pourquoi un moulin à vent n'est pas un ventilateur". *Langue française,* 86, 103-185.

---. (1991a) "La détermination zéro quelques propiétés". *Langages,* 102, 103-124.

---. (1991b) "L'article zéro sous préposition". *Langue Française,* 91, 24-39.

ANTICH, J. (1973) "Actualización de la terminología en citogenética humana". *Medicina Clínica,* 60/2, 173-179.

ARECHAGA, J. I AL. (1980) "Hacia un léxico científico universal en citología, histología y embriología. I Nómina histológica". *Morfología formal y patología, Sección A* I, 1-83.

ARÉCHAGA, J.; GUIRAO, M. (1987) "Hacia un léxico científico universal en

citología, histología y embriología. II Nómina embryológica ". *Anales del Desarrollo,* 31/69-70, 43-103.

ARPPE, A. (1995) "Term extraction from unestricted text". *http://www.lingsoft.com,* 7 d'octubre.

ASSAL, A. (1994) "La métaphorisation terminologique". *Terminologie et Traduction,* 2, 235-242.

ASSAL, A.; DELAVIGNE, V. (1993) "Le découpage des unités terminologiques complexes: limites des critères linguistiques". *Actes de la IVème Journée ERLA-CLAT: Les langues de spécialité: pratiques, outils, théories,* 175-193.

ASSAL, A.; GAUDIN, F.; GUESPIN, L. (1992) "Sémantique et terminologie: sens et contextes". *Terminologie et Traduction,* 2/3, 411-421.

AUGER, P.; ROUSSEAU, L-J. (1987) *Metodologia de la recerca terminològica*. Barcelona: Generalitat de Catalunya. Traduït i adaptat per M. T. Cabré.

BACH, C. I AL. (1997) *El corpus de l'IULA: descripció*. Barcelona: Papers de l'IULA, Sèrie Informes, 17, 1-66. Barcelona: IULA, Universitat Pompeu Fabra.

BADIA I MARGARIT, A. (1995) *Gramàtica de la Llengua Catalana*. Barcelona: Proa.

BADIA, T. (1994) *Aspectes del sintagma nominal en català des de la perspectiva de la traducció automàtica.* Barcelona: Publicacions de l'Abadia de Montserrat.

---. (1996) "Bancos de conocimientos". *Terminómetro,* 1997, número especial/2, 71-75.

BALAGUER, E. (1974) "Las nomenclaturas en documentación clínica". *Medicina Española,* 71, 191-200.

BARCIA, J. (1980) "Expresiones y términos incorrectos en las ciencias neurológicas". *Medicina Española,* 79, 377-382.

---. (1983) "Los orígenes de la terminología anatómica en las lenguas catalana y valenciana". *Medicina Española,* 82, 121-137.

BARKEMA, H. (1996) "Idiomaticity and Terminology: A Multi-dimensional

Descriptive Model". *Studia Linguistica,* 50/2, 125-160.

BARONA, J. L. (1990) *Introducció a la medicina*. València: Servei de Publicacions de la Universitat de València.

---. (1993) "Teorías médicas y la clasificación de las causas de muerte". *Boletín de la Asociación de Demografía Histórica,* 11/3, 49-64.

BASILI, R. I AL. (1997) "Contextual word sense tuning and disambiguation". *Applied Artificial Intelligence,* 11, 235-262.

BAYLON, C.; MIGNOT, X. (1995) *Sémantique du langage.* París: Nathan Université.

BECKER, T. (1992) "Compounding in German". *Rivista di Linguistica,* 4/1, 5-36.

BENAVENT, A.; ISCLA. A. (1997) "Vicios del lenguaje y defectos del estilo científico en las comunicaciones del IV Congreso Nacional de Documentación Médica". *Papeles Médicos,* 6/3, 5-13.

BERNABEU, J. I AL. (1995) *El llenguatge de les ciències de la salut: introducció a la formació de termes mèdics*. València: Generalitat Valenciana.

BERNAD, J. A. (1995) "Análisis y representación del conocimiento: Aportaciones de la pscicología cognitiva". *Scire,* 1/1, 57-80.

BERNARD, J.; DRUON, M. (1979) "La langue française et la médecine". *Revue des Deux-Mondes,* 7, 43-51.

BERTHONNEAU, A-M. (1991) "Pendant et pour, variations sur la durée et donation de la référence". *Langue française,* 91, 102-124.

BLAIS, R. (1993) "Le phraséologie. Une hypothèse de travail". *Terminologies Nouvelles,* 10, 50-54.

BLANK, I. (1995 (en premsa)) "Méthodes pour l'extraction de terminologie bilingue". *Actes de les IVèmes Journées scientifiques du Réseau Lexicologie, Terminologie, traduction.*

BONET, S.; SOLÀ, J. (1986) *Sintaxi generativa catalana*. Barcelona: Enciclopèdia Catalana.

BOOIJ, G. (1992) "Compounding in Dutch". *Rivista di Linguistica,* 4/1, 37-60.

BORDONI, L.; ANZALDI, C. (1996) *Prototipo di thesaurus per l'energia e*

l'ambiente tramite il sistema SBIC. Roma: ENEA.

BORRÀS, L. (1997) *Los zoónimos y su descripción lexicográfica en español. Estado de la cuestión*. Universitat Pompeu Fabra. Institut Universitari de Lingüística Aplicada: Treball de recerca de doctorat.

BOSQUE, I. (1991) *Las categorías gramaticales: relaciones y diferencias*. Madrid: Síntesis.

---. (1993) "Sobre las diferencias entre los adjetivos relacionales y los calificativos". *Revista Argentina de Lingüística,* 3, 10-48.

BOSREDON, B.; TAMBA, I. (1991) "Verre à pied, moule à gaufres: préposition et noms composés de sous-classe". *Langue française,* 91, 40-55.

BOUGHEDAQUI, M. (1996) "Essai de catégorisation sémantique des adjectifs composés". *Les Cahiers de l'APLIUT,* XVI/2, 37-54.

BOULANGER, J-C. (1988) "Le syntagme en informatique: un projet de recherche". *Terminogramme,* 46, 22-23.

BOURIGAULT, D. (1993) "Analyse syntaxique locale pour le reperage de termes complexes dans un texte". *TAL*, 2, 105-117.

---. (1994) *Lexter, un logiciel d'Extraction de TERminologie. Application à l'acquisition des connaissances à partir des textes*. École des Hautes Études en Sciences Sociales: tesi doctoral.

---. (1995 (en premsa)) "Conception et exploitation d'un logiciel d'extraction de termes: problèmes théoriques et méthodologiques". *Actes de les IVèmes Journées scientifiques du Réseau Lexicologie, Terminologie, Traduction*.

BOURIGAULT, D.; CONDAMINES, A. (1995 (en premsa)) "Réflexion sur le concept de Base de Connaissances Terminologiques". *Actes de les 5èmes Journées Nationales du PRC GDR Intelligence Artificielle*.

BOURIGAULT, D. I AL. (1996) "LEXTER, a Natural Language Processing Tool for Terminology Extraction". *Actes del 7th EURALEX International Congress,* EURALEX'96, Göteborg, *771-780*. Göteborg: Göteborg University.

BOUTIN-QUESNEL, M. I AL. (1985) *Vocabulaire systématique de la terminologie*. Quebec: Les Publications du Québec.

BOUVERET, M. (1998) "Approche de la dénomination en langue spécialisée". *Meta,* XLIII/3, 393-409.

BOVÉ, A.; CERVERA, R.; GALOFRÉ, J. (1989) "Prevalencia del latín en el lenguaje científico". *Medicina Clínica,* 93, 705-708.

BOWDEN, P. I AL. (1998) "Automatic Acronym Acquisition in a Knowledge Extraction Program". *Actes de Coling'96. First Workshop on Computational Terminology,* 43-49.

BROWN, R. I AL. (1998) "Translingual Information Retrieval: Learning from Bilingual Corpora". *Artificial Intelligence Journal. Special issue: Best of IJCAI'97.* ?

BÜHLER, H. (1992) "Of terms and texts". *Terminologie et Traduction,* 2-3, 423-428.

CAAMAÑO, A. (1998) "Nomenclatura, símbolos y escritura de las magnitudes fisicoquímicas". *Alambique,* 17, 47-57.

CABRÉ, M. T. (1992) *La terminologia. La teoria, els mètodes, les aplicacions*. Barcelona: Empúries.

---. (1994a) *A l'entorn de la paraula (I). Lexicologia general*. València: Servei de Publicacions de la Universitat de València.

---. (1994b) *A l'entorn de la paraula (II). Lexicologia Catalana*. València: Servei de Publicacions de la Universitat de València.

---. (1994c) "Terminologie et dictionnaires". *Meta,* 39/4, 589-597.

---. (1995a) "On diversity and terminology". *Terminology,* 2, 1-16.

---. (1995b) "Terminologia i diccionaris II". *Estudis de Llengua i Literatura Catalanes XXXI. Miscel·lània Colon,* 277-305.

---. (1996a) "Diversidad en la terminología: de la disciplina a su funcionalidad". *Sendebar,* 7, 89-96.

---. (1996b) "Terminology today". *Terminology, LSP and Translation* Studies in language engineering in honour of Juan Carlos Sager, 15-33.

---. (dir.) (1996c) *E. Wüster: Terminologia*. Barcelona: Servei de Llengua de la Universitat de Barcelona.

---. (1997a) "Standardization and Interference in Terminology". *American*

Translators Association Scholarly Monograph, Series IX, 49-74.

---. (1997b) "Importancia de la terminología en la fijación de la lengua: la planificación terminológica". *Terminologie et Traduction,* 2, 96-117.

---. (1998a) "Elementos para una teoría de la terminología: hacia un paradigmo alternativo". *El Lenguaraz,* 1, 59-78.

---. (1998b (en premsa)) "El discurs especialitzat o la variació funcional determinada per la temática: noves perspectives". *Caplletra: variació lingüística.*

---. (1998c) "Précision sur le discours de spécialité". *Des mots en liberté. Mélanges offerts à Maurice Tournier.* Textes réunis par P. Fiala et P. Lafon.

---. (1998d) "La noció de normalització terminològica per al treball documental". *Anuari SOCADI de Documentació i Informació,* 113-121.

---. (1998e (en premsa)) "Una nueva teoría de la terminología: de la denominación a la comunicación". *Actas del VI Simposio Iberorománico de Terminología, Cuba.*

--. (1999 (en premsa)) "Is there a need for an autonomous theory of terms?". *Terminology*

CABRÉ, M. T.; DE YZAGUIRRE, LL. (1995) "Stratégie pour la détection semiautomatique des néologismes de presse". *Technoletes et dictionnaires,* VIII/2, 89-100.

CABRÉ, M. T.; ESTOPÀ, R. (1997) "Formar en terminología: una nueva experiencia docente". *TradTerm*, 4/1, 175-202.

CABRÉ, M. T.; ESTOPÀ, R.; LORENTE, M. (1996) "Criterios de reconocimiento de la fraseología a partir del análisis de corpus". *Actes del V Simposio Iberoamericano de Terminología,* 67-81. Ciudad de México: Unión Latina, El Colegio de México, ENEP Acatlán e Instituto de Ingeniería de la UNAM, Organización Mexicana de Traductores, Asociación Mexicana de Lingüística Aplicada.

CABRÉ, M. T.; RIGAU, G. (1985) *Lexicologia i semàntica.* Barcelona: Enciclopèdia Catalana.

CABRÉ, M. T.; ROJO, A. (1996) "Specialized knowledge representation: towards a new hypertextual, multimedia proposal". *Actes del TKE'96*

Terminology and Knowledge Engineering, 424-430.

CADIOT, P. (1991) "A la hache ou avec la hache? Représentation mentale, expérience située et donation du référent". *Langue française*, 91, 7-23.

CADIOT, P.; NEMO, F. "Propiétés extrinsèques en sémantique lexicale". *French Languages Studies*, 1997, 7, 127-146.

CALDEIRO, M. A. I AL. (1993) *Medicina Clínica. Manual de estilo.* Barcelona: Doyma. Publicaciones biomédicas.

CALONGE, J. (1995) “El lenguaje científico y técnico”. Seco, M.; Gregorio, S. (coord.) (1995) *La lengua española, hoy*. Madrid: fundación Juan March, 175-185.

CANDEL, D. (1979) "La présentation par domaines des emplois scientifiques et techniques dans quelques dictionnaires de langue". *Langue française*, 43, 100-115.

CARBONNIER, J. (1980) *Préface au Dictionnaire des principaux sigles utilisés dans le mode juridique de A à Z*. Gendrel, M. (1980). París: Le cours de Droit Rontchrestien, 2-8.

CÁRDENAS, E. (1996) *Terminología médica.* México: McGraw-Hill Interamericana.

CASASSAS, E. (1998) "La nomenclatura de sustancias inorgánicas". *Alambique*, 17, 37-46.

CASASSAS, O.; RAMIS, J. (1993) *La cardiologia mot per mot*. Barcelona: Generalitat de Catalunya.

---. (1994) *Els mots de la dermatologia*. Barcelona: Generalitat de Catalunya.

---. (1995) *La pneumonologia al peu de la lletra*. Barcelona: Generalitat de Catalunya.

CHABNER, D. (1996) (5a ed.) *The language of Medicine: a write-in text explaining medical terms*. Philadephia: Saunders company.

CHARLET, J. I AL. (1996) “Ontologie et réutilisabilité: expérience et discussion”. Aussenac-Gilles, N. i al. (1996) *Acquisition et ingénierie des connaissances: tendances actuelles*. Tolosa: Lépaduès Éditions, 69-87.

CHAROLES, M. (1978) "Introduction aux problèmes de la cohérence des

textes". *Langue française,* 38, 7-41.

CHETOUAMI, L. (1997) *Vocabulaire général d'enseignement scientifique.* París: L'Harmathan.

CHEVALIER, J.; COSTAGLIOLA, J. (1987) "La nouvelle nomenclature anatomique. On touche pas à mon anatomie!". *Prospective et Santé,* 44, 63-64.

CHURCH, K.; HANKS, W. (1989) "Word association norms, mutual information and lexicography". *Actes del 27th Annual Meeting of the ACL,* 76-83.

CLAS, A. (1994) "Collocations et langues de spécialité". *Meta,* XXXIX/4, 576-580.

CLIMENT I AL. (1997) "Definition of the links and subsets for nouns of the EuroWordNet project". *Paper de treball,* Versió 6, 1-123.

CODINA, L. (1996) "Publicación digital yrepresentación del conocimiento". Quark: ciencia, medicina, comunicación y cultura, 5, 33-43.

COHEN, B. (1996) "Le vocabulaire des fractales, une évolution en terminologie". *Circuit,* 51, 29-30.

COLLET, T. (1997) "La réduction des unités terminologiques complexes de type syntagmatique". *Meta,* XLII/1, 193-206.

COLSON, J-P. (1992) "Ébauche d'une didactique des expressions idiomatiques en langue étrangère". *Terminologie et traduction,* 2/3, 165-179.

COMISIÓN DE NOMENCLATURA DE LA QUÍMICA ORGÁNICA DE LA IUPAC (adaptació castellana per Férnanadez Álvarez i Fariña Pérez) (1987) *Nomenclatura de la química orgánica.* Madrid: CSIC.

CONDAMINES, A. (1995) "Terminologie et représentation des connaissances". *Intelligence artificielle,* 1-2-3, 29-44.

CONDAMINES, A.; REBEYROLLE, J. (1997) "Point de vue en langue spécialisée". *Meta,* XXXXII, 1, 174-184.

CONENNA, M. (1988) "Sur un lexique-grammaire comparé de proverbes". *Langages,* 90, 99-118.

CONGOST, N. (1994) *Problemas de la traducción técnica. Los textos médicos en inglés.* Alicante: Universidad de Alicante.

CONTRERAS, L. (1982) "La Torre de Babel del léxico sanitario". *Revista Sanitaria de higiene Pública,* 56, 311-340.

COPECK, T. I AL. (1992) "Parsing and Case Analysis in TANKA". *Actes de la 15ème Conférence Internationale de Linguistique Informatique, Coling'92,* Nantes.

CORBIN, D. (1991) "La morphologie lexicale: bilan et perspectives". *Travaux de linguistique,* 23, 33-56.

---. (1992) "Hypothèses sur les frontières de la composition nominale". *Cahiers de grammaire,* 17, 26-55.

---. (1997a) "Locutions, composés, unités polylexématiques: lexicalisation et mode de construction". *Actes du colloque de 1994 "La Locution, entre langue et usage".* Textes reúnis par Michel Martins-Baltar/ENS Éditions Fontenay/Saint-Cloud, 1-29.

---. (1997b) "La répresentation d'une famille de mots dans le Dictionnaire dérivationnel du français et ses corrélats théoriques, méthodologiques et descriptifs". *Recherches Linguistiques,* 26

CORBIN, D.; PLÉNAT, M. (1994) "Nouvelle note sur l'haplologie dans les mots construits". *Cahiers de grammaire,* 19, 139-166.

CORPAS, G. (1997) *Manual de fraseología española.* Madrid: Gredos.

COURTOIS, B. (1990) "Un système de dictionnaires eléctroniques pour les mots simples du français". *Langue française,* 87, 11-22.

CRYSTAL, D. (1992) *An Encyclopedie Dictionary of Language and Languages.* Oxford: Charlesworth Group.

CURRÁS, E. (1995) "Concierto y desconcierto en la organización del conocimiento actual y su intersección con el mundo de la información". *Scire,* 1/1, 4-27.

DAGAN, I.; CHURCH, K. (1994) "Termight: Identifying and Translating Technical Terminology". *Actes de la 4th Conference on Applied Natural Language (ANLP'94).*

DAILLE, B. (1994) *Approche mixte pour l'extraction de terminologie: statistique lexicale et filtres linguistiques.* Université Paris VII: Tesi doctoral.

---. (1995) "Repérage et extraction de terminologie par une approche mixte statistique et linguistique". *TAL,* 36/1-2, 101-118.

DAILLE, B. I AL. (1996) "Empirical observation of term variations and principles of their description". *Terminology,* 3/2, 197-257.

DANLOS, L. (1988) "Les phrases à verbe support être Prép". *Langages,* 90, 23-37.

DARBELNET, J. (1979) "Réflexions sur le discours juridique". *Meta,* 21/1, 26-34.

DAVID, S. (1993) *Les unités nominales polylexicales. Éléments de description et reconnaissance automatique.* Tesi doctoral. París: Université Denis Diderot.

DAVID, S.; PLANTE, P. (1991) "Le progiciel TERMINO: de la nécessité d'une analyse morphosyntaxique pour le dépouillement terminologique des textes". *Les industries de la langue: perspectives des années 1990. Actes du Colloque de Montréal, 1990,* 1, 71-88.

DAVIDSON, L. I AL. (1998) "Semi-automatic Extraction of Knowledge-Rich Contexts from Corpora". *Actes de Coling'96. First Workshop on Computational Terminology*, 50-56.

DE YZAGUIRRE, L. (1996) "Ingeniería lingüística y terminología". *Terminómetro,* 1997, número especial/2, 69-71.

---. (1998) *"DIGIT"*. Ponència presentada a les Jornades de docència del professorat de la Facultat de Traducció i Interpretació (juny de 1997). Barcelona: Universitat Prompeu Fabra.

DEL HOYO, J. (1985) *L'asma*. Barcelona: Proa.

DESMET, I. (1994) "Propositions pour la recherche en phraséologie contrastive". *Banque des mots,* 6, 45-59.

DESMET, I.; BOUTAYEB, S. (1993) "Terme et mot: Propositions pour la terminologie". *Le banque des mots,* 5, 5-32.

DI SCICULLO A.; WILLIAMS, E. (1987) *On the Definition of Word*. Massachusetts: Massachusttes Institute of Technology.

DOMÈNECH, M. (1998) *Unitats de coneixement i textos especialitzats: primera proposta d'anàlisi.* Universitat Pompeu Fabra. Institut Universitari

de Lingüística Aplicada: Treball de recerca de doctorat.

DRASKAU, J.; PICHT, H. (1985) *Terminology: an Introduction*. Guildford: University of Surrey.

DROUIN, P. (1997) "Une méthodologie d'identification automatique des syntagmes terminologiques: l'apport de la description du non-terme". (en premsa), *Meta,* XLII/1, 45-54.

DUBOIS, J. (1990) "Incomparabilité des dictionnaires". *Langue française,* 87, 5-10.

DUBOIS, J.; I AL. (1979) *Diccionario de Lingüística*. Madrid: Alianza Editorial.

---. (1991) *Dictionnaire de Linguistique*. Canada: Larousse.

DUBUC, R. (1985) *Manuel pratique de terminologie*. Montreal: Linguatech.

DUGAS, A. (1990) "La création lexicale et les dictionnaires électroniques". *Langue française,* 87, 23-29.

EL ATENEO (1992 (9a ed.)) *Diccionario de Ciencias Médicas Dorland*. Buenos Aires: El Ateneo.

ELHADAD, M. (1996) "Lexical Choice for Complex Noun Phrases: structure, modifiers, and determiners". *Machine Translation,* 11, 159-184.

ENCICLOPÈDIA CATALANA (1970) *Diccionari jurídic català*. Barcelona: Enciclopèdia Catalana.

---. (1993 (3a ed.)) *Diccionari de la Llengua Catalana*. Barcelona: Enciclopèdia Catalana.

---. (1993) *Hiperdiccionari català-castellà-anglès en CD-ROM*. Barcelona: Enciclopèdia Catalana.

ENGUEHARD, C.; PANTERA, L. (1994) "Automatic Natural Acquisition of a Terminology". *Journal of Quantitative Linguistics,* 2/1, 27-32.

ESPINOSA, J. (1997) "Unidades sintácticas, relaciones sintagmáticas y funciones sintácticas oracionales". *Lingüística Española Actual,* XIX/2, 137-154.

ESTOPÀ, R. (1996a) "Noms que formen part d'un determinant complex". *Papers de l'IULA. Sèrie Monografies,* 1, 1-26. Barcelona: IULA, Universitat Pompeu Fabra.

---. (1996b) *Les unitats terminològiques polilexemàtiques en els lèxics especialitzats: dret i medicina.* Universitat Pompeu Fabra. Institut Universitari de Lingüística Aplicada: Treball de recerca de doctorat.

---. (1999) “El léxico especializado en los diccionarios de lengua general: las marcas temáticas”. *Revista de la Sociedad de Lingüística Española*, 18.

ESTOPÀ, R.; GELPÍ, C. (1996) "Els termes a través dels reculls lèxics". *Articles*, 9, 71-82.

ESTOPÀ, R.; SAURÍ, R. (1996) "La estación de trabajo del terminólogo". *Terminómetro*, 1997, número especial 2, 75-76.

ESTOPÀ, R.; VIVALDI, J. (1998) “Systèmes de détection automatique de (candidats à) termes: vers une proposition intégratrice”. *Actes des 7èmes Journées ERLA-GLAT*, 385-410. Brest: Faculté des lettres et Sciences Sociales Victor Ségalen.

ESTOPÀ, R.; VIVALDI, J.; CABRÉ, M. T. (1998) "Sistemes d'extracció automàtica de (candidats a) termes: estat de la qüestió". *Papers de L'IULA. Sèrie Informes*, 22, 1-66. Barcelona: IULA, Universitat Pompeu Fabra.

EVANS, D.; LEFFERTS, R. (1995) "CLARIT-TREC Experiments". *Information Processing and Management*, 31/3, 385-395.

EVANS, D.; ZHAI, C. (1996) "Noun-phrase Analysis in Unrestricted Text for Information Retrieval". *Actes del 34Th Annual Meeting of ACL*, 17-24.

FABRA, P. (1956 (1932)) *Gramàtica Catalana.* Barcelona: Teide.

FABRE, C. (1996) *Interprétation automatique des séquences binominales en anglais et en français. Aplication à la recherche d'informations.* Université de Rennes I: Tesi doctoral.

FABRE, C. I AL. (1998 (en premsa)) "La polysémie dans la langue générale et les discours spécialisés". *Sémiotiques.*

FARRERAS, P.; ROZMAN, C. (1997 (13a ed.)) *Medicina interna.* Madrid: Harcout Brace.

FAULSTICH, E. (1996) "Spécificités linguistiques de la lexicologie et de la terminologie. Nature épistémologique". *Meta*, XLI/2, 237-246.

FELBER, H. (1984) *Terminology Manual*. París: Unesco-Infoterm.

---. (1987) *Manuel de terminologie*. París: Organisation des Nations Unies pour l'Education.

FELBER, H.; PICHT, H. (1984) *Métodos de terminografía y principios de investigación terminológica*. Madrid: CSIC i Instituto Miguel Cervantes.

FELLBAUM, C. I AL. (1993) "Adjectives in WordNet". *Paper de treball*, 26-38.

FIALA, P.; LAFON, P.; PIGUET, M-F. (ed.) (1997) *La locution: entre lexique, syntaxe et pragmatique*. París: INALF.

FORCADA, V.; DE CARRASCO, A.; SAGER, J-C. (ed.) (1996) *Estudios computacionales del español y el inglés*. Madrid: Instituto Cervantes.

FOUCAULT, M. (1966 (1963)) *El nacimiento de la clínica. Una arqueología de la mirada médica*. Madrid: Siglo XXI.

FRANTZI, K.; ANANIADOU, S. (1995) "Statistical measures for terminological extraction". *Working Papers of Department of Computing of Manchester Metropolitan University*.

FUNDACIÓ BARCELONA; TERMCAT (1992) *Diccionari de lingüística*. Barcelona: Fundació Barcelona i TermCat.

---. (1993) *Diccionari d'anatomia*. Barcelona: Fundació de Barcelona i TermCat.

FUNDACIÓ ENCICLOPÈDIA CATALANA (1990) *Enciclopèdia de Medicina i Salut*. Barcelona: Fundació Enciclopèdia Catalana.

FUNDACIÓ JOAQUIM TORRENS IBERN (1992) *L'ús del Català científic i tècnic*. Barcelona: Publicacions de l'Abadia de Montserrat.

GAATONE, D. (1987) "Les préfixes négatifs avec les adjectives et noms verbaux". *Cahiers de Lexicologie*, 50/1, 79-90.

GABRIELI, E. (1986) "Construction of a Biomedical Nomenclature". *Meta*, 31/1, 22-25.

GALINSKI, C. (1990) "Terminology and Phraseology". *IITF Journal*, 1/1-2, 70-82.

GALLARDO, N.; MAYORAL, R.; KELLY, D. (1992) "Reflexiones sobre la

traducción científico-técnica". *Sendebar,* 3, 185-191.

GAMBIER, Y. (1992) "Phraséologie et terminologie en traduction et interprétation". *Multilingua,* 11/3, 325-329.

---. (1993) "Socioterminologie et phraseologie: pertinence théorique et méthodologique". *Terminologie et Traduction,* 2/3, 397-409.

GAMPER, J. (1997) "CATEx, Computer Assisted Terminology Extraction". *http: //www.eurac.edu/,* 28 d'Agost, 1-8.

GARCÍA, A.; BERTOMEU, J-R. (1998) "Lenguaje, ciencia e historia: una introducción histórica a la terminología química". *Alambique,* 17, 20-36.

GARCÍA PALACIOS, J. (1996) "La terminología en los manuales de Enseñanza Media: hacia la determinación de la terminología básica del español". *Actas del V Simposio Iberoamericano de Terminología,* 150-157. Ciudad de México: Unión Latina, El Colegio de México, ENEP Acatlán e Instituto de Ingeniería de la UNAM, Organización Mexicana de Traductores, Asociación Mexicana de Lingüística Aplicada.

GARCÍA YEBRA, V. (1982) *Teoría y práctica de la traducción*. Madrid: Gredos.

GARRIDO, A. (1984) *Diccionario de abreviaturas médicas inglés-español*. Barcelona: DIPSA.

GAUDIN, F. (1991) "Terminologie et travail scientifique: mouvement des signes, mouvement de connaissances". *Cahiers de Linguistique Social,* 18, 111-131.

---. (1992) "Terminologie et démocratisation du savoir: à propos de dictionnaires scientifiques". *Le Langage et l'Homme,* 27/2-3, 123-129.

GENTILHOMME, Y. (1995) "Contribution à une réflexion sur les locutions mathématiques". *Cahiers de Lexicologie,* 66/1, 5-37.

GILE, D. (1986) "La compréhension des énoncés spécialisés chez le traducteur: quelques réflexions". *Meta,* 31/1, 26-30.

GOFFIN, R. (1992) "Du synthème au phraséolexème en terminologie différentielle". *Terminologie et Traduction,* 2/3, 431-438.

GÓMEZ, F. (1998) "Linking WordNet Verb Classes to Semantic Interpretation". *Paper de treball, Orlando F1 32816,* 1-7.

GOODRICH, P. (1987) *Legal Discourse.* Hong Kong: MacMillan.

GOUADEC, D. (1990) *Terminologie. Constitution de données*. París: AFNOR.

GRAITSON, M. (1975) "Identification et transformation automatique des morphèmes dans le lexique médical français". *Cahiers de Lexicologie,* 26, 85-109.

GREFENSTETTE, G. (1994) *Explorations in Automatic Thesaurus Discovery*. Boston: Kluwer Academic Press.

GROSS, G. (1988) "Degré de figement de noms composés". *Langages*, 90, 57-72.

---. (1990) "Définition des noms composés dans un lexique-grammaire". *Langue française,* 87, 84-90.

---. (1991) "Syntaxe du complément de nom". *Linguisticae investigationes,* XV/2, 255-284.

---. (1996) *Les expressions figées en français*. París: OPHRYS.

GROSS, G.; VIVÈS, R. (1986) "Les constructions nominales et l'élaboration d'un lexique-grammaire". *Langue française,* 69, 5-27.

GROSS, M. (1988) "Les limites de la phrase figée". *Langages*, 90, 7-22.

---. (1990) "Le programme d'extension des lexiques électroniques". *Langue française,* 87, 123-127.

---. (1993) "Les phrases figées en français". *L'information grammaticale,* 59, 36-41.

GUILBERT, L. (1965) *Le vocabulaire de l'Astronautique.* París: Publication de l'Université de Rouen.

---. (1973) "La spécificité du terme scientifique et technique". *Langue française,* 17, 5-17.

---. (1975) *La creativité lexicale.* París: Larousse.

GUILLE, B. (1978) *Histoires des techniques.* París: Gallimand.

GUILLET, A. (1990) "Reconnaissance des formes verbales avec un dictionnaire minimal". *Langue française,* 87, 52-58.

GUÍO, Y. (1992) "Medicina popular y medicina científica: ¿dos discursos nosológicos y una traducción imposible?. Algunas reflexiones sobre el problema de la integración cultural en América latina desde esta problemática". *Asclepio*, 1, 327-346.

GUTIÉRREZ, B. (1998) *La ciencia empieza en la palabra*. Barcelona: Ediciones Península.

GUYTON, A. (1987) *Fisiología humana*. México: Nueva Editorial Interamericana.

HABERT, B. I AL. (1996) "Symbolic word clustering for medium-size corpora". *Actes de Coling'96*, 490-495.

---. (1997) "Recyclage d'analyses syntaxiques automatiques pour le repérage de variantes de termes". *Atelier des projets franco-canadiens.*

HABERT, B.; NAZARENKO, A.; SALEM, A. (1997) *Les linguistiques de corpus*. París: Armand Colin.

HACKEN, P. (1994) *Informatik und sprache*. Hamburg: Olms.

HEID, U. I AL. (1996 (en premsa)) "Term extraction with standard tools for corpus exploration. Experience drom German". *Actes del 4th International Congress on Terminology and knowledge Engineering, TKE'96, Viena*.

HEISTER, J. (1989) *Dictionary of abreviations in medical sciences*. Berlín : Springer Verlag.

HERMANS, A. (1995) "Sociologie des discourse scientifiques. Quelques réflexions". *Meta*, 40/2, 224-228.

HERSH, W.; HICKAM, D. (1991) "Evaluation of SPHIRE". *Procedings of the 15th Annual Symposium on Computer Application in Medical Care*, 808-812.

---. (1992) "A comparasison of two methods for indexing and retrieval from a full text medical database". *Procedings of the 55th Annual Meeting of the American Society for Information Science*, 221-230.

HOFFMAN, L. (1979) "Towards a theory of LSP. Elements of a methodology of LSP analysis". *Fachprach*, 1/1-2, 12-17.

HOOF, H. VAN (1986) "Les éponymes médicaux: essai de classification". *Meta*,

31/1, 59-84.

HUMPHREY, B.; LINDBERG, D. (1989) "Building the Unified Medical Language System". *Procedings of the 6th Annual SCAMC*, 475-480.

HUOT, H. (1997) "A propos des nominalisations en -ion: mots-thèmes et lacunes dans les séries dérivationnelles du français". *Travaux de linguistique*, 34, 5-19.

IACOBINI, C. (1992) *La prefissazione nell'italiano contemporaneo.* Univerità degli Studi di Roma La Sapienza: Tesi de doctorat.

INSTITUT D'ESTUDIS CATALANS (1995) *Diccionari de la Llengua Catalana*. Barcelona, Palma de Mallorca, València: Edicions 3 i 4, Edicions 62, Editorial Moll, Enciclopèdia Catalana, Publicacions de l'Abadia de Montserrat.

INTERNATIONAL UNION OF BIOCHEMISTRY (1979) *Enzyme nomenclature: recommendations of the nomenclature committee of the International Union of Biochemistry on the nomenclature and classification of enzymes*. New York: Academic Press.

IRMAY, S. (1998) "La terminologie scientifique en hébreu moderne". *Meta*, XLIII/1, 27-30.

IUPAC (1979) *Nomenclature of organic chemistry sections A, B, C, D, E, F and H*. Oxford: Pergamon Press.

JABLONSKI, S. (1991) *Jablonski's dictionary of syndromes and eponymic diseases*. Malabar: Krieger Publishing Company.

JACQUEMIN, C. (1991) "Une grammaire d'unification des noms composés contrôlée par l'acceptabilité". *Cahiers de grammaire*, 16, 51-71.

---. (1994) "Recycling Terms into a Partial Parser". *Actes de la 4th Conference on Applied Natural Language (ANLP'94)*, 113-118.

---. (1996) "What is the tree we see through the window: A linguistic approach to windowing and term variation". *Information Processing and Management*, 32/4, 445-458.

---. (1997) *Variation terminologique: Reconnaissance et acquisition automatique de termes et leurs variants en corpus*. Institute de Recherche en informatique de Nantes. Université de Nantes: Habilitation à diriger des recherches.

JACQUIN, C.; LISCOUET, M. (1996) “Terminology extraction from texts corpora: application to document keeping via Internet”. *TKE'96: Terminology and Knowledge Engineering*. Berlín: Index Verlag, 74-83.

JAMMAL, A. (1988) "Les vocabulaires des spécialités médicales: pourquoi et comment les fabrique-t-on?". *Meta,* 33/4, 535-541.

JAMMAL, A. I AL. (1986) "L'epidémiologie et les mots pour le dire". *Meta,* XXXI/1, 34-57.

JÀUREGUI, S. (1973) "La terminología científica en la traducción". *Lenguaje y ciencia,* 13/2, 82-87.

JEFFREY, C. (1976) *Nomenclatura biológica*. Madrid: Blume.

JIMS (1993) *Diccionario oncológico. Abreviaturas, siglas y acrónimos*. Barcelona: Jims.

JING, H. AND TZOUKERMANN, E. (1998) "Improving Retrieval with Semantics and Morphology". *Actes de Coling'96. First Workshop on Computational Terminology*, 76-80.

JORGE, G. (1992) "Les expressions idiomatiques correspondantes: analyse comparative". *Terminologie et Traduction,* 2/3, 127-134.

JUNG, R. (1990) "Remarques sur la constitution du lexique des noms composés". *Langue française,* 87, 91-97.

JUSTESON, J.; KATZ, S. (1995) "Technical terminology: some linguistic properties and an algorithm for identification in text". *Natural Language Engineering,* 1/1, 9-27.

KAGEURA, K.; UMINO, B. (1996) "Methods of Automatic Term Recognition: a review". *Terminology,* 3, 2, 259-289.

KANTER, S. I AL. (1994) "Using POSTDOC to recognize biomedical concepts in medical school curricular documents". *Bulletin Medical Library Association,* 82/3, 283-287.

KARLSSON, F. (1990) "Constraint grammar as a framework for parsing running text". *Actes de la 13th International Conference on computational Linguistic,* 3, 168-173.

KIEFER, F. (1992) "Compounding in Hungarian". *Rivista di Linguistica,* 4/1, 61-78.

KISTER, L. (1993) *Groupes nominaux complexes et anaphores: possibilites de reprise pronominale dans un "N1 de (det) N2"*. Université de Nancy: Tesi de doctorat.

KJAER, A. L. (1990) "Methods of describing word combinations in language for specific purposes". *IITF Journal*, 1/1-2, 3-32.

KLEIBER, G. (1990) *La sémantique du prototype*. París: PUF.

---.(1994) "Métaphore: le problème de la déviance". *Langue française,* 101, 35-65.

---. (1996) "Noms propres et noms communs: un problème de dénomination". *Meta,* XLI/4, 567-589.

KLEIN, J. R.; LAMIROY, B. (1994) "Lexique-grammaire du français de belgique: les expressions figées". *Linguisticae Investigationes,* XVIII/2, 285-320.

KOZLOWSKA, C. D. (1991) "English Adverbial Collocations". *International Journal of Lexicography*, 1993, 6/4, 300-304.

KRIEGER, M-G. (1996) "Environmental Law Dicionary: from theory to practice". *Meta*, XLI/2, 259-264.

L'HOMME, M-C. (1996a) "Formes verbales de temps et texte scientifique". *Le Langage et l'Homme*, XXXI/2-3, 107-123.

---. (1996b) "Sélection des prépositions dans les termes complexes Nom (Prép.) Nom à partir de leur structure conceptuelle". *Cahiers de Lexicologie,* 68/1, 25-43.

---. (1996c) “A computerized Model fot Processing Lexical Combinations in Technical Language”. *Actes del 7th EURALEX International Congress,* EURALEX'96, Göteborg, *797-806*. Göteborg: Göteborg University.

---. (1998) "Le Statut du verbe en langue de spécialité et sa description lexicographique”. *Cahiers de Lexicologie,* 73/2, 61-84.

L'HOMME, M-C. i al. (1996) "Definition of an evaluationgrid for ter-extraction software". *Terminologyu*, 3, 2, 291-312.

LABELLE, J. (1988) "Lexiques-grammaires comparés: formes verbales figées en français du Québec". *Langages*, 90, 73-97.

LAINÉ, C.; PAVEL, S.; BOILEAU, M. (1992) "La phraséologie: nouvelle

dimension de la recherche terminologique. Travaux du module canadien du Rint". *L'Actualité terminologique,* 25/3, 5-9.

LA MAISON DU DICTIONNAIRE (1992) *Dictionnaire de sigles*. París: La Maison du Dictionnaire.

LAPAGE, M. (1975) *International code of nomenclature of bacteria and Statutes of International Comittee on systematic Bacterology.* Wagshington: International Association of Microbiological Societies.

LAPORTE, E. (1988) "Le reconnaissance des expressions figées lors de l'analyse automatique". *Langages,* 90, 119-128.

---. (1990) "Le dictionnaire phonémique DELAP". *Langue française,* 87, 59-70.

---. (1997) "Les mots. Un demi-siecle de traitements". *TAL,* 38/2, 47-68.

LARIVIÈRE, L. (1989) "Vers un produit unifié en terminologie et en documentation: le thésaurus terminologique". *Meta,* XXXIV/3, 457-467.

LECLÈRE, C. (1990) "Organisation du lexique-grammaire des verbes français". *Langue française,* 87, 112-122.

LEECH, G. (1974) *Semantics*. Middlesex: Penguin Books.

LEEMAN, D. (1990) "Verbes en tables et adjectifs en -able". *Langue française,* 87, 30-51.

---. (1991) "Hurler de rage, rayonner de bonheur: remarques sur une constructuion en de". *Langue française,* 91, 56-79.

LERAT, P. (1983) *Sémantique descriptive*. París: Hachette Université.

---. (1990) "L'Hyperonymie dans la structuration des terminologies". *Langages,* 98, 79-86.

---. (1994) "Composé syntagmatique, dénomination, terminologie". *Cahiers de Lexicologie,* 65/2, 151-158.

---. (1995) *Les langues spécialisées.* París: PUF.

LIEBER, R. (1992) "Compounding in English". *Rivista di Linguistica,* 4/2, 79-96.

LIN, D. (1998) "Extracting Collocations from Text Corpora". *Actes de Coling'96. First Workshop on Computational Terminology,* 57-63.

LIPPERT, H.; LEHMANN, H. (1980) *Sistema Internacional de unidades en medicina: introducción al Sistema Internacional de unidades.* Barcelona: JIMS.

LOFFLER-LAURIAN, A-M. (1984) "Vulgarisation scientifique: formulation, reformulation, traduction". *Langue française,* 64, 109-125.

---. (1994) "Réflexions sur la métaphore dans les discours scientifiques de vulgarisation". *Langue française,* 101, 72-79.

LÓPEZ PIÑERO, J. M.; TERRADA FERRANDIS, M. L. (1990) *Introducción a la terminología médica*. Barcelona: Salvat editores.

LORENTE, M. (1994) *Aspectes de lexicografia: representació i interpretació gramaticals*. Universitat de Barcelona: Tesi doctoral.

LORENTE, M.; BEVILACQUA, C.; ESTOPÀ, R. (1998 (en premsa)) “El análisis de la fraseología especializada mediante elementos de la lingüística actual”. *Actas del VI Simposio Iberorománico de Terminología, Cuba.*

LORENZO, E. (1986) "Tecnicismos y traducción". *Telos,* 5, 90-95.

LOVE, G.; DAVIS, P. (1990) *Curso rápido de terminología médica*. México: Limusa.

MACEDO, M-E. (1992) "Noms composés: traitement automatique, traduction". *Terminologie et Traduction,* 2/3, 119-126.

MAINGUENEAU, D.; SALVADOR, V. (1995) *Elements de lingüística per al discurs literari*. València: Tàndem Edicions.

MAKAGAWA, H.; TATSUNORI, M. (1998) "Nested Collocation and Compound Noun for Term Extraction". *Actes del Computerm'98. First Workshop on Computational Terminology*, Montréal, 64-70.

MANTECA, A. (1985) "Sintaxis del compuesto". *Lingüística Española,* 1987, IX/2, 333-346.

MANUILA, A. (1975) *Dictionnaire Français de Médicine et de Biologie*. París: Masson.

MARCHAND, H. (1960) *The categories and types of present-day English word-formation*. Wiesbaden: Otto Harrassowitz.

MARCOVECCHIO, E. (1993) *Dizionario etimologico storico dei termini medici*. Florència: Festina Lente.

MARQUET, L. (1995) *El llenguatge científic i tècnic.* Barcelona: Col·legi d'Enginyers Industrials de Catalunya.

MARTÍN, M. A. (1997) "Formación de palabras y lenguaje técnico". *Revista Española de Lingüística,* 27/2, 317-340.

MARTÍN MUNICIO, A. (1992) "La metáfora en el lenguaje científico". *BRAE,* 72, 221-249.

MARTIN-VALIQUETTE, L. (1986) "Les traquenards de la traduction médicale ou l'interaction texte-traducteur-dictionnaire". *Meta,* XXXI/1, 31-33.

MARTÍN VIDE, C. (ed.) (1996) *Elementos de lingüística*. Barcelona: Octaedro Universidad.

MARTIN, W. (1992) "Remarks on Collocations in Sublanguages". *Terminologie et Traduction,* 2/3, 157-164.

MARTINS-BALTAR, M. (ed.) (1997) *La locution entre langue et usage.* París: ENS Éditions Fontenay/Saint-Cloud.

MATHIEU-COLAS, M. (1990) "Orthographe et informatique: établissement d'un dictionnaire électronique des variantes graphiques". *Langue française,* 87, 104-111.

---. (1996) "Essai de typologie des noms composés français". *Cahiers de lexicologie,* 69/2, 71-125.

MATTHEWS, R. (1982) *Clasificación and Nomenclature of Viruses. Fourth Report of the International Commitee on Taxonomy of Viruses*. Basilea: Karger.

MAYNARD, D.; ANANIADOU, S. (1998) "Acquiring Contextual Information for Term Disambiguation". *Actes de Coling'96. First Workshop on Computational Terminology*, 86-90.

MEER VAN DER, A. (1998) "Collocations as one particular type of conventional word combinations, their definition and character". *EURALEX'98,* 1, 313-322.

MELBY, A. (1990) "Benefits and limitations of formal systems in technical writing". *TKE'90,* 1, 23-30.

MERLO, J-C. (1993) "Terminología y lingüística informática". *Voces*, 1994, 3, 2-5.

MESTRE, J. M, I AL. (1995) *Manual d'estil. La redacció i l'edició de textos*. Barcelona: Eumo ed. i al.

MEUNIER-CRESPO, M. (1995 (en premsa)) "Les locutions nominales dans les dictionnaires de spécialités ". *Actes del IVèmes Journées du Réseau Lexicologie, Terminologie, Traduction*.

MILLNER, G. (1990) "Nouns in WordNet: A Lexical Inheritance System". *International Journal of Lexicography*, 3/4, 245-264.

MILLNER I AL. (1993) "Introduction to WordNet: An On-line Lexical Database". *Paper de treball*, 1-9.

MOESCHLER, J. (1992) "Idiomes et locutions verbales". *Terminologie et Traduction*, 2/3, 135-147.

MONTERO, B. (1998) "Compounds vs. Complex Nominals". *Terminology Science and research*, 9, 1, 33-42.

MOREAU, A. (1986) "La traduction médicale: réflexions de praticiens. Enquête d'André Moreau". *Meta*, XXXI/1, 98-105.

MOREL, J. I AL. (1997) *El Corpus de l'IULA: etiquetaris*. Barcelona: Papers de l'IULA, Sèrie Informes, 18, 1-76. Barcelona: IULA, Universitat Pompeu Fabra.

MORENO, J. C. (1987) *Fundamentos de sintaxi general*. Madrid: Síntesis.

---. (1994) *Curso universitario de lingüística general*. Madrid: Síntesis.

MOTTA, E. I AL. (1991) "Methodological fondations of KEATS, the Knowledge Engineer's Assistant". *Knowledge Acquisition*, 3.

MUNNIER, M. (1994) "La composition nominale, une microsyntaxe. Les locutions nominales en espagnol". Fiala, P. i al. (1994) *La locution: entre lexique, syntaxe et pragmatique*. París: Publication de l'INALF, 69-76.

NAULLEAU, E. (1998) *Apprentissage et filtrage syntatico-sémantique des syntagmes nominaux pour la recherche documentaire*. Université Paris VIII: Tesi doctoral.

NAVARRO, F. (1995a) "La nomenclatura de los fármacos (I). ¿Qué es y para

qué sirve la denominación común internacional?". *Medicina Clínica,* 105, 344-348.

---. (1995b) "La nomenclatura de los fármacos (II). Las denominaciones comunes internacionales en España". *Medicina Clínica,* 105, 382-388.

---. (1995c) "La nomenclatura de los fármacos (III). Propuesta de normalización ortográfica de las denominaciones comunes internacionales y adaptación del inglés al castellano". *Medicina Clínica,* 105, 420-427.

---. (1997) *Traducción y lenguaje en medicina*. Barcelona: Doyma. Fundación Dr. Antonio Esteve.

NAVARRO, X. (1996) *Curs pràctic de terminologia mèdica.* Bellaterra: Servei de Publicacions de la Universitat Autònoma de Barcelona.

NOALLY, M. (1989) "Le nom composé: us et abus d'un concept grammatical". *Cahiers de grammaire,* 14, 109-126.

NORMAS UNE 50 001 (1995) *Clasificación Decimal Universal (CDU).* Madrid: AENOR.

NYCKEES, V. (1998) *La sémantique.* París: Belin.

OAKES, M.; PAICE, C. (1998) "Term Extraction for Automatic Abstracting". *Actes de Coling'96. First Workshop on Computational Terminology*, 91-95.

OLLER, L.; FERRER, L. (1979) *L'epilèpsia*. Barcelona: Acadèmia de Ciències Mèdiques de Catalunya i de Balears.

ONIGA, R. (1992) "Compounding in Latin". *Rivista di Linguistica,* 4/2, 97-116.

ORDÓÑEZ, A. (1992a) "Lenguaje médico 1992". *Medicina Clínica,* 99, 781-783.

---. (1992b) *Lenguaje médico. Estudio sincrónico de una jerga.* Madrid: Ediciones de la Universidad Autónoma de Madrid.

---. (1994) *Lenguaje médico. Modismos, tópicos y curiosidades*. Madrid: Noesis.

ORDÓÑEZ, A.; GARCÍA C. (1987) "Diversos aspectos del lenguaje médico (los modismos al uso)". *Medicina Clínica,* 89, 419-421.

---. (1989) "Las metáforas médicas". *Medicina Clínica,* 93, 374-375.

ORGANIZACIÓ MUNDIAL DE LA SALUT (1975) *Las unidades SI para las profesiones de la salud*. Ginebra: OMS.

OTMAN, G. (1991) "Des ambitions et des performances d'un système de dépouillement terminologique assisté par ordinateur". *La banque des mots,* 4, 59-96.

---. (1996) *Les représentation sémantiques en terminologie*. París: Masson.

---. (1997) "Les bases de connaissances terminologiques: les banques de terminologie de seconde génération". *Meta,* XLII/2, 244-256.

PACAK, M.; DUNHAM, G. (1973) *Automated Morphosemantic Analysis of Compound Words Forms in Medical language.* National Institutes of Health: Internal Report, Division of Computer Research and Tecnology.

PARRA, J. (1991) *Manual de términos médicos con nombre propio*. Madrid: Luzán.

PAVEL, S. (1993) *Bibliographie de la phraséologie (1905-1992)*. Montreal: Bureau de la Traduction.

PEARSON, J. (1998) *Terms in Context*. Amsterdam: John Benjamins.

PÉREZ PEÑA, F. (1994) "Deterioro del lenguaje médico. El imperio de las siglas". *Organización Médica Colegial,* 38, 4-5.

PERRON, J. (1988) "L'analyseur syntaxique: un outil tout à fait adéquat?". *Terminogramme,* 46, 29-31.

---. (1989) "Termino: un système de dépouillement terminologique". *Terminogramme,* 54, 3-9.

---. (1991) "Présentation du progiciel de dépouillement terminologique assisté par ordinateur: Termino". *Industries de la langue: perspectives des années 1990. Actes du Colloque de Montréal 1990,* 1991, 2, 715-755.

PESANT, G.; THIBAULT, E. (1993) "Terminologie et coocurrence dans la langue du droit". *Terminologies Nouvelles,* 10, 23-35.

PETERSON, W. (1987) *Formulación y nomenclatura química inorgánica. Según la normativa IUPAC*. Barcelona: EDUNSA.

PETRECA, F. (1992) "Taxonomía científica y discurso lexicográfico". *BRAE*, 72, 251-267.

PICHT, H. (1987) "Terms and their LSP Environment, LSP Phraseology". *Meta,* XXXII/2, 149-155.

---. (1990) "LSP phraseology from the terminological point of view". *IITF Journal,* 1/2, 33-48.

---. (1991) "Fraseología LSP (1) desde el punto de vista terminológico". *Sendebar,* 2, 91-113.

PICOCHE, J.; HONESTE, M-L. (1997) "Les figures éteintes dans le lexique de haute fréquence". *Langue française,* 101, 112-124.

PICONE, M. D. (1991) "L'impulsion synthétique: le français poussé vers la synthèse par la technologie". *Le français moderne*, 1991, LIX/2, 148-163.

PIERREL, J-M. (1989) "Lexique et compréhension automatique de la parole". *Lexiques,* 8, 137-165.

PILZ, K. (1978) *Phraseologie, Versucheiner interdissziplinären Abgrenzung*. Göppingen: Bände.

PINEIRA-TRESMONTANT, C. (1992) "Reconnaissance automatique des unités syntagmatiques". *Paper de treball*, 1-13.

PIOT, M. (1988) "Conjonctions de subordination et figement". *Langages,* 90, 38-56.

PIVAUT, L. (1989) "Les dictionnaires électroniques: un projet de représentation des noms composés ". *Linguisticae investigationes,* XIII/1, 117-145.

PLANAS, A. (1994) "AUTOLEX: Sistema para la gestión de bases de datos terminológicas y herramienta para la traducción asistida por computadora". *Ciencias de la información,* 25, 6-61.

PLANAS, J. (1985) *Elementos de biología*. Barcelona: Omega.

PLANTE, P.; DUMAS, L. (1988) "Le dépoulliment terminologique assisté par ordinateur". *Terminogramme,* 46, 24-28.

POTTIER, B. (1993) *Semántica general*. Madrid: Gredos.

POTVIN, D. (1982) "Le découpage du terme". *Terminogramme,* 14, 1-3.

PRATT, A. (1973) “Medicine, computers and linguistic”. *Advances in Biomedical Engineering,* 3, 97-140.

PRATT, A.; PACAK, M. (1969) "Identification and transformation of terminal morphemes in medical english". *Methods of Information in Medicine,* 8/2, 84-90.

PUERTA, J-L.; MAURI, A. (1995) *Manual para la redacción, traducción y publicación de textos médicos*. Barcelona: Masson.

PUSTEJOVSKY, J. (1995) *The Generative Lexicon.* Massachusetts: Massachusetts Institute of Techonology.

PUSTEJOVSKY, J. I AL. (1993) "Lexical Semantic Techniques for Corpus Analysis". *Computational Linguistics,* 19/2, 331-358.

QUEMADA, B. (1971) "A propos de la néologie: essai de délimitation des objectifs et des moyens d'action". *La banque des mots,* 2, 137-150.

---. (1978) "Technique et langage". *Guille, B. (1978) Histoires des techniques. París: Gallimard*, 52-64.

QUINTANA, J. M. (1989) *La terminología médica a partir de sus raíces griegas*. Madrid: Dykinson.

RAINER, F.; VARELA, S. (1992) "Compounding in Spanish". *Rivista di Linguistica,* 4/2, 117-142.

RALLI, A. (1992) "Compounding in Modern Greek". *Rivista di Linguistica,* 4/2, 143-174.

RASTIER, F. (1987) *Sémantique interprétative.* París: PUF.

---. (1995) “Le défigement des expressions figées et leur interprétation” Fall, K. (ed.) (1995) *Polysémie et construction du sens.* Montpeller: Université Paul-Valéry, 17-24

REALITER (1997) *Taula de formants cultes.* Barcelona: IULA, Universitat Pompeu Fabra.

REY, A. (1988) "Terminology et lexicographie". *Parallèles,* 10, 27-35.

---. (1992, (1979 1a ed.)) *La terminologie: noms et notions*. París: Presses Universitaires de France.

RIEGEL, M. (1988) "Les séquences composées N1 - N2: une catégorie floue". *Studia romanica posnaniensia,* 13, 129-138.

---. (1993) "Grammaire et référence: à propos du statut sémantique de l'adjectif qualificatif". *L'information grammaticale,* 58, 5-10.

RIERA, C. (1993) "Terminologia mèdica". *Annals de Medicina,* 10, 223-224.

---. (1994) *El llenguatge científic català: antecedents i actualitat.* Barcelona: Barcanova.

---. (1998) *Curs de lèxic científic*. Barcelona: Claret.

ROBERTS, R. (1993) "La phraséologie: état des connaissances". *Terminologies Nouvelles,* 10, 36-42.

RODRÍGUEZ ADRADOS, F. (1997) "Los orígenes del vocabulario científico". *Revista Española de Lingüística,* 27/2, 299-316.

ROJO, A. (1995) "La invención de máquinas simuladoras de los procesos heurísticos de pensamiento". *Anthropos,* 164, 33-40.

ROMANA, L. (1997) *Da linguagem jurídica à linguagem documental: metodologia e construçao de um microtesauro de direito administrativo.* Universiade Nova de Lisboa: Tesi doctoral.

RONDEAU, G. (1984) *Introduction à la terminologie.* Québec: Gaëtan Morin Éditeur.

---. (ed.) (1979) *Table ronde sur les problèmes du découpage du terme. 5ème congrès International de Linguistique Appliquée, Montréal, 1978*. Montreal: OLF.

RONDEAU, G.; FELBER, H. (ed.) (1981) *Textes choisis de terminologie I. Fondaments théoriques de la terminologie.* Quebec: Université Laval.

ROULEAU, M. (1995) "La langue médicale: une langue de spécialité à emprunter le temps d'une traduction". *TTR: Technolectes et Dictionnaires,* VIII/2, 29-49.

ROUSSEAU, L-J. (1993) "Terminologie et phraséologie, deux composantes indissociables des langes de spécialités". *Terminologies Nouvelles,* 10, 9-11.

RUIZ GURILLO L. (1997) *Aspectos de fraseología teórica española*. València: Universitat de València, Cuadernos de Filología, XXIV.

RUIZ LÓPEZ, R. I AL. (1991) "El lenguaje del dolor". *Medicina Clínica,* 96, 106.

RUIZ, R. I AL. (1988) *Nuevo diccionario médico.* Barcelona: Teide.

SABAH, G. (1997) "Le sens dans les traitements automatiques des langues". *TAL,* 38/2, 91-133.

SAGER, J-C. (1980) "Standardization of nomenclature". *International of the Sociology of Language,* 23, 25-32.

---. (1990) *A pratical course in Terminology Processing.* Amsterdam: John Benjamins.

---. (1993) *Curso práctico sobre el procesamiento de la terminología.* Madrid: Fundación Germán Sánchez Ruipérez.

---. (1998) "Terminología y traducción". *Conferència llegida el 03/05/1998 a la Universitat Pompeu Fabra*, 1-16.

SAGER, J-C.; KAGEURA, K. (1994) "Concept Classes and Conceptual Structures: their role and necessity in terminology". *Terminology and LSP Linguistic, Studies in Specialised Vocabulaires and Texts. Actes de Langue Française de Linguistique (ALFA),* 1994, 191-216.

SALTON, G.; BUCKLEY, C. (1988) "Term-Weighting Approaches in Automatic Text Retrieval". *Information Processing and Management,* 24/X, 513-523.

SANS, M. (1990) "La importancia del inglés como vehículo de comunicación e información científica y su enseñanza en las ciencias de la salud". *Revista Clínica Española,* 187, 25-28.

SANTACANA, A. (1990) *Lèxic bàsic de la visita mèdica.* Barcelona: Associació professional d'informadors tècnics sanitaris.

SANTAMARIA, C. (1998) *Les definicions dels termes d'especialitat en els diccionaris generals: diversitat i sistematització.* Universitat Pompeu Fabra. Institut Universitari de Lingüística Aplicada: Treball de recerca de doctorat.

SCALISSE, S. (1992) "Compounding in Italian". *Rivista di Linguistica,* 4/2, 175-200.

SCHMIDT, G.; WETTER, T. (1990) "Towards knowledge Acquisition in Natural Language Dialogue". *Procedings of the 3rd European Workshop on*

Konwledge Acquisition for Knowledge-Based Systems (EKAW), Amsterdam.

SILBERZTEIN, M. (1990) "Le dictionnaire électronique des mots composés". *Langue française,* 87, 71-83.

---. (1993) "Les groupes nominaux productifs et les noms composés lexicalisés". *Linguisticae investigationes,* XVII/2, 405-425.

SIBLOT, P. (1995) “La polysémie en question: une question mal posée?”. Fall, K. (ed.) (1995) *Polysémie et construction du sens*. Montpeller: Université Paul-Valéry, 41-62.

SINCLAIR, J. (1991) *Corpus, concordance, collocation*. Oxford: Oxford University Press.

SISTEMA INTERNATIONAL DE UNIDADES (1974) Madrid: Comisión Nacional de Metrología y Metrotecnia.

SLODZIAN, M. (1995) "Comment revisiter la doctrine terminologique aujourd'hui?". *Le banque des mots,* 7, 11-18.

SMITH, G. L.; DAVIS, P. E. (1989) *Terminología médica. Texto programado*. México: Limusa-Wiley.

SOLER, C. (1997) *Desajustes léxicos nominales y su representación en una base de conocimiento léxico. Valores sémanticos de los adjetivos*. Universitat Politècnica de Catalunya: tesi doctoral.

SONNEVELD, H.; LOENING, L. (1993) *Terminology. Aplications in interdisciplinary communication*. Amsterdam: John Benjamins.

SORGI, M.; HAWKINS, C. (1990) *Investigación médica*. Barcelona: Medici.

SPENCER, A. (1995) *Morphological Theory*. Massachusetts: Blackwell Publishers.

SPYNS, P. (1996) "Natural Language Processing in Medicine: An Overview". *Methods of Information in Medicine,* 35, 285-301.

STAMBUK, A. (1998) "Metaphor in scientific communication". *Meta*, XLIII/3, 373-379.

STANASZEK, W. I AL. (1996) *Análisis y comprensión de la terminología médica*. Barcelona: Rasgo Editorial S.L.

TABLA PERIÓDICA (1976) Barcelona: Reverté.

TAGNIN, E. (1992) "What's in a verbal Colligation?". *Terminologie et Traduction,* 2/3, 149-156.

TALEB, S-A. (1993) "Rapport de la phraséologie avec la terminologie". *Terminologies nouvelles,* 10, 13-15.

TAMBA, I. (1991) "Organisation hiérarchique et relations de dépendance dans le lexique". *L'information grammaticale,* 50, 43-47

---..(1994a (1988)) *La sémantique.* París: PUF.

---. (1994b) "Une clé pour différencier deux types d'interprétation figurée, métaphorique et métonymique". *Langue française,* 101, 26-34.

TATILON, C. (1980) "Traitement des unités lexicales". *Meta,* 1982, XXVII/2, 167-172.

THOIRON; P. I AL. (1996) “Notion d’archi-concept et dénomination”. Meta, 1996, XXXXI/4, 512-524.

TEBÉ, C. (1996) *Els conceptes en la teoria terminològica: anàlisi i revisió crítica.* Universitat Pompeu Fabra. Institut Universitari de Lingüística Aplicada: Treball de recerca de doctorat.

VARANTOLA, K. (1986) "Special Language and General Language: Linguistic and didactic aspects". *ALSED-LSP Newsletter,* 9/2, 23-36.

VARELA, S. (1990a) "Composición nominal y estructura temática". *Revista de Lingüística,* XX/1, 136-161.

---. (1990b) *Fundamentos de Morfología*. Madrid: Síntesis.

VIANA, A.; DE LA MORENA, J. (1998) "Abreviaturas o siglas en los informes de alta en Medicina Interna". *Anales de Medicina Interna,* 15/4, 194-196.

VILANUEVA, A. (1986) "Siglas: abreviatura o confusión?". *Revista Española de Enfermedades del Aparato Digestivo,* 70/2, 160.

VILLALVA, A. (1992) "Compounding in Portuguese". *Rivista di Linguistica,* 4/2, 201-220.

VILLAR, J. (1988) "El inglés, idioma internacional en Medicina". *Medicina Clínica,* 91, 23-24.

VILLARRASA, J. (1984) *Introducció a la nomenclatura química (inorgànica i orgànica)*. Barcelona: EUNIBAR.

VIVALDI, J. (1996a) "Corpus especializados y terminología". *Terminómetro,* 1997, número especial/2, 68-69.

---. (1996b) “Proyectos del IULA: Corpus técnico”. Forcada, V.; De Carrasco, A.; Sager, J-C. (ed.) (1996) *Estudios computacionales del español y el inglés.* Madrid: Instituto Cervantes, 227-241.

VIVALDI I AL. (1996) *Marcatge estructural i morfosintàctic del corpus tècnic amb l'estandard SGML*. Barcelona: Papers de l'IULA, Sèrie Informes, 1, 36. Barcelona: IULA, Universitat Pompeu Fabra.

VIVÈS, R. (1990) "Les composés nominaux par juxtaposition". *Langue française,* 87, 98-104.

VOSSEN, P. I AL. (1997) "The EuroWordNet Base Concepts and Top Ontology". *Document de treball LE-4003-D-017, D-034, D-036,* 1-45.

VOUTILAINEN, A. (1993) "NPtool, a detector of english noun phrases". *Proceedings of Workshop on Very Large Corpora. Columbus, Ohio State University.*

WACHOLDER, N.; BYRD, R. (1994) "Retrieving Information from Full Text Using Linguistic Knowledge". *Procedings of the 15th National On Line Meeting*, 441-447.

WAKABAYASHI, J. (1996) "Teaching medical translation". *Meta,* 41/3, 356-365.

WALCZAK, B. (1981) "La terminologie dans les dictionnaires généraux (en s'appuyant sur l'exemple du Dictionnaire de la langue polonaise sous la réd. de Mieczyslaw Szymczak)". *Neoterm*, 13/16, 126-130.

WEISSENHOFER, P. (1995) *Conceptology in Terminology Theory, Semantics and Word-Formation*. Viena: International Network for Terminology, IITF-Series 6.

WIERZBICKA, A. (1986) "What's in a Noun? (Or: How do Nouns Differ in Meaning from Adjectives?)". *Studies in Languages,* 10/2, 353-389.

---. (1988) *The Semantics of grammar*. Amsterdam: John Benjamins.

WÜSTER, E. (1998) *Introducción a la teoría general de la terminología y a la lexicografía terminológica*. Barcelona: IULA, Universitat Pompeu

Fabra.

ZWANENBURG, W. (1992) "Componding in French". *Rivista di linguistica*, 4/1, 221-240.