Relacions conceptuals i terminologia: anàlisi i proposta de detecció semiautomàtica

Judit Feliu i Cortès

2004

INSTITUT UNIVESITARI DE LINGÜÍSTICA APLICADA
UNIVERSITAT POMPEU FABRA

Programa de doctorat: Lingüística Aplicada (lèxic)
Bienni 1996-1998
Tesi doctoral

Direcció: Dra. M. Teresa Cabré Castellví

Per optar al títol de doctora per la Universitat Pompeu Fabra

Dipòsit legal: B.28046-2004
ISBN: 84-688-6999-6

Potser filosofar consisteix a intentar aclarir els embolics produïts pel llenguatge que emprem. Un d'aquests embolics és el fet de suposar que a cada paraula hi ha de correspondre en el món «alguna cosa» substantiva i tangible, quan moltes paraules tot plegat designen posicions, relacions o principis abstractes.

Fernando Savater
Les preguntes de la vida

Agraïments

Els agraïments solen ser el punt final d'un procés de treball molt llarg. Quan arriba aquest moment, és l'hora de repassar mentalment tota l'ajuda rebuda al llarg dels anys i reconèixer el suport dels altres per tirar endavant aquesta tasca.

Primerament, vull donar les gràcies a la directora d'aquesta tesi, la Dra. M. Teresa Cabré, per haver-me donat la possibilitat de treballar amb ella al llarg dels anys i per haver acceptat de guiar-me, corregint-me i animant-me, durant tot aquest temps.

També he d'agrair el suport i la confiança de tot el professorat d'aquesta Universitat des de l'any 1992 quan vaig arribar per primer cop a aquesta casa i, sobretot, als professors del doctorat *Lingüística aplicada (lèxic)*, del bienni 1996-1998, per despertar en mi l'interès per la recerca.

Així mateix, he de donar les gràcies a tots els companys de l'IULA amb qui he compartit moments durs i moments feliços, sobretot al Jesús per la seva disponibilitat i ajuda en l'etapa final d'edició d'aquest treball. I gràcies també als amics, a tots els que ho han estat fermament durant aquest període de temps, en especial, al Jordi, pels seus consells durant les converses que hem mantingut, i al Raúl, al John i a l'Elisabet.

Deixo per al final l'agraïment a la meva família sense el suport i la confiança de la qual no me n'hauria sortit. Per aquest motiu, gràcies al meu pare, en Francesc; als meus pares, la Núria i en Fèlix; al meu germà, en Dani, pels seus comentaris esperonadors; i a tota la resta de la família, des de l'avi gran que cada diumenge em preguntava si havia fet els deures fins a la meva fillola que, quan somriu, reparteix energia per tirar endavant.

Finalment, moltes gràcies al Xavier que, després de cinc anys sentint-me parlar de relacions, continua pensant que la relació més important és la nostra.

ÍNDEX DE CONTINGUTS

ÍNDEX DE FIGURES I TAULES

Capítol II

Capítol III

Capítol IV

Capítol V

Capítol 1

Introducció

Capítol I

1 Introducció

1.1 Antecedents[1]

Aquesta tesi doctoral neix d'un primer treball titulat *Relacions conceptuals i variació funcional: elements per a un sistema de detecció automàtica*, que va constituir el treball de recerca del Doctorat en Lingüística Aplicada cursat a l'Institut Universitari de Lingüística Aplicada durant el bienni 1996-1998. D'aquest primer treball van sorgir diverses futures línies de recerca. Una d'aquestes possibles vies de recerca s'orientava cap a la detecció semiautomàtica o assistida de les relacions conceptuals en terminologia, propòsit que ha esdevingut el tema central d'aquesta tesi de doctorat.

Les relacions conceptuals en terminologia, enteses com a element clau d'organització del coneixement, prenen un paper essencial quan ens proposem de recollir, i d'extreure, el coneixement especialitzat contingut en un discurs també especialitzat. Les unitats que trobem en el discurs vehiculen aquest coneixement però ho fan, sens dubte, mitjançant les relacions que s'estableixen entre aquestes unitats i sobre les quals farem una proposta de detecció semiautomàtica a partir dels textos.

[1] Aquesta tesi doctoral s'emmarca parcialment en el programa de concessió de beques predoctorals del Comissionat per a Universitats i Recerca (Direcció General de Recerca) de la Generalitat de Catalunya, del qual he estat beneficiària durant el període 1997-2000 (1997FI 00893 APLP).

1.2 Objectius

Aquesta tesi té com a objectiu final establir un prototip de sistema de detecció semiautomàtica de les relacions conceptuals. Per tal d'aconseguir aquest objectiu, el treball segueix diverses etapes, algunes de caire més teòric i d'altres sota un fil conductor més aplicat i que, en suma, ens permeten d'establir les línies mestres del prototip de sistema de detecció semiautomàtica de relacions conceptuals.

Per arribar a l'objectiu final, ens hem marcat dos objectius específics que detallem seguidament. En primer lloc, ens proposem validar empíricament en un corpus més ampli la tipologia de relacions conceptuals establerta a Feliu (2000). En aquell treball de recerca es va arribar a una primera tipologia de relacions conceptuals a partir de l'etiquetatge manual d'un corpus especialitzat sobre cardiopaties. En el marc d'aquest treball, comprovem que la tipologia de relacions conceptuals establerta continua sent vàlida no només en un conjunt de documents superior en nombre al del treball inicial sinó també en textos que pertanyen a un altre àmbit temàtic.

En aquest sentit, i amb la finalitat de validar la tipologia de relacions conceptuals, apliquem tots els marcadors lingüístics explícits que vehiculaven cada tipus i subtipus de relació conceptual al nou corpus i establim una nova llista de marcadors de relació conceptual. La llista definitiva que apliquem en aquest treball conté un nombre tancat de marcadors lingüístics explícits, essencialment verbals (acompanyats en alguns casos de preposicions i elements nominals de suport), que expressen els diferents tipus de relacions conceptuals que es donen en textos especialitzats.

El segon objectiu específic que conforma una de les aportacions essencials d'aquest treball es materialitza en l'establiment d'estratègies sintacticosemàntiques que ens permetin detectar amb el màxim de precisió els marcadors que expressen cada tipus de relació conceptual. Per fer-ho, proposem una llista de les unitats que hem establert *a priori* com a marcadors de relació conceptual, acompanyades d'una informació sintàctica mínima que respon als patrons sintàctics que es donen de manera repetida per a cada marcador. Seguidament, apliquem l'estratègia que anomenem *factor del context*, la qual ens ha de permetre, amb l'anàlisi semàntica de les unitats que

apareixen en el context d'una frase, determinar amb un grau de fiabilitat elevat si un marcador en particular expressa una relació conceptual o una altra. Aquesta estratègia de caire semàntic es basa en comprovar el lligam dels possibles candidats a unitats terminològiques, que apareixen vinculades per una determinada relació, amb una ontologia pròpia de l'àrea temàtica que treballem.

Més concretament, i orientats en tot moment a la construcció del sistema de detecció semiautomàtica de relacions conceptuals, establim les principals estratègies que ens porten a detectar una determinada relació a partir de marcadors verbals explícits, atribuir-li una o més d'una etiqueta de relació conceptual entre com a mínim dues unitats terminològiques i contribuir a enriquir la informació semàntica continguda en una ontologia a partir dels fragments textuals especialitzats que continguin unitats terminològiques i relacions conceptuals entre aquestes. La detecció de les unitats terminològiques combina estratègies automàtiques i la detecció manual, mentre que la detecció de les relacions conceptuals es basa en la informació sintacticosemàntica derivada de cadascun dels contextos especialitzats.

1.3 Hipòtesis de partida

De manera molt sintètica volem indicar en aquest apartat quina és la nostra hipòtesi de partida. En el marc de la Teoria Comunicativa de la Terminologia en què avança la nostra recerca, creiem que el coneixement especialitzat contingut en un text es transfereix mitjançant les unitats de coneixement especialitzat més prototípiques (a partir d'ara unitats terminològiques [UT]) i les relacions conceptuals que lliguen aquestes unitats en el discurs.

La nostra hipòtesi de partida és que el nombre i el tipus de relacions conceptuals que es poden donar en un document especialitzat és una llista tancada. Són, en canvi, les diverses manifestacions d'aquesta tipologia de relacions conceptuals les que poden expressar-se molt diversament ja sigui a través del lèxic, la tipografia o els connectors textuals. Així, podem partir d'una tipologia de relacions conceptuals molt ben definida, tipificada i tancada però, atesa la riquesa de la llengua, difícilment

podrem acotar mai tots els recursos que fem servir per expressar cada tipus de relació conceptual.

De tota manera, volem deixar palès que el tipus de relació conceptual, és a dir, la noció d'una determinada relació conceptual és diferent del marcador lingüístic, i si es prefereix lèxic, que l'expressa en un text. Per tant, en un sistema de formalització de la informació especialitzada continguda en un document, partim de la base que els conceptes i els tipus de relacions conceptuals entre aquestes unitats s'han de recollir en una ontologia, en un mòdul ontològic, el nucli de conceptes inicial del qual es pot establir prèviament a partir de l'ajuda de l'especialista i de diccionaris. La detecció de les relacions conceptuals a partir de fragments textuals especialitzats reals ens permet d'incorporar progressivament les diverses relacions que s'estableixen entre els conceptes continguts en l'ontologia. I, d'altra banda, és en el mòdul lèxic on es recullen les unitats lingüístiques, és a dir, les unitats terminològiques que contenen i transfereixen el coneixement especialitzat i, a més, tot i que separadament, les unitats lingüístiques explícites que vehiculen cada tipus de relació conceptual i uneixen d'una determinada manera cadascuna d'aquestes unitats.

1.4 Estructura de la tesi

En aquesta tesi doctoral, hem estructurat el contingut per capítols que segueixen l'ordre dels objectius específics que hem establert anteriorment. Així, el capítol segon recull els resultats essencials a què es va arribar en el treball de recerca de doctorat. Es recorda quina és la noció de relació conceptual amb què hem treballat des dels inicis de la recerca en aquest tema, es reprèn la tipologia de relacions conceptuals establerta en aquell treball i es presenten breument les principals característiques de cada tipus de relació.

En el capítol tercer, i per tal de demostrar el lligam estret entre ontologia i lèxic, es detalla quin és el tractament que fan de les relacions conceptuals les principals ontologies utilitzades per a la posterior recuperació d'informació. Es tracta de deixar palès que no se sol distingir entre la noció de relació i la unitat lèxica que expressa aquesta relació. Creiem que la distinció entre un mòdul lèxic i un mòdul conceptual

afavoreix la proposta posterior de sistema de detecció semiautomàtica de relacions conceptuals i millora notablement el procés d'actualització i enriquiment de l'ontologia sobre la qual treballarem.

En el capítol quatre, presentem el procés de constitució del corpus amb què hem treballat al llarg de la tesi. Per a cada tipus de relació conceptual, establim els possibles marcadors lingüístics que hem aplicat sobre el material textual i indiquem quins han estat els resultats obtinguts pel que fa a la precisió i al soroll segons les dades. Seguidament, presentem quines són les estructures més significatives per a cada tipus de relació que esdevenen el punt de partida per a la nostra proposta de sistema de detecció semiautomàtica.

En el capítol cinquè descrivim detalladament quines estratègies creiem que poden refinar el sistema de detecció semiautomàtica de relacions conceptuals i establim un possible lligam entre la detecció de les relacions conceptuals i la detecció de candidats a unitats terminològiques. Creiem que la suma de dues eines orientades a la detecció de candidats a termes i la detecció de possibles relacions entre candidats a termes ens ha de dur a obtenir un primer esquelet del coneixement especialitzat contingut en un text i, per tant, aquest coneixement pot ser reaprofitat en la compleció del mòdul lèxic d'una base de coneixements lligada a una ontologia.

En el capítol sisè presentem breument la proposta de sistema de detecció de relacions conceptuals de la manera més sistematitzada possible i, per tant acostada a l'automatització, tenint en compte que el perfil de l'autora d'aquest treball cobreix l'àmbit lingüístic però, més difícilment, els coneixements informàtics requerits per a una implementació d'aquesta índole.

Finalment, el capítol setè recull les conclusions d'aquesta tasca investigadora, en el capítol vuitè indiquem les referències bibliogràfiques que de manera directa o indirecta han alimentat el curs d'aquesta recerca i el CD-ROM annex conté el treball de recerca previ a aquesta tesi doctoral, el corpus d'anàlisi i la base de dades de marcadors verbals indicadors de relació conceptual.

Capítol 2

La noció de relació conceptual en terminologia: definició i tipus

Capítol II

2 La noció de relació conceptual en terminologia: definició i tipus

2.1 Descripció i antecedents

Aquesta tesi doctoral té com a antecedent immediat el treball de recerca de doctorat defensat al mes de maig de l'any 2000[1]. En aquest capítol volem descriure breument les principals línies de recerca abordades en aquell treball i els aspectes teòrics essencials que s'han de prendre com a punt de partida de la tesi que ara presentem.

De manera general, la hipòtesi de partida era que les relacions conceptuals descrites tradicionalment en la teoria terminològica eren insuficients, quant a nombre i diversitat, per donar compte dels lligams que efectivament s'estableixen entre les unitats de coneixement especialitzat que apareixen en els textos de qualsevol àrea temàtica. Per tal de demostrar aquesta hipòtesi, vam revisar totes les aportacions teòriques fetes des de la Terminologia pel que fa a la noció de relació conceptual. De fet, vam constatar que es tractava d'una noció poc definida atès que aquesta semblava caracteritzar-se principalment pels seus diferents tipus, és a dir, per

[1] El treball de recerca Feliu, Judit (2000) *Relacions conceptuals i variació funcional: elements per a un sistema de detecció semiautomatica*. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra es troba disponible i accessible en la seva totalitat en el CD-ROM que adjuntem com a annex al final d'aquest volum. Volem fer especial referència als capítols 2 i 3 pel que fa a la revisió bibliogràfica teòrica sobre el tractament de les relacions conceptuals en terminologia i a la definició de relació conceptual que vam establir en aquell treball i que mantenim com a punt de partida d'aquesta tesi doctoral.

nombroses tipologies que es reprenien les unes a les altres al llarg del temps sense presentar-ne una revisió teòrica aprofundida que qüestionés la validesa i la utilitat real de les relacions.

És per aquest motiu que vam ampliar la consulta amb propostes provinents d'altres camps teòrics, principalment la semàntica lèxica i la intel·ligència artificial i vam proposar una primera tipologia de relacions conceptuals que era el resultat híbrid entre la majoria d'aportacions que havíem analitzat. Aquesta tipologia va servir com a punt de partida per analitzar el corpus que havíem constituït sobre cardiopaties. Procedírem d'aquesta manera i vam poder corroborar la nostra hipòtesi inicial atès que la rigidesa de les tipologies presentades no abraçava la totalitat i diversitat de relacions conceptuals que s'establien en el corpus. Per aquest motiu, ens vam veure obligats a proposar una nova tipologia de relacions conceptuals, menys jerarquitzadora i basada en gran mesura en la realitat discursiva amb què ens havíem trobat.

En els apartats següents detallem de manera força més selectiva la informació que acabem de proporcionar al lector. En primer lloc, reprenem la nostra definició de relació conceptual perquè creiem que es tracta d'un element clau per a la comprensió posterior de tot el treball. Seguidament, presentem la tipologia de relacions conceptuals a què vam arribar en el treball de recerca i, per a cada relació conceptual, introduïm una descripció més aprofundida, sobretot pel que fa al catàleg de marcadors lingüístics que vam trobar que expressaven cada tipus de relació conceptual.

A mesura que la recerca en aquest camp ha anat avançant, hem considerat necessaris alguns petits canvis en la tipologia de relacions conceptuals. Específicament, i pel que fa als subtipus de relacions conceptuals, esmentem tres canvis o especificacions que hem considerat menors però necessaris per refinar també la nostra proposta de manera més adequada a les estructures trobades en els textos especialitzats.

Creiem, finalment, que, de la llista de marcadors possibles inicials de què partíem, ens trobem en el moment just per indicar quins s'han mantingut en una primera exploració del corpus que hem utilitzat en el marc d'aquesta tesi. Per tant, aquells

marcadors que s'han demostrat específics del treball anterior, han quedat al marge de l'anàlisi que presentem en el capítol quart d'aquesta tesi doctoral.

2.2 Les relacions conceptuals en terminologia

El nostre primer interès per estudiar els conceptes i les relacions conceptuals en aquest treball deriva del fet que aquests dos elements, juntament amb les classes conceptuals, són dos dels punts essencials dels esquemes cognitius, per tal com són elements clau del procés d'adquisició de coneixement (com aprehenem la realitat), de la categorització (com interioritzem la realitat) i de l'organització del coneixement (com emmagatzemem la informació al cervell). S'ha descrit àmpliament que, en qualsevol procés cognitiu, els humans detectem o construïm objectes individuals a partir de la realitat. D'aquests objectes individuals formem classes d'objectes per mitjà d'operacions diverses del pensament (com són la percepció[2], l'observació, la comparació i l'abstracció, entre d'altres). Es produeix, doncs, un procés de conceptualització, és a dir, de construcció de conceptes, que estableixen entre ells determinats tipus de relacions. El conjunt de les relacions que els conceptes d'un mateix àmbit especialitzat mantenen entre ells constitueix l'estructuració conceptual d'una matèria o camp de coneixement sobre un objecte. A través d'una selecció de conceptes i de relacions, organitzats en predicacions, verbalitzem el coneixement que tenim d'una matèria en concret.

2.2.1 Definició de relació conceptual

> «*Cada unidad terminológica corresponde a un nudo cognitivo dentro de un campo de especialidad y el conjunto de dichos nudos conectados por relaciones específicas (causa-efecto, todo-parte, contigüidad, anterioridad-posterioridad, etc.) constituye la representación conceptual de dicha especialidad. Si ello es así, no cabe duda de que mediante la terminología representamos la realidad especializada. Paralelamente a*

[2] En relació a les operacions del pensament, sobretot la sensació i la reflexió a partir de l'observació i l'experiència, vegeu J. Locke (1975: 104-401 i 720-721).

> *la representación de la realidad, categorizada en clases de conceptos relacionados, las unidades terminológicas sirven también para la transmisión de este conocimiento, es decir, para la comunicación*.» Cabré (1999b: 48)

Assumint el que es diu en el text anterior i partint de la idea que no existeix cap àmbit d'especialitat que no estigui estructurat i que l'estructura s'expressa a través d'un conjunt de relacions entre conceptes o nusos de coneixement expressats mitjançant unitats terminològiques, passem a analitzar i intentar definir a continuació què és una relació conceptual i quins tipus de relacions s'han postulat.

Otman (1996: 55-56) fa les següents aportacions a la noció de relació conceptual:

a) Una relació conceptual és un lligam conceptual entre diversos conceptes.

b) En un model relacional, un concepte es defineix pel conjunt de les relacions que manté amb els seus conceptes veïns.

c) Segons Otman, tota relació conceptual consta:

- d'un nom o **identificador**, que especifica el tipus de relació;
- de l'especificació del **tipus d'objectes** que aquesta relació admet;
- de l'**atribució** de determinades **propietats** a aquests objectes;
- *de vegades*, de determinades **condicions de validesa**.

Per tant, atès que una relació conceptual estableix sempre un lligam com a mínim binari, aquesta es pot representar (independentment de quina sigui la forma gràfica) per una fórmula del tipus:

a R b

on *R* representa la relació i *a* i *b* els dos conceptes implicats pertanyents a classes conceptuals determinades.

Com podem observar, de la proposta d'Otman es desprèn que les relacions conceptuals són més que simples lligams que uneixen els conceptes d'un determinat àmbit, perquè cada una d'aquestes relacions:

a) té un contingut semàntic en ella mateixa, és a dir, transmet una determinada informació;

b) presenta restriccions;

c) permet configurar una predicació específica entre els conceptes que enllaça.

Si, a més, considerem la possible recursivitat d'una relació conceptual, la formula anterior es modifica de la manera següent:

a R b, *n*

on la variable *n* indica que no es tracta només, ni sempre, de relacions binàries, sinó que aquestes poden tenir fins a *n* elements, tenint en compte que *n* no és il·limitat.

El repàs fet a totes les aportacions teòriques sobre relacions conceptuals i, concretament, l'estudi més detallat sobre relacions hiponímiques i meronímiques, ens porta a extreure algunes conclusions parcials abans de fer una primera proposta de tipologia de les relacions conceptuals:

— Es pot observar que el nombre de relacions conceptuals descrites en les aportacions tant en el camp de la terminologia com de la semàntica lèxica no és gaire extens.

— Els noms amb què es denominen aquestes relacions varien, en general, entre les aportacions des de la terminologia i les que es fan des de la semàntica lèxica, però en tots dos camps de coneixement, les relacions més ben descrites són la relació jeràrquica d'hiponímia/hiperonímia (anomenada d'inclusió de classe en semàntica lèxica) i la relació meronímica.

— Dins de l'àmbit de la terminologia, les aportacions de Wüster des de la teoria clàssica s'han anat reprenent per part dels autors posteriors però ni aquests, ni els organismes normalitzadors, no han arribat a establir una definició i una

classificació estables i funcionals de les relacions conceptuals. En la majoria de casos, per contra, s'inclou un calaix "altres" o "etc." que demostra que els autors no poden encabir totes les relacions conceptuals que, de fet, es donen en la comunicació especialitzada[3].

— Dins de l'àmbit de la semàntica lèxica, les aportacions de Lyons i Cruse han estat acceptades i represes sovint en treballs posteriors d'altres autors. Aquestes aportacions han permès donar compte de les relacions que s'estableixen entre els elements lèxics i, per tant, sembla que la terminologia se'n podria beneficiar directament, sobretot en el cas de les relacions meronímiques de les quals en lingüística s'ha fet una anàlisi molt més aprofundida que no pas en terminologia.

2.2.2 Tipologia de relacions conceptuals

La nostra primera proposta partia, bàsicament, de les aportacions fetes per les obres teòriques de terminologia, completades amb les relacions tipificades des de la semàntica lèxica quan aquestes no han estat recollides des de la terminologia o quan sembla que, pel fet d'haver estat descrites amb més detall, la seva extrapolació beneficia la llista de les relacions conceptuals que hauria de donar compte de qualsevol àmbit especialitzat. Amb aquesta descripció podrem comprovar si aquesta llista és o no suficient per donar raó de les relacions que apareixen en el corpus i si les descriuen adequadament.

Més concretament, per fer aquesta síntesi vàrem partir de la classificació de Wüster (que dóna compte de les relacions establertes entre els referents de les unitats lèxiques). Aquesta classificació basada en la referència coincideix només *grosso modo* amb les relacions generals de significat descrites per la semàntica lèxica.

---

[3] Aquesta afirmació es demostra, per exemple, en el llibre de G. Otman, sobre la representació semàntica en terminologia, on recull principalment les relacions genèriques (*sorte-de*) i les relacions partitives (*partie-de*), amb la finalitat d'establir i autormatitzar xarxes semàntico-terminològiques. Ara bé, aquest autor també descriu altres tipus de relacions, com ara: *fonction-de; proximité-de; contraste-avec; équivalent-de*. Aquest fet ens indica que les relacions proporcionades per les obres de teoria terminològica clàssiques no li permeten donar compte de la realitat que troba en els textos especialitzats i, per tant, n'ha de proposar de noves.

A més, vam incorporar la noció de *relació semàntica* (i no merament referencial) de D. Cruse i, doncs, vam assumir que les relacions conceptuals s'estableixen entre els significats de com a mínim dues unitats lingüístiques (lèxiques, sintagmàtiques o oracionals), i no entre els seus referents[4]. Així, vam observar que es poden donar quatre tipus de relacions semàntiques diferents entre els significats de les unitats. Només un d'aquests tipus es correspon amb una de les relacions tipificades per Wüster, la relació lògica.

La materialització de les relacions semàntiques en el discurs, analitzada i aplicada per Cruse a les relacions entre ítems lèxics, ens va servir directament en el nostre treball atès que partíem de textos especialitzats, i és a través de les expressions lingüístiques explícites o implícites d'aquests textos, que identifiquem les relacions que s'estableixen entre els conceptes. Pensem, però, que la nostra proposta complementa les aportacions de Cruse en el sentit que analitzem no només ítems lèxics relacionats semànticament, sinó també altres tipus d'unitats de coneixement especialitzat que poden aparèixer a la dreta i a l'esquerra d'una relació. La primera proposta de tipologia que aplicàrem en l'anàlisi de les dades recull, per tant, les quatre relacions semàntiques bàsiques de Cruse però més detallades amb la finalitat d'obtenir el màxim d'informació quan les contrastem amb les dades del corpus. Vegem, a continuació, la tipologia de relacions que inicialment vam aplicat al corpus que havíem constituir d'unes 12.000 paraules sobre cardiopaties:

TIPUS DE RELACIÓ CONCEPTUAL	SÍMBOL	SUBTIPUS DE RELACIÓ CONCEPTUAL	SÍMBOL	AUTORS	ESPECIFICACIÓ DE LES UNITATS RELACIONADES (a, b)	PROPIETATS[5]
Semblança	SEMBL.	positiva: - *equivalència total* o *sinonímia* - *equivalència parcial* o *semblança*	++ +	- CRUSE - CHAFFIN & HERRMANN		+ Sim. + Trans.
		negativa: - *oposició* - *contrast*	-- -	- ARISTÒTIL - CRUSE - OTMAN		+ Sim. + Trans.

[4] Cruse només fa servir la referència per confirmar que dues unitats presenten una determinada relació. Per això, aquest autor estableix l'existència de relacions semàntiques a partir de condicions de veritat i sobre la base de la materialització de les unitats en el discurs.

[5] La codificació seguida és la següent: (+) indica que es dóna una determinada propietat; (-) indica que no es dóna una determinada propietat; (~) indica que no necessàriament es dóna una determinada propietat; i (#) indica la no pertinència d'una determinada propietat, o de cap de les tipificades.

TIPUS DE RELACIÓ CONCEPTUAL	SÍMBOL	SUBTIPUS DE RELACIÓ CONCEPTUAL	SÍMBOL	AUTORS	ESPECIFICACIÓ DE LES UNITATS RELACIONADES (a, b)	PROPIETATS
Inclusió	INCL.	de classe o hiponímica	CLASS.	- WÜSTER - CRUSE - CHAFFIN & HERRMANN	*gènere-espècie*	- Sim. + Trans.
		topològica o espacial	TOPOL.	- WINSTON *et al.*		- Sim. + Trans.
Seqüencialitat	SEQ.	espacio-temporal: - *simultaneïtat* - *anterioritat-posterioritat*	ESP-TMP SIM. ANT-POS.	- WÜSTER - ARNZT & PICHT		- Sim. + Trans.
		causal	CAUS.	- WÜSTER - ARNZT & PICHT -NUOPPONEN	*causa-efecte*	- Sim. ~ Trans.
		procedural: - *amb afectació* - *sense afectació*	PROC. > <	- ISO/1087	*procés-resultat*	- Sim. ~ Trans.
		instrumental	INST.	- ARNZT & PICHT - DIN/2331	*instrument-ús de l'instrument*	- Sim. - Trans.
Meronímia	MER.	part-tot	P-T.	- WÜSTER - OTMAN - CHAFFIN & HERRMANN - WINSTON *et al.*	*component-objecte*	- Sim. ~ Trans.
					membre-col·lecció	- Sim. ~ Trans.
					porció-massa	- Sim. ~ Trans.
					material-objecte	- Sim. ~ Trans.
					característica-activitat	- Sim. ~ Trans.
					lloc-àrea	- Sim. ~ Trans.
Argumental	ARG.	paper-temàtic	ROL.	- CHAFFIN & HERRMANN	*agent-objecte* *agent-instrument* *agent-acció* *objecte-acció* *instrument-acció* *lloc-acció*	- Sim. - Trans.
Possessió	POSS.	atribució	ATRIB.	- WINSTON *et al.*	*element-qualificació*	- Sim. - Trans.
		adjunció	ADJ.		*element-localització*	- Sim. ~ Trans.
		pertinença	PERT.		*element simple-element complex*	- Sim. + Trans.

Taula 2-1. Tipologia inicial de relacions conceptuals (Feliu, 2000).

L'observació de les dades d'aquest petit corpus i la primera anàlisi de les relacions que s'hi donaven ens va fer constatar que:

— hi havia relacions que no quedaven recollides a la primera tipologia proposada que havíem pres com a punt de partida (*les troballes electrocardiogràfiques s'han de correlacionar sempre amb l'anamnesi, l'exploració física i, si pot ser, amb una radiografia de tòrax PA i lateral*, on es dóna una relació associativa expressada amb el verb *correlacionar*);

— que hi havia relacions que se solapaven les unes amb les altres (*l'onda T positiva pot presentar una negativitat final, encara que el segment ST estigui supradesnivellat*, on podem tenir una relació possessiva atributiva d'element-qualificació però, també, una relació meronímica de característica-activitat) i que, per tant, el nombre de relacions conceptuals tipificades es podria reduir.

Amb aquestes observacions, fruit de l'aplicació a les dades dels esquemes lingüístics prototípics que havien d'encaixar amb els fragments que expressaven alguna relació conceptual en el text, vàrem modificar la primera tipologia de relacions proposada al final del capítol segon del treball de recerca, procedent del recull de les aportacions teòriques especialment des de la terminologia i la semàntica, en el sentit de simplificar-la, resoldre els solapaments i afegir una relació conceptual associativa que apareix en els textos especialitzats de medicina i de la qual no podíem donar compte amb la tipologia anterior. Les relacions conceptuals recollides a partir de les aportacions teòriques i usades inicialment en l'observació de les dades queda doncs modificada de la manera següent:

TIPUS DE RELACIÓ CONCEPTUAL	SÍMBOL	SUBTIPUS DE RELACIÓ CONCEPTUAL	SÍMBOL	ESPECIFICACIÓ DE LES UNITATS RELACIONADES (a, b)	PROPIETATS[6]
Semblança	SEMBL.	positiva: - *equivalència total* o *sinonímia* - *equivalència parcial* o *semblança*	++ +		+ ⇔ + ⇒
		negativa: - *total o oposició* - *parcial o contrast*	-- -		+ ⇔ + ⇒
Inclusió	INCL.	de classe o hiponímica	CLASS.	*gènere-espècie*	- ⇔ + ⇒
Seqüencialitat	SEQ.	espacial: - *localització* - *direcció*	ESP. *loc.* ESP. *dir.*		- ⇔ ~ ⇒
		temporal: - *simultaneïtat* - *anterioritat-posterioritat*	TMP. *sim.* TMP. *ant-pos.*		- ⇔ + ⇒

[6] Reprenent les propietats descrites anteriorment, indicarem amb el símbol ⇔ la simetria entre dos elements i amb el símbol ⇒ la transitivitat que es dona entre, com a mínim, dos elements diferents. Els símbols + / - / ~ ens indiquen la presència, l'absència i la no necessarietat d'una determinada propietat.

TIPUS DE RELACIÓ CONCEPTUAL	SÍMBOL	SUBTIPUS DE RELACIÓ CONCEPTUAL	SÍMBOL	ESPECIFICACIÓ DE LES UNITATS RELACIONADES (a, b)	PROPIETATS
Causalitat		causal: - *causa-efecte* - *procés-resultat: amb afectació* - *procés-resultat: sense afectació*	CAUS. CAUS- PROC. > <		- ⇔ ~ ⇒
Instrumental	INST			*instrument-funció*	- ⇔ - ⇒
Meronímia	MER.	part-tot	P-T.	*component-objecte*	- ⇔ ~ ⇒
				membre-col·lecció	- ⇔ ~ ⇒
				porció-massa	- ⇔ ~ ⇒
				material-objecte	- ⇔ ~ ⇒
				etapa-procés	- ⇔ ~ ⇒
				característica-activitat	- ⇔ ~ ⇒
				lloc-àrea	- ⇔ ~ ⇒
Associació	ASS.				~ ⇔ ~ ⇒

Taula 2-2. Tipologia de relacions conceptuals revisada (Feliu, 2000).

Les materialitzacions lingüístiques considerades per a cada tipus de relació que serviren per a l'etiquetatge definitiu de les dades són[7]:

- Relació de semblança:
  - positiva: *ser semblant a;*
  - negativa: *ser diferent de.*
- Relació d'inclusió:
  - de classe o hiponímica: *ser (un tipus) de*.
- Relació de seqüencialitat:
  - espacial: *ser en / ser davant / ser darrere / anar de x a y*;
  - temporal: *ser simultani / anterior / posterior a*.
- Relació de causalitat:
  - causal: *causar / ser la causa de / ser l'efecte de*;
  - causal-procedural: *produir / fer que*.

[7] Com es pot observar, queden resolts els solapaments i les possibles paràfrasis ambigües que apareixien en la primera proposta de les materialitzacions lingüístiques prototípiques per a cada tipus de relació conceptual.

- Relació instrumental: *servir per a / fer-se amb*.
- Relació meronímica: *ser una part / element de; tenir + SN / estar format / fet per; incloure; constar de / pertànyer a*.
- Relació d'associació: *correlacionar-se amb*.

La tipologia que vam proposar com a definitiva difereix de la classificació inicial en els aspectes següents:

— Dins de les relacions d'inclusió, s'ha suprimit la relació topològica o espacial perquè creiem que la distinció establerta en la bibliografia entre aquesta relació i la relació meronímica lloc-àrea no es basa, realment, en una distinció entre dos tipus diferents de relació sinó entre els elements *a* i *b* que entren en joc en la relació. No es tracta, doncs, de dues relacions conceptuals diferents sinó que, en el cas de la inclusió topològica, la part és separable del tot mentre que en la relació de lloc-àrea, la part no és separable. Per tant, quan un element té una determinada localització però no forma part del lloc on es troba la relació que s'estableix és de seqüencialitat espacial locativa.

— Pel que fa a les relacions de seqüencialitat, s'han desdoblat les relacions espacio-temporals en dos subtipus diferents: d'una banda, tenim les relacions seqüencials espacials (i dins d'aquestes, les que indiquen localització i les que indiquen direcció) i, de l'altra, les relacions seqüencials temporals, que es mantenen de manera anàloga a com apareixien en la primera tipologia presentada. Tal i com acabem d'esmentar, aquest desdoblament permet donar compte de les relacions espacials que es donen en els fragments que hem recollit en el corpus quan la relació que s'hi estableix no és una relació meronímica lloc-àrea i, a més, evita la redundància que provocava la presència de la relació de possessió que indica adjunció (element-localització).

— Es pot observar, a més, que s'han separat les relacions de seqüencialitat espacials i temporals de les relacions causals i de les instrumentals. I, més específicament, la relació de seqüencialitat causal engloba, en la nostra proposta, dos subtipus de relacions: la relació causal pròpiament dita i la relació procedural (amb afectació o sense). Efectivament, una relació causal sempre indica una determinada

seqüencialitat però ho fa implícitament. En canvi, el contingut semàntic explícit de la relació és de causalitat i és per aquest motiu que hem introduït la relació causal i la relació causal procedural. En el primer cas, existeix una causa explícita que dóna lloc a un determinat efecte mentre que, en el segon cas, existeix un procés que proporciona un resultat encara que aquest procés no es concebi com una causa.

— S'ha mantingut també la relació instrumental, tot i que la seva freqüència d'aparició no sigui gaire elevada. En aquest cas, presentem l'especificació de l'element conceptual *b* com a *funció*, i no com a *ús de l'instrument*. Aquest canvi es justifica per una variació del punt de vista, ja que l'etiqueta *ús de l'instrument* requereix l'existència o l'aparició en el discurs d'un agent, mentre que *funció* es refereix directament a l'*instrument*, sense cap altra mena d'especificació.

— En el cas de les relacions meronímiques, s'ha mantingut la distinció inicial de les unitats que es poden vehicular per mitjà d'aquesta relació proposada per Winston *et al*. Ara bé, la nostra proposta diferencia explícitament les unitats que expressen etapa-procés d'aquelles unitats que expressen característica i activitat, per tal com hem considerat relacions meronímiques els casos en que el verb *tenir* se segueix d'un sintagma nominal que expressa característiques d'un tot.

— La nostra proposta introdueix una relació conceptual nova, la relació d'associació, per donar compte de les correlacions que s'estableixen entre dos o més elements. Aquesta relació no s'ha d'entendre com la semblança entre alguna o algunes de les característiques dels elements lligats, sinó com la simple existència d'algun punt de contacte entre aquests dos elements que fa possible que aquests es puguin connectar dins d'una mateixa unitat discursiva.

— Finalment, s'ha considerat que la inclusió de relacions argumentals i de relacions de possessió no és necessària perquè aquestes queden subsumides per les altres relacions tenint en compte que els matisos que alguns autors han considerat essencials per a incloure-les es poden especificar amb una anàlisi més detallada dels tipus d'unitats relacionades i de les seves propietats.

2.3 Resultats de l'aplicació de la tipologia de relacions conceptuals

2.3.1 Catàleg de relacions conceptuals

A partir de l'anàlisi de les dades, hem pogut extreure alguns elements de base per a la construcció d'una primera proposta de catàleg dels set tipus majors de relacions conceptuals establertes en el primer nivell de la tipologia proposada per a l'anàlisi de les dades. En el catàleg, per a cada tipus de relació tenim les informacions següents:

Símbol de la relació
/Descripció de la relació + nom complet/
Propietats de la relació
/Observacions sobre la RC/
Nombre d'arguments relacionals (2 / > 2)
Recursivitat de *b* (materialitzada en *n*)
Característiques de a (classe conceptual)
Característiques de b (classe conceptual)
/Observacions sobre *a* i *b*/
Expressió de la relació (unitat lingüística)

SÍMBOL: **SEMBL.**
DESCRIPCIÓ: /Relació de semblança: relació que s'estableix per equivalència o oposició entre dos o més elements/
PROPIETATS: *Simètrica i transitiva*.
OBSERVACIONS: /Presenta subtipus diversos: positiva i negativa/

ARGUMENTS: 2 o > 2

RECURSIVITAT: sí

CARACTERÍSTIQUES a*: entitat, activitat*

CARACTERÍSTIQUES b*: entitat, activitat*

/Observacions (*a*/*b*): necessàriament idèntiques

Expressió de la relació: *o*; *(_)*; *és a dir*; *assemblar-se a*; *al contrari que; ser el contrari a*.

SÍMBOL: **INCL.**

DESCRIPCIÓ: /Relació d'inclusió: relació que s'estableix per la inclusió d'algunes característiques d'un determinat element en un altre element/

PROPIETATS: *No simètrica i transitiva*.

OBSERVACIONS: /Presenta un únic subtipus: de classe o hiponímia/

ARGUMENTS: 2 o > 2

RECURSIVITAT: sí

CARACTERÍSTIQUES a*: entitat, activitat*

CARACTERÍSTIQUES b*: entitat, activitat*

/Observacions (*a*/*b*): necessàriament idèntiques

Expressió de la relació: *ser (+ SN)*; ∅ *(disposició textual)*; *considerar-se*; *com (x o y)*; [:].

SÍMBOL: **SEQ.**

DESCRIPCIÓ: /Relació de seqüencialitat: relació que s'estableix per localització o successió en l'espai o en el temps dels elements que uneix/

PROPIETATS: *No simètrica i transitiva (segons el subtipus)*

OBSERVACIONS: /Presenta subtipus: espacial i temporal/

ARGUMENTS: 2 o > 2 segons els subtipus

RECURSIVITAT: sí

CARACTERÍSTIQUES a*: entitat, activitat*

CARACTERÍSTIQUES b*: entitat, activitat*

/Observacions (*a*/*b*): no necessàriament idèntiques

Expressió de la relació: *iniciar-se en; produir-se en; tenir lloc a nivell de; quedar encarat amb; realitzar-se sobre; situar sobre; registrar-se en / a / des de; evidenciar-se a; originar a; veure's en; ocórrer (en presència de); aparèixer fins que: propagar-se a través de x / cap a y; continuar per / fins; arribar a / a través de; mesurar-se des de x fins a y; apropar-se a; allunyar-se de; produir-se cap a; correspondre primer a x i posteriorment a y; localitzar (primer) i després; (veure's) abans i després de; transcórrer des de x fins a y; ser seguit de / seguir-se de.*

SÍMBOL: **CAUS.**

DESCRIPCIÓ: /Relació de causalitat: relació que s'estableix entre una causa i el seu efecte/

PROPIETATS: *No simètrica i no necessàriament transitiva*

OBSERVACIONS: /Presenta subtipus: causal i procedural/

ARGUMENTS: 2 o > 2 segons els subtipus

RECURSIVITAT: sí

CARACTERÍSTIQUES a*: entitat, activitat, propietat*

CARACTERÍSTIQUES b*: entitat, activitat,propietat*

/Observacions (*a*/*b*): no necessàriament idèntiques

Expressió de la relació: *dependre de; fer que; ser la causa (principal) de; deure's a; implicar; aparèixer; contribuir a; dependre de; emmascarar; donar lloc a; <cond.> trobar:; renforçar; provocar; augmentar; produir; transformar-se en; augmentar; obtenir-se (+ ger.); donar lloc a; permetre (+inf.)*.

SÍMBOL: **INST.**
DESCRIPCIÓ: /Relació d'instrumentalitat: relació que s'estableix entre un instrument i la seva funció/
PROPIETATS: *No simètrica i no transitiva*
OBSERVACIONS: /No presenta subtipus/
ARGUMENTS: 2
RECURSIVITAT: sí
CARACTERÍSTIQUES a*: entitat, activitat*
CARACTERÍSTIQUES b*: entitat, activitat*
/Observacions (*a*/*b*): no necessàriament idèntiques
Expressió de la relació: *servir com a; realitzar-se amb.*

SÍMBOL: **MER.**

DESCRIPCIÓ: /Relació de meronímia: relació que s'estableix entre un element que constitueix un tot i els elements que conformen les seves parts/

PROPIETATS: *No simètrica i no necessàriament transitiva*.

OBSERVACIONS: /Presenta un únic subtipus: part-tot/

ARGUMENTS: 2 o > 2 segons l'especificació de les unitats relacionades

RECURSIVITAT: sí

CARACTERÍSTIQUES a*: entitat, activitat, propietat*

CARACTERÍSTIQUES b*: entitat, activitat, propietat*

/Observacions (*a*/*b*): no necessàriament idèntiques però una *propietat* requereix una entitat.

Expressió de la relació: *definir-se (tres grups):; constar de (les següents parts i fases); [:]; tenir; mostrar; incloure; caracteritzar-se per (la presència de); presentar; (dues zones):*.

SÍMBOL: **ASS.**

DESCRIPCIÓ: /Relació d'associació: relació que s'estableix per la correlació entre dos o més elements/

PROPIETATS: *No necessàriament simètrica ni transitiva*.

OBSERVACIONS: /No presenta subtipus/

ARGUMENTS: 2 o > 2

RECURSIVITAT: sí

CARACTERÍSTIQUES a*: entitat, activitat*

CARACTERÍSTIQUES b*: entitat, activitat*

/Observacions (*a*/*b*): no necessàriament idèntiques

Expressió de la relació: *correlacionar-se amb; correspondre a / considerar-se com a corresponent a; representar-se amb; reflectir; registrar / registrar-se en; ser imprescindible per realitzar; determinar; confondre's amb; simular; intervenir (en); suggerir; ser indicatiu de; veure's en; manifestar-se com; indicar / ser indicatiu de; presentar-se com a*.

2.3.2 Catàleg de marcadors lingüístics de relació conceptual

Presentem en aquest apartat els resultats obtinguts després d'aplicar la tipologia de relacions conceptuals al corpus de cardiopaties.

Quant a les unitats verbals que efectivament vehiculen una relació conceptual, constatem que:

- les relacions es poden expressar, d'una banda, per elements verbals lèxics autònoms —en general, verbs transitius (*causar*, *provocar*)—, i de l'altra, per verbs que requereixen l'aparició d'una o més preposicions (*correlacionar-se amb*, *veure's en*, *evidenciar-se en*, *transcórrer des de* x *fins a* y);

- alguns verbs transmeten un valor semàntic constant, és a dir, la unitat verbal transmet el significat expressat per la relació i només per a un tipus de relació (*causar*, *augmentar*), mentre que altres verbs tenen, segons les unitats conceptuals que relacionen, valors semàntics diferents (*registrar*, *implicar*).

Un cop fetes aquestes observacions, detallem a continuació els marcadors que representen en el corpus del treball de recerca cada una de les relacions conceptuals que detectades a partir de l'anàlisi dels materials. Presentem les unitats verbals i, en alguns casos no verbals, que vehiculen cada una de les relacions conceptuals en sis quadres, seguint cada un dels tipus de relació conceptual majors de la nostra tipologia d'anàlisi. Les taules inclouen tots els subtipus i especificacions dels elements de les relacions conceptuals malgrat que algunes no apareguin en els textos. Aquests buits —indicats en nota per a les relacions meronímiques— ens serviran, posteriorment, per detectar quines relacions o especificacions de les unitats no han aparegut en el corpus d'anàlisi.

Així, per a la relació de semblança trobem:

RELACIÓ DE SEMBLANÇA

Pos. ++ (total)	pos. + (parcial)	neg. -- (total)	neg. - (parcial)
o *(_)* *és a dir (4)*	*assemblar-se a*	*al contrari que* *ser el contrari a*	*diferenciar-se de*

Distingim entre la relació de semblança positiva i negativa i dins d'aquests dos grups, entre les que expressen una semblança positiva total (sinonímia) i les que expressen una semblança positiva parcial (diferents graus d'equivalència). I el mateix per a les relacions de semblança negativa total o parcial que expressen, respectivament, una oposició o un contrast. Segons les dades, disposem de pocs casos de semblança parcial, tant positiva com negativa, per poder fer-ne comentaris concloents. Tanmateix, sembla força destacable el paral·lelisme que s'estableix entre les relacions positives i les negatives. En els casos de positivitat o negativitat totals, aquestes relacions s'estableixen sempre entre dues unitats de coneixement especialitzat materialitzades en una unitat terminològica (UCE_R_UCE). En canvi, en el cas de les relacions parcials, el lligam s'estableix igualment entre dues unitats de coneixement especialitzat però es produeix un salt semàntic abans de la introducció de la segona unitat conceptual (UCE_R_Salt semàntic_UCE). Més concretament, les unitats *o, (_)* i *és a dir*, d'una banda, i *(ser) el contrari (que / a)* semblen funcionar de manera idèntica mentre que *assemblar-se a* i *diferenciar-se de* introdueixen un salt semàntic en la relació (per exemple: *si és positiva el vector s'apropa a ella i si es negativa s'allunya*). Abans de passar a la relació d'inclusió, volem comentar breument que tant el parèntesi com l'expressió *és a dir* presenten, en les nostres dades, un doble valor. Aquestes unitats poden vehicular relacions metalingüístiques però, també, relacions de semblança entre dues unitats conceptuals diferents.

La relació d'inclusió de classe s'expressa, en el nostre corpus d'anàlisi, per *ser + SN*; per la conjunció *com*, que indica "tipus de"; per una condició seguida d'una unitat verbal introductòria (*considerar-se*); per la presència dels dos punts (:); i, majoritàriament, per la disposició visual de la informació que separa un primer element conceptual (superordinat) dels seus tipus (subordinats) que es presenten en línies successives en el text, les quals segueixen un format d'esquema, tal com s'observa en el quadre següent:

RELACIÓ D'INCLUSIÓ

incl.
ser (+SN) *∅ (disposició textual)* <_> *considerar-se* *∅ com (x o y)* *:*

Observem que la relació d'inclusió s'estableix entre elements genèrics i els seus específics (hiperònim/hipònim), però també, entre unitats no lexicalitzades. Més concretament, i pel que fa a les relacions d'inclusió prototípiques, on els elements relacionats estan lexicalitzats, trobem:

— casos en què la disposició textual de la informació i/o l'ús dels dos punts introdueixen una llista tancada d'hipònims;

— casos en què una unitat lingüística introduïda en el discurs (p. ex. *com*) apareix seguida d'una llista oberta d'unitats de coneixement especialitzades que constitueixen l'hipònim de la unitat primera.

Encara dins dels casos d'inclusió prototípica, observem que la relació entre diferents elements pot estar restringida per una condició (per exemple: *Basant-nos en el registre d'una onda Q, un infart pot considerar-se: septal: xxx // anterior: xxx // anteroseptal: xxx // lateral: xxx // anteroseptal i lateral: la suma dels dos anteriors*) i, seguidament, introduïda per una unitat verbal (*considerar-se*). Aquesta condició representa una faceta o punt de vista determinat sobre un element a partir del qual es poden establir els seus subtipus. Remarquem, a més, que els tipus s'expressen aleshores per mitjà d'adjectius sense la repetició explícita de l'hiperònim.

Pel que fa a les relacions d'inclusió entre elements lexicalitzats, observem que la unitat conceptual primera pot ser una unitat lexicalitzada o no ser-ho, és a dir, pot ser una unitat de coneixement especialitzat fixada o una expressió oberta. En tots aquests casos, però, la unitat conceptual relacionada en segon lloc, és a dir, els hipònims, corresponen a unitats de coneixement poc o gens especialitzades. A més, aquestes unitats *b* apareixen amb una determinació explícita del tipus *últims llocs, situació que..., canvis més comuns, causes més usuals,* que podrien tenir un valor voluntàriament divulgatiu per situar el receptor.

Per al tercer tipus de relació conceptual, la relació de seqüencialitat, presentem seguidament les unitats verbals que expliciten els diferents subtipus d'aquesta relació:

RELACIÓ DE SEQÜENCIALITAT

esp. loc	esp. dir.	tmp. sim.	tmp. ant.-pos.
-iniciar-se en *-produir-se en* *-tenir lloc a nivell de* *-quedar encarat amb* *-realitzar-se sobre* *-situar sobre* *-registrar-se en / a / des de* *-evidenciar-se a* *-originar a* *-veure's en* *-ocórrer (en presència de)* *-aparèixer fins que (=quan)*	*-propagar-se a través de x / cap a y* *-continuar per / fins* *-arribar a / a través de* *-mesurar-se des de x fins a y* *-apropar-se a* *-allunyar-se de* *-produir-se cap a*		*-correspondre primer a x i posteriorment a y* *-localitzar (primer) i després* *-(veure's) abans i després de* *-transcórrer des de x fins a y* *-ser seguit de / seguir-se de*

De manera general, observem que la relació seqüencial locativa es manifesta essencialment mitjançant verbs diàdics. Els verbs triàdics apareixen, en canvi, en els casos en què la relació espacial indica direcció i quan la relació temporal expressa anterioritat i posterioritat on, necessàriament, els verbs requereixen l'aparició de més de dos arguments i, per tant, de diversos elements conceptuals lligats per aquesta noció de seqüencialitat.

Començant per la relació espacial locativa, distingim entre una localització espacial literal, és a dir, en un lloc concret material (*iniciar-se en el nòdul sinusal*) i les expressions de localització metafòrica, el significat de les quals és "lloc on es produeix una determinada manifestació" (*veure's en persones vagotòniques*). Aquest segon tipus de localització, materialitzada majoritàriament per la unitat verbal *veure's en*, s'ha de distingir, al seu torn, dels casos en què aquesta mateixa unitat, tal i com veurem més endavant, introdueix una relació associativa, però no de localització.

Encara dins de les relacions espacials, la divisió entre les relacions de localització i les de direcció respon a la distinció entre la situació d'un determinat element en un punt de l'espai o en un interval. Així, en els casos de localització, determinades

unitats expressen una localització exacta (*iniciar-se en*, *evidenciar-se a*) mentre que d'altres indiquen una localització aproximada (*tenir lloc a nivell de*)[8]. Per expressar l'interval, en canvi, trobem algunes unitats verbals que, per elles mateixes, transfereixen aquesta noció (*propagar-se a través de*) i d'altres que requereixen la presència de dues preposicions per expressar-la (*mesurar-se des de x fins a y*).

Sovint, les relacions seqüencials de direcció parteixen d'una relació seqüencial locativa per indicar l'origen i, a partir d'aquí, s'encadenen les altres unitats que expressen l'interval.

Passant ara a les relacions temporals, volem indicar que en el conjunt de les dades no apareix cap relació seqüencial temporal que indiqui simultaneïtat, és a dir, que dos elements coocorrin o coapareguin en un mateix moment. Contràriament, la relació de seqüencialitat temporal que indica anterioritat i posterioritat és força freqüent. Les dades ens mostren que aquest tipus de relació es pot expressar per mitjà d'un verb la semàntica del qual transfereix aquesta noció d'interval en el temps, però també per verbs menys autònoms que requereixen la presència de preposicions o d'elements adverbials per tal de vehicular aquesta noció. Alguns exemples del primer cas són *seguir-se de*, *transcórrer des de x fins a y* (en aquest cas, tot i la presència de preposicions, el verb ja indica un interval). Per al segon cas, trobem unitats del tipus *(veure's) abans i després de*, *correspondre primer a x i posteriorment a y*.

Vegem a continuació les unitats que expressen la relació causal, sigui en la seva expressió merament causal, sigui quan expressa un procés que dóna un resultat que pot presentar una afectació o mantenir l'estat inicial.

RELACIÓ DE CAUSALITAT

caus.	caus.-proc. >	caus.-proc. <
-dependre de *-fer que* *-ser la causa (principal) de* *-deure's a* *-implicar* *-aparèixer* *-contribuir a*	*-augmentar* *-produir* *-transformar-se en* *-augmentar*	*-obtenir-se (connectant / desplaçant / unint x amb y / prolongant)* *-donar lloc a* *-permetre (construir)*

[8] Per a una anàlisi més detallada, des d'una perspectiva cognitiva, sobre la semàntica de les preposicions vegeu Cuenca i Hilferty (1999), p. 143-145 i 200-208.

-dependre de *-emmascarar* *-donar lloc a* *-<cond.> trobar :* *-renforçar* *-provocar*		

Partint de la base que els tres subtipus tenen en comú una noció de causalitat més o menys explícita, observem que les unitats recollides en la primera columna indiquen una relació causal explícita prototípica en què *a* causar/provocar *b*. Altres unitats comporten, però, un canvi d'ordre dels elements conceptuals connectats i, en primer lloc, apareix l'efecte (*b*) i, seguidament, la causa (*a*), com per exemple *b* dependre de *a*. Encara en el marc de les relacions causals, s'observa que de vegades, la causa s'expressa o es restringeix per una condició i, seguidament, el verb que vehicula la noció de causalitat. En aquests casos també es produeix un canvi en l'ordre d'aparició de les unitats conceptuals relacionades.

Per a les relacions de causalitat procedural, constatem que en els casos en què el resultat ha patit una afectació, el verb mateix vehiculador de la relació indica aquest canvi. Per contra, en els casos en què el resultat no es veu afectat, la unitat verbal transfereix la noció de procés però, en cap cas, s'indica una alteració explícita de l'efecte resultant. Voldríem afegir que les unitats recollides en aquesta columna mereixen un comentari particular ja que en cinc dels sis casos el verb apareix seguit d'una altra unitat verbal que refina el seu significat, és a dir, *obtenir-se* es veu modificat per formes no personals, més concretament gerundis, que especifiquen el procés. Hem recollit aquestes unitats com a vehiculadores de la relació conceptual perquè restringeixen l'abast del verb principal que expressa la relació però aquestes s'haguessin pogut materialitzar en forma de nominalitzacions i, aleshores, aquestes nominalitzacions haurien passat a constituir la unitat conceptual *b*, com per exemple, *obtenir-se amb/mitjançant (la connexió de, la unió de*, etc.).

Molt breument, i pel que fa a la relació d'instrumentalitat, les unitats verbals que l'expressen són:

RELACIÓ D'INSTRUMENTALITAT

inst.
-servir com a *-realitzar-se amb*

Tal com hem comentat anteriorment, la freqüència d'aparició d'aquesta relació en el nostre primer corpus és molt baixa. Creiem, tanmateix, que és útil de mantenir aquest tipus de relació conceptual perquè, tot i que s'hauria de confirmar, probablement en textos especialitzats d'altres àmbits aquesta relació té una presència major. Segons les nostres dades, les dues úniques unitats que transmeten la noció d'instrumentalitat (instrument-funció) són les referenciades en la taula anterior. Podem comentar, però, que anàlogament a alguns casos de causalitat, es produeix una inversió de l'ordre de les unitats conceptuals vehiculades per l'expressió verbal *realitzar-se amb*, fet que, d'altra banda, no ocorre amb l'expressió *servir com a* (equivalent a *servir per*).

Per a les relacions meronímiques, i més concretament, segons les especificacions dels elements relacionats[9], les dades ens aporten les següents unitats verbals com a transmissores d'aquesta relació:

RELACIÓ DE MERONÍMIA

P-T. comp.-obj.	P-T. etapa-procés	P-T. caract.-activ.	P-T. lloc-àrea
-definir-se (tres grups):	*-constar de (les següents parts i fases):* *- :*	*-tenir* *-mostrar* *-incloure* *-caracteritzar-se per (la presència de)* *-presentar*	*-(dues zones) :*

Destaquem que la presència de la relació meronímica en l'àmbit temàtic de la cardiopatia queda gairebé restringida a les unitats que indiquen característiques de determinades activitats i a les etapes que es donen o cal seguir per acomplir un determinat procés[10]. Les altres especificacions possibles de les unitats relacionades

[9] No recollim una columna per a les especificacions de *membre-col·lecció*, *porció-massa*, i *material-objecte* perquè aquestes no apareixen en les dades del nostre corpus d'anàlisi.

[10] Inicialment, hem separat aquesta relació proposada per Winston *et al*. en dues relacions diferents perquè, d'una banda, s'expressen per mitjà de materialitzacions lingüístiques diferents i, de l'altra, i principalment, perquè hem considerat parts d'un tot les característiques, enteses com a propietats, que una determinada unitat pugui tenir i que es materialitzen en els textos generalment especificades per unitats de mesura (alguns exemples són: *durada*, *llargada*, *negativitat*).

per una relació meronímica són gairebé inexistents i, per tant, podem afirmar que la relació preeminent en aquest àmbit mèdic concret és la relació meronímica part-tot, en què els elements indiquen característica-activitat. Ara bé, cal indicar que aquesta relació s'ha considerat tradicionalment prototípica, ja que tot i que s'expressa majoritàriament amb el verb *tenir*, la unitats conceptuals que representen les parts són qualitats mesurables.

Les relacions meronímiques pròpiament dites es materialitzen en unitats que expressen *component-objecte*, *etapa-procés* i *lloc-àrea*. En aquests casos, els fragments analitzats permeten observar que pot ser que s'avanci el nombre d'elements de l'enumeració posterior (sobretot quan s'introdueix per mitjà d'expressions del tipus *dues zones: / tres grups*); o bé que no s'avanci el nombre d'elements que constitueixen l'enumeració següent (per exemple amb *incloure, : , caracteritzar-se per*).

Finalment, i pel que fa a la relació associativa, és important de destacar l'alta freqüència d'aparició d'aquest tipus de lligam en aquest àmbit temàtic concret. Aquesta relació s'estableix sobre la base inicial d'una determinada proximitat semàntica o pragmàtica entre els elements connectats. Aquesta proximitat permet, per mitjà d'unitats verbals força diverses, unir dos elements que poden no compartir trets semàntics però que es troben relacionats entre si temàticament. Les unitats que expressen aquesta relació conceptual són:

RELACIÓ D'ASSOCIACIÓ

ass.
-correlacionar-se amb
-correspondre a / considerar-se com a corresponent a
-representar-se amb
-reflectir
-registrar / registrar-se en
-ser imprescindible per realitzar
-determinar
-confondre's amb
-simular
-intervenir (en)
-suggerir
-ser indicatiu de
-veure's en
-manifestar-se com
-indicar / ser indicatiu de
-presentar-se com a

D'entre totes aquestes unitats podem distingir dues tendències diferents. D'una banda, trobem expressions verbals del tipus *correlacionar-se amb* o *correspondre a* que expressen una relació recíproca entre els elements que vehiculen entre els quals s'estableix una relació simètrica. D'altra banda, els altres verbs vehiculen una relació entre les unitats conceptuals connectades però aquesta no és necessàriament simètrica (per exemple: *registrar*, *simular*, *representar*).

Volem destacar, finalment, que algunes de les unitats recollides per a aquest tipus de relació han aparegut prèviament per a algunes de les relacions anteriors. Així, el verb *implicar*, per exemple, introdueix una relació de causalitat en alguns casos i, d'associació en d'altres; el l'expressió verbal *veure's en* serveix per indicar una localització real o per expressar només una associació entre dos elements. En aquests casos, són els elements relacionats els que ens han servit per decidir si es tracta d'un o altre tipus de relació conceptual.

2.4 Primera exploració sobre el corpus: canvis i desaparicions

2.4.1 Sobre les relacions conceptuals

En aquest apartat, volem indicar tres petites modificacions que adapten la tipologia de relacions conceptuals al treball fet al llarg del temps sobre un nombre variat de textos i en la tasca diària de representar les relacions en una eina de modelització de la informació especialitzada com és una ontologia.

En primer lloc, i seguint l'ordre seqüencial de la tipologia presentada, creiem que efectivament cal mantenir la relació de causalitat però considerem que no és necessari distingir entre les relacions causals i les relacions procedurals atès que, en tota relació causal, existeix un procés. La relació procedural quedava inicialment justificada perquè s'intentava indicar si el resultat del procés, és a dir, l'efecte en una relació causal, havia rebut una afectació o si bé quedava, tot i haver patit un procés,

igual que al principi. En aquest cas, ens trobàvem davant d'una informació que, per força, feia necessària l'aparició de tres elements: el concepte a (objecte causal) i el concepte b (objecte causat) i un tercer concepte c (canvi). Aquesta última informació no apareix majoritàriament en els contextos analitzats. Per aquest motiu, hem decidit englobar sota el tipus únic de relació causal els casos en què es tracti d'un procés que té un resultat i els casos en què es vegi molt clarament que es tracta d'una causa i el seu efecte ja que, en totes dues ocasions, ens trobem davant d'una relació binària, que lliga dos conceptes, i que té una noció semàntica de base idèntica. Dit amb altres paraules, l'afectació o no del resultat és una informació no sempre explícita i, per tant, l'etiquetatge del tipus d'informació esdevé una tasca força difícil. És per aquest motiu que hem cregut més adient etiquetar els contextos amb la noció de relació causal sabent que, en alguns casos, apareix efectivament la noció de causa mentre que, en d'altres casos, es produeix un procés que dóna lloc a un resultat, considerats sinònims de la noció de causalitat.

En segon lloc, creiem que en el cas de la relació meronímica no és necessari mantenir la distinció entre el subtipus *etapa-procés* i el subtipus *característica-activitat*. Si reprenem la definició que per a aquest últim subtipus ens donen els màxims estudiosos de la noció de meronímia (Winston, Chaffin i Herman, 1987), veiem que el subtipus característica-activitat serveix per designar les característiques o les fases d'activitats i processos (per exemple: *pagar és part de comprar*). En un primer moment, havíem proposat aquesta distinció de donar compte d'estructures del tipus *l'edifici té una alçada de vuit metres i una amplada de nou*. Ara bé, tenint sempre com a objectiu final la formalització de la informació especialitzada, hem vist que aquest tipus d'informació es pot considerar de manera metafòrica com a *elements* o *parts* no físiques d'un tot però, per motius de modelització de la informació en una ontologia, la informació sobre els conceptes que pot aparèixer amb el marcador lingüístic explícit *tenir* és una informació abocable directament com a atributs del concepte *a* i, si som afortunats, el mateix context d'anàlisi ens proporcionarà el valor d'aquest atribut. Així, en el cas del concepte 'edifici', aquest tindria un atribut que és 'alçada' amb un valor de 'vuit metres' i un altre atribut que és 'amplada' que té com a valor 'nou metres'.

Finalment, creiem que dins la relació d'associació podem establir com a mínim dos tipus. Un primer subtipus de relació associativa general, que es vehicula a través de marcadors lingüístics explícits propis de la llengua en general i que no canvien de significat encara que canviem d'àrea temàtica o de corpus d'anàlisi. I un segon subtipus de relació associativa que és especialitzada i que ve determinada per l'àrea temàtica amb què estem treballant. Així, per al subtipus general, obtenim marcadors lingüístics del tipus *mostrar* o *indicar* mentre que per al subtipus específic tenim, per a l'àmbit de les cardiopaties, per exemple, *correlacionar-se amb*, *veure's en* i per a l'àrea temàtica del genoma humà tenim unitats com *determinar*.

De tot el que acabem de dir, es desprèn que el catàleg que hem presentat i la llista de patrons lingüístics utilitzada són igualment vàlids però, tanmateix, volem indicar explícitament com quedaria la tipologia de relacions conceptuals ara sí definitiva:

TIPUS DE RELACIÓ CONCEPTUAL	SÍMBOL	SUBTIPUS DE RELACIÓ CONCEPTUAL	SÍMBOL	ESPECIFICACIÓ DE LES UNITATS RELACIONADES (a, b)	PROPIETATS
Semblança	SEMBL.	positiva: - *equivalència total* o *sinonímia* - *equivalència parcial* o *semblança*	++ +		+ ⇔ + ⇒
		negativa: - *total o oposició* - *parcial o contrast*	-- -		+ ⇔ + ⇒
Inclusió	INCL.	de classe o hiponímica	CLASS.	*gènere-espècie*	- ⇔ + ⇒
Seqüencialitat	SEQ.	espacial: - *localització* - *direcció*	ESP. *loc.* ESP. *dir.*		- ⇔ ~ ⇒
		temporal: - *simultaneïtat* - *anterioritat-posterioritat*	TMP. *sim.* TMP. *ant-pos.*		- ⇔ + ⇒
Causalitat	CAUS.	causal	CAUS.	*causa-efecte* *procés-resultat*	- ⇔ ~ ⇒
Instrumental	INST.			*instrument-funció*	- ⇔ - ⇒
Meronímia	MER.	part-tot	P-T.	*component-objecte*	- ⇔ ~ ⇒
				membre-col·lecció	- ⇔ ~ ⇒
				porció-massa	- ⇔ ~ ⇒
				material-objecte	- ⇔ ~ ⇒
				etapa-procés	- ⇔ ~ ⇒
				lloc-àrea	- ⇔ ~ ⇒
Associació	ASS.	general	ASS. gen.		~ ⇔ ~ ⇒
		especialitzada	ASS. esp.		

Taula 2-3. Tipologia definitiva de relacions conceptuals.

2.4.2 Sobre els marcadors lingüístics de relació conceptual

De la llista proposada en aquest capítol, hi ha alguns marcadors que no apareixen en l'anàlisi del nou corpus que hem constituït sobre el genoma humà. Així, en el capítol quart, després de la descripció acurada del corpus de treball, presentem una anàlisi detallada de cada marcador lingüístic de relació conceptual, tant pel que fa als resultats de soroll que dóna com al grau de precisió en vehicular la relació conceptual

que transmet. D'aquesta anàlisi extraurem els marcadors verbals per a la proposta de detecció semiautomàtica de relacions conceptuals.

A mode de representació gràfica dels marcadors que han de ser estudiats de nou en un corpus de diferent àrea temàtica i amb un nombre superior d'ocurrències per poder afirmar sens dubte que aquests no s'han de tenir en compte per a futures aplicacions, presentem la taula següent on, per a cada tipus i subtipus de relació conceptual, indiquem el marcador que no es manté del corpus de cardiopatia al corpus del genoma humà.

RELACIÓ CONCEPTUAL	MARCADOR LINGÜÍSTIC
Relació de semblança	*ser el contrari*
Relació d'inclusió	∅
Relació de seqüencialitat: *Espacial-locativa* *Espacial-direcció* *Temporal-simultània* *Temporal-anterioritat/posterioritat*	 *produir amb; registrar + preposició* ∅ ∅ *anar seguit de; estar seguit de*
Relació de causalitat	*fer aparèixer; emmascarar*
Relació instrumental	∅
Relació meronímica	∅
Relació associativa	*(en)registrar; ser indicatiu de*

Taula 2-4. Marcadors verbals no presents en el corpus de genoma humà.

La informació continguda en aquesta taula constata que són pocs els marcadors que no podran ser descrits i exemplificats en el capítol quart, que és on tractem amb força detall quins són els marcadors lingüístics de relació conceptual més productius i que esdevenen l'eix fonamental per al disseny del sistema de detecció semiautomàtic o assistit de relacions conceptuals a partir de textos especialitzats.

Capítol 3

Relacions conceptuals i ontologia

Capítol III

3 Relacions conceptuals i ontologia

3.1 Introducció[1]

L'objectiu d'aquest capítol és avaluar algunes de les ontologies existents per demostrar que en aquest camp, l'ús de les relacions conceptuals té un paper essencial que, en molts casos, queda curt a l'hora de donar compte efectivament de les relacions que s'estableixen en un determinat àmbit temàtic. A més, en algunes de les ontologies analitzades s'observa un solapament entre la noció de relació mateixa que es vol recollir en l'ontologia i la unitat de la llengua que la vehicula. Ens proposem, doncs, d'indicar en quins casos creiem que una tipologia de relacions conceptuals al darrera d'una ontologia permetria organitzar de manera més adequada el coneixement d'un determinat àmbit especialitzat i, a més, ens proposem de justificar la distinció que hem adoptat entre el contingut de l'ontologia i el contingut que caldria trobar en un mòdul lèxic, on el lligam amb l'ontologia es presenta com a essencial.

És àmpliament sabut i compartit que les noves tecnologies de la informació són utilitzades per al tractament i la transferència de gran quantitat de documents i, conseqüentment, de la informació que aquests contenen. Considerant l'augment incessant de dades, un dels principals objectius de la gestió de la informació és tenir

[1] En el camí de la recerca duta a terme en el transcurs d'aquesta tesi doctoral destaca la feina feta en la descripció del vincle entre les relacions conceptuals i l'ontologia, aspecte que s'ha estudiat en el marc d'un projecte finançat dins de l'Institut Universitari de Lingüística Aplicada.

accés i recuperar els documents pertinents. Els sistemes de recuperació d'informació (RI) actuals tenen una efectivitat limitada per tal com gairebé només utilitzen informació estadística i, en el millor dels casos, es basen en una anàlisi lingüística de baix nivell. En cap cas, però, tenen accés a cap tipus d'informació semàntica.

Un dels projectes actuals de l'Institut Universitari de Lingüística Aplicada, anomenat Banc de coneixements GENOMA[2], agrupa els resultats de dos projectes finançats. En el marc d'aquesta recerca es pretén anar un pas més enllà en l'anàlisi discursiva, gramatical i semàntica dels textos especialitzats. Els esforços s'orienten principalment a la caracterització de les unitats lèxiques (simples o complexes) que es troben en el discurs especialitzat i que constitueixen la terminologia d'un determinat àmbit. L'objectiu final és construir un sistema de detecció automàtica de les estructures cognitives subjacents en els textos especialitzats. Per arribar a obtenir un primer prototip d'aquest sistema, es treballa en la detecció assistida d'unitats terminològiques, en el mapeig semiautomàtic dels nodes cognitius presents en un text i, amb la recerca presentada en aquesta tesi, en la representació i detecció dels lligams entre aquests nodes, això és, en les relacions conceptuals en els textos especialitzats. En el vessant més aplicat, es treballa per construir un sistema de recuperació d'informació capaç de millorar els sistemes actuals utilitzant el control terminològic. Per fer-ho, s'està aprofitant tota la informació gramatical, semàntica i pragmàtica associada a les unitats que transmeten coneixement especialitzat, i una part important d'aquesta informació és la que proporcionen les relacions conceptuals.

La metodologia utilitzada en aquest cas combina una eina capaç de dur a terme el processament del llenguatge natural atès que inclou marcatge estructural, anàlisi morfològica i sintàctica, tècniques de desambiguació i un sistema de detecció de candidats a termes basat en patrons formals i en una ontologia lèxica.

2 La construcció i explotació del Banc de coneixements GENOMA es duu a terme gràcies al finançament de dos projectes públics finançats: texterm i ricoterm. TEXTERM: Textos especializados y terminología: selección y recuperación automática de la información (BFF2000-0841), dirigit per M. T. Cabré; i RICOTERM: Sistema de recuperación de información con control terminológico y discursivo (TIC2000-1191), dirigit per M. Lorente.

La figura següent mostra l'arquitectura general del projecte Base de Coneixements GENOMA, on podem veure com un dels elements centrals l'ús d'una ontologia que lligui totes les unitats lèxiques de la base de dades terminològica al coneixement recollit en aquesta formalització (vegeu Figura 3-1).

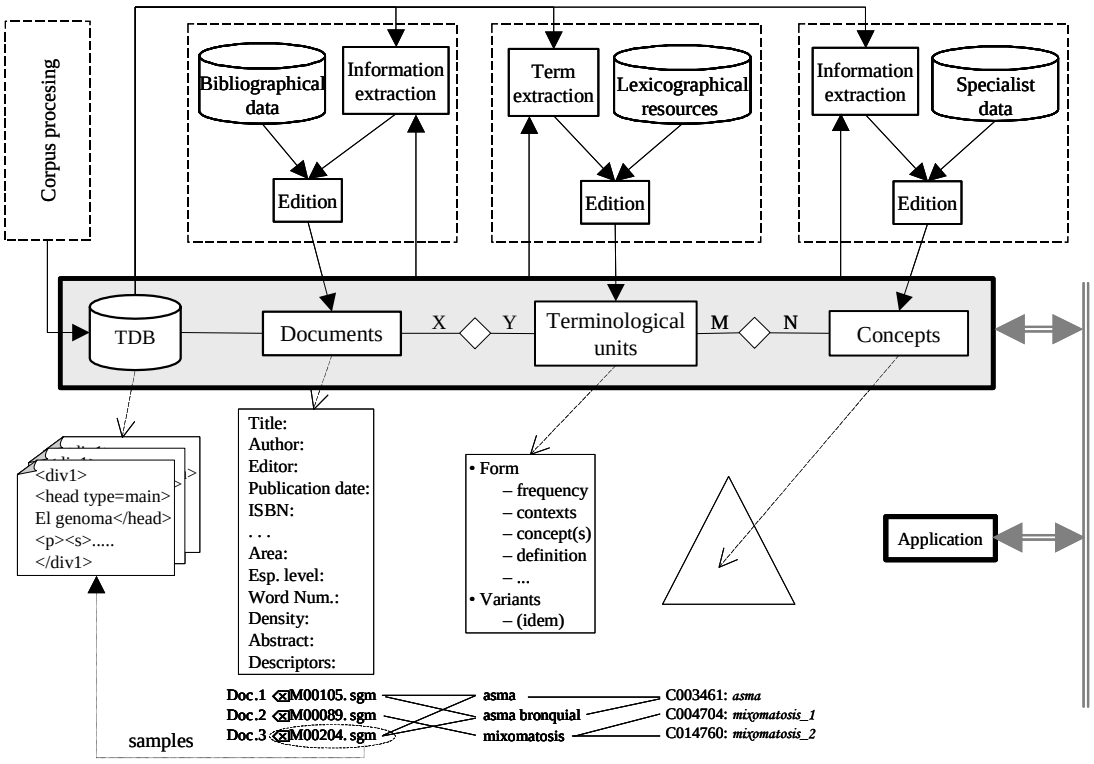


Figura 3-1. Descripció general: Base de Coneixements GENOMA.

Com es pot observar, la Figura 3-1 mostra l'estreta relació entre la base de dades terminològica, això és, les unitats de coneixement especialitzat, i els conceptes i els documents relacionats amb l'àmbit del genoma humà, que constitueix l'àrea temàtica central del projecte. L'ontologia s'utilitza per classificar i estructurar el coneixement especialitzat extret del corpus i, per dur a terme aquesta tasca de la manera més fidel possible a la informació continguda en els documents, cal que les relacions conceptuals previstes en l'ontologia siguin efectivament les relacions que trobem en l'àrea temàtica del genoma humà, és a dir, que existeixi una tipologia de relacions conceptuals general per al coneixement especialitzat que sigui útil per mapejar en una ontologia el coneixement especialitzat recuperat a partir del discurs especialitzat.

Després d'aquesta breu descripció dels projectes en què s'emmarca aquesta tesi doctoral, el capítol presenta algunes consideracions generals sobre què és una ontologia (secció 2). Seguidament, descrivim i analitzem les cinc ontologies més conegudes i utilitzades en projectes vigents (secció 3), i passem a presentar una anàlisi comparativa entre aquestes cinc ontologies avaluades d'acord amb els paràmetres de disponibilitat, capacitat de maneig (ampliació i modificació), expressivitat, àmbit d'aplicació, grandària, granularitat, completesa i tractament de les relacions conceptuals.

Finalment, en aquest capítol, mostrem l'estat actual de l'ontologia sobre el genoma humà tal com es troba en el moment de la recerca, posant especial èmfasi en el tipus de relacions conceptuals que s'hi han inclòs, algunes mostres de relacions entre conceptes i el tractament que rebrà el lèxic en relació a aquesta ontologia, tant pel que fa als conceptes que expressen entitats o propietats com pel que fa als conceptes que expressen relacions.

3.2 Ontologies: consideracions generals

> «An ontology is an explicit specification of a conceptualization. The term is borrowed from philosophy, where an ontology is a systematic account of Existence. For knowledge-based systems, what "exists" is exactly that which can be represented. When the knowledge of a domain is represented in a declarative formalism, the set of objects that can be represented is called the universe of discourse. This set of objects, and the describable relationships among them, are reflected in the representational vocabulary with which a knowledge-based program represents knowledge.
>
> Thus, we can describe the ontology of a program by defining a set of representational terms. In such an ontology, definitions associate the names of entities in the universe of discourse (e.g., classes, relations, functions, or other objects) with human-readable text describing what the

> names are meant to denote, and formal axioms that constrain the interpretation and well-formed use of these terms.»
>
> (T. Gruber, 1993)

Segons T. Gruber, una ontologia és una especificació explícita d'una conceptualització. L'autor parla "del que existeix" que, en el nostre cas, esdevé el que trobem en els textos especialitzats, és a dir, els conceptes expressats mitjançant unitats de coneixement especialitzat (o termes), relacions conceptuals i altres tipus d'informacions lingüístiques útils per a propòsits terminològics. Tot aquest material cognitiu ocuparà un lloc essencial en l'ontologia i el conjunt de textos especialitzats del nostre corpus serà el nostre univers de discurs, d'on extraurem tota aquesta informació.

La paraula *ontologia* s'ha manllevat de la filosofia i s'ha extrapolat a la Intel·ligència Artificial (IA). Les ontologies es conceben com a artefactes dissenyats o formulats per a propòsits específics i avaluats tenint sempre en compte els seus objectius explícits en el criteri de disseny. Des de la IA, la funció de les ontologies és donar suport a les activitats d'intercanvi d'informació i, per aquest motiu, els investigadors d'aquest àmbit intenten establir criteris que hauran de guiar el desenvolupament de les ontologies orientades a la representació, la transferència i la recuperació d'informació.

En un sistema de representació de coneixement, el terme *ontologia* es refereix a tots els conceptes d'un determinat àmbit temàtic, A més, aquest terme sovint s'usa per descriure els conceptes, les seves relacions i les restriccions que cal tenir en compte per tal d'obtenir una modelització d'un àmbit[3]. En aquest cas, una ontologia es pot entendre com la representació formal d'un àmbit especialitzat que conté la seva pròpia terminologia.

[3] Per a més informació sobre la relació entre representació del coneixement i terminologia, vegeu la següent url: http://www.biomath.jussieu.fr/~pz/Publications/ZweigenbaumISIS99/isis99.html.

Terminologia es defineix, en una de les seves accepcions, com el conjunt de termes en un àmbit. I un terme es defineix, així mateix, tenint en compte el nombre de relacions centrades en una unitat lèxica, en la majoria de casos lexicalitzada. Per aquest motiu, sembla lògic utilitzar ontologies amb l'objectiu de mapejar el coneixement especialitzat contingut en un corpus temàticament específic i, per tant, per descriure les unitats de coneixement especialitzat transferides per les unitats lingüístiques (o termes) i les relacions entre aquestes unitats.

El mapeig del coneixement especialitzat és una tasca difícil i, per això, cal dedicar el màxim d'esforç en el disseny de l'ontologia. El criteri adoptat per al disseny de l'ontologia esdevé un punt de partida essencial. Algunes de les decisions bàsiques que cal prendre concerneixen:

a) La cobertura que es requereix a l'ontologia, és a dir, el nombre de conceptes que ha de recollir.

b) La finalitat de l'aplicació que utilitzarà l'ontologia, per tal com les característiques de l'ontologia (àmbit, representació dels nodes, etc.) darrere d'una eina en particular es veuran condicionades per les restriccions de l'aplicació final. Els requisits que es demanen a una ontologia que s'usi en una *semantic web* o en un sistema de traducció automàtica són molt diferents.

c) Els nodes superiors. Tradicionalment, els nodes superiors de les ontologies han estat entitats, propietats i relacions. Tanmateix, en alguns casos, el nombre de nodes superiors pot incrementar-se i diferir (per exemple, WordNet [WN] utilitza onze *top nodes* sense relacions entre ells).

d) Les relacions conceptuals que es permeten a l'ontologia. La relació d'hiponímia-hiperonímia "*is-a*" és la relació bàsica de qualsevol ontologia però també són possibles altres tipus de relacions, i en algunes aplicacions, més que possibles, les relacions conceptuals esdevenen necessàries. Sembla doncs que, *a priori*, el nombre de relacions conceptuals no és una llista tancada. Augmentar el nombre de relacions enriqueix l'ontologia però s'esdevé més difícil mantenir-ne la consistència.

e) Ús de l'herència. L'herència és un mecanisme general per afegir informació a un node en particular d'una manera compacte i molt fàcil de mantenir. D'acord amb aquest mecanisme, una informació concreta és compartida per un determinat node i tots els seus hipònims. L'herència monotònica simple és el mecanisme més senzill. Consisteix en l'herència de les propietats d'un node només d'un antecessor i aquest valor no pot ser eliminat ni substituït en cap punt de l'ontologia. El mètode de l'herència té, no obstant, alguns problemes per manejar les situacions reals. Aquesta situació es pot superar utilitzant el que es coneix com a herència múltiple (cada node pot heretar propietats de més d'un pare) i/o mitjançant l'herència per defecte (un node pot sobreescriure localment el valor d'una propietat heretada). El problema sorgeix quan un node hereta valors incompatibles per a una propietat en particular, valors que provenen de diferents pares. Els mecanismes que acabem d'esmentar no solucionen aquests problemes però es poden adoptar algunes solucions basades en un control profund de la jerarquia. Per exemple, l'herència ortogonal suggereix agrupar la informació i permetre l'herència múltiple només des d'alguns grups.

f) Representació dels nodes. Els conceptes, que configuren els nodes de l'ontologia s'han d'indicar per mitjà d'una etiqueta o convenció (majúscules, números, etc.) o utilitzant informació estructurada (estructures de trets).

Tradicionalment, i al marge dels criteris que acabem d'esmentar, les ontologies s'han classificat seguint diversos paràmetres com a:

- generals (p.ex., WN [Fellbaum, 1998]) o específiques (p.ex., ULMS [NLM, 1998]),

- genèriques (p.ex., WN) o construïdes per a una aplicació específica (p. ex., μKosmos)

- episòdiques o enciclopèdiques (p.ex., Cyc [Lenat *et al.*, 1990]),

- lèxiques (p.ex., WN) o conceptuals (p.ex., Cyc).

Finalment, voldríem esmentar l'existència de les anomenades metaontologies, és a dir, formalismes orientats a construir i reutilitzar ontologies preexistents (p.ex., Ontolingua).

Abans d'analitzar les ontologies més comunament utilitzades, és interessant d'esmentar l'existència d'ontologies de nivell superior que intenten representar tot tipus de coses existents al món i les relacions entre aquestes (LE3-4244, 1999). Tot i que les aplicacions per al processament del llenguatge natural (NLP) en diferents àmbits semblen requerir conceptualitzacions lligades a domini, existeix l'esperança que els conceptes independents comuns i les seves relacions siguin comuns a totes les aplicacions i propòsits. En aquest sentit, la càrrega de feina necessària per a cada aplicació individual es reduiria en alguna mesura. Aquest tipus de representacions poden ser útils per a una varietat d'eines centrades en la comprensió i generació del llenguatge natural en general, com ara la desambiguació sintàctica, la resolució de la coreferència, etc. Aquest tipus d'ontologies no són recursos lèxics ja que utilitzen una conceptualització particular de la llengua. Per aquesta raó, és necessari proporcionar algun lligam entre l'ontologia i el lèxic tal com passa amb els mòduls lèxics a ontologies com Cyc, μKosmos i UMLS.

3.3 Ontologies: descripció

En aquest apartat, analitzarem les ontologies més utilitzades amb l'objectiu de determinar quines són les seves característiques essencials en relació amb el seu disseny i la seva estructura general. La nostra anàlisi cobreix les cinc ontologies següents:

- Cyc
- EuroWordNet
- μKosmos
- SIMPLE
- UMLS

És interessant d'esmentar que, tret d'UMLS que és una ontologia específica d'un àmbit temàtic en concret, les altres quatre ontologies es presenten com a estructuracions no lligades a domini tot i que μKosmos està desenvolupada principalment per a l'àmbit de l'economia (i serveix com a suport d'un sistema de traducció automàtica basat en el coneixement).

Per a cada ontologia, incloem una descripció del recurs, una mostra i una anàlisi final[4]. Pel que fa a la mostra, hem seleccionat el concepte de *cèl·lula* "cell", un dels conceptes essencials en la nostra ontologia sobre el genoma humà. Reprenent la voluntat de construir una ontologia en el nostre projecte, val a dir que és interessant de destacar que la majoria d'ontologies no disposen d'una eina de gestió. Tan sols μKosmos presenta aquesta eina, que descriurem breument en l'apartat anterior on es mostra l'estat de l'ontologia sobre el genoma en el marc de l'Institut.

3.3.1 Cyc

Descripció

Cyc és una ontologia d'alt nivell desenvolupada per Cycorp, una empresa privada amb la participació de les empreses informàtiques nord-americanes més importants. El seu desenvolupament va començar a principis de 1980. Durant els últims anys, l'equip que treballa amb Cyc hi ha afegit una gran quantitat de coneixement de base fonamental: fets, regles i heurística sobre els esdeveniments de la vida diària.

Malauradament, només tenim un coneixement general sobre aquest recurs i molt poca informació detallada sobre les especificacions de l'arquitectura del sistema. Recentment, Cycorp ha arribat a gairebé 3.000 termes que intenten representar els conceptes més generals a partir del consens humà. Aquesta ontologia es coneix com a *upper Cyc ontology* i, d'acord amb els responsables, pot ser el nucli d'ontologies posteriors ja que satisfà dos criteris essencials: és una ontologia universal (qualsevol

[4] Cal dir que no disposem del mateix nivell d'informació per a totes les ontologies analitzades. Així, EWN i UMLS són públiques, μKosmos ha estat consultada a través de l'eina de gestió OntoTerm i la seva descripció a Moreno (2000), hem obtingut una mostra de Cyc a http:www.cyc.com i, finalment, per a l'anàlisi de SIMPLE hem seguit la descripció de Bel *et al.* (2000).

concepte imaginable es pot afegir a l'ontologia) i està articulada (les distincions en l'ontologia es fan sobre el criteri de necessarietat i suficiència). De tota manera, no deixa de ser un recurs força petit i, per tant, necessàriament parcial atès que inclou només una petita part de tot el coneixement de base. A més, no presenta un lèxic lligat a l'ontologia ni cap altra eina per al processament del llenguatge natural.

Aquesta ontologia utilitza CycL, un llenguatge de representació del coneixement específic utilitzat per representar els conceptes. Es tracta d'un llenguatge formal la sintaxi del qual deriva del càlcul de predicats de primer ordre (el llenguatge de la lògica formal) i alguns recursos de segon ordre com, per exemple, la quantificació de predicats en alguns casos. El vocabulari de CycL consisteix en termes que poden ser: constants semàntiques, termes complexos, variables, números, cadenes, etc. Els termes es combinen amb expressions significatives de CycL, que acaben formant frases en CycL. Cada terme té, com a mínim, un vincle del tipus "*is-a*" a la seva superclasse corresponent, de la qual és una instància; i un vincle '*genls*' (general) a la superclasse de la qual és una subclasse. Dues de les categories més importants de Cyc són les col·leccions i les relacions (predicats i funcions).

Cada concepte de l'ontologia es representa mitjançant una constant. Una constant pot fer referència a una col·lecció, un objecte individual, una paraula en la llengua natural, un quantificador (per exemple *there exists*), una relació (un predicat, una funció, un valor, un atribut, etc.), entre d'altres.

Tot i que la mateixa descripció de Cyc afirma tenir un sistema de llenguatge natural propi, l'única informació concreta de què disposem és del lèxic en anglès que, en el moment de la nostra recerca, conté 14.000 entrades, que recullen «the usual sorts of linguistic information».

Mostra

En aquesta secció, mostrem la informació disponible per al concepte 'cell', informació que no ens sembla gaire completa:

#$Cell

The collection of living cells; a subset of #$BiologicalLivingObject. Each element of #$Cell is one of the basic structural units of nearly all living things, consisting (at least) of cytoplasm bounded by a cell membrane. Only the living structures viruses, mitochondria, and plastids are not composed of cells.
isa: #$ExistingObjectType

genls: #$BiologicalLivingObject

Figura 3-2. Representació del concepte 'cell' a Cyc.

Anàlisi

La característica més rellevant de l'ontologia Cyc és l'intent de reunir tot els tipus de «common sense knowledge», és a dir, el coneixement comú, i proveir un saber essencial que puguin utilitzar altres programes fent-lo més flexible. Dissortadament, no és un recurs públic i la part que s'ha fet pública és força restringida i, a més, no inclou el mòdul lèxic que hem esmentat. Tanmateix, aquest mòdul lèxic és només per a l'anglès i enlloc no s'indiquen les possibilitats d'extensió cap a d'altres llengües.

3.3.2 EuroWordNet (EWN)

Descripció

EWN (Vossen, 1999) és una base de dades lèxica multilingüe de propòsit general, basada en el recurs WordNet fet a Princeton (Fellbaum, 1998) i que cobreix diverses llengües europees[5]. Cada llengua té la seva pròpia estructura wordnet, és a dir, la seva pròpia estructuració i tots els wordnets estan units per una estructura comuna. Tot i tractar-se d'un projecte en principi finalitzat, es tracta d'un recurs molt viu en diverses aplicacions per al processament del llenguatge natural.

[5] EWN és un projecte finançat per la Unió Europea que, inicialment, cobria l'espanyol, l'holandès, l'anglès i l'italià. Posteriorment, es va estendre al francès, l'alemany, l'estoni i el txec. Algunes extensions, que reben finançament local, són el català, el basc i el grec.

Un wordnet està estructurat en unitats lexicosemàntiques, o synsets, que s'uneixen seguint relacions semàntiques bàsiques (sinonímia, meronímia, hiperonímia i, en alguns casos, algunes relacions morfològiques). Un synset és un conjunt de paraules sinònimes –en el sentit del WordNet de Princetown[6]– que es poden intercanviar en determinats contextos. La sinonímia i la hiponímia són les relacions bàsiques tant per a WN com per a EWN. La primera relació, la sinonímia, s'utilitza per crear synsets i la segona, la hiperonímia, defineix la relació bàsica entre synsets diferents.

Un exemple clar (extret de WN 1.5) el forma el següent grup de paraules {car, auto, automobile, motorcar}. Qualsevol d'aquestes paraules es pot utilitzar en una frase sense canviar-ne el significat bàsic. Aquest synset es relaciona amb d'altres tal com es mostra a la Figura 3-3.

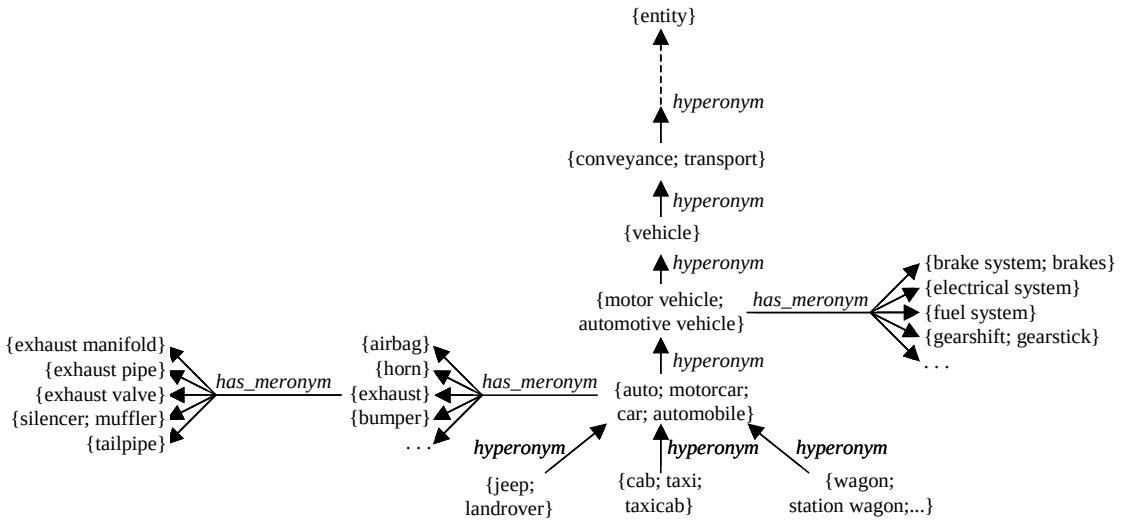


Figura 3-3. Synsets relacionats amb el primer sentit de la paraula 'car'.

Tant EWN com WN estructuren el conjunt de noms en diverses jerarquies, cadascuna de les quals té el seu tot que s'anomena *top*. Cada jerarquia correspon a un camp semàntic autònom i distintiu, com per exemple: {act, activity}, {artifact} i {cognition, knowledge}, així fins a 25. Dins de les jerarquies, cada synset està lligat al seu hiperònim formant una mena de cadena. Una jerarquia lèxica es pot construir seguint les relacions hiperonímiques, per exemple: {bronchus} → {cartilaginous

[6] La definició de sinonímia és: «two expressions are synonymous in a linguistic context C if the substitution of one for the other in C does not alter the truth value» (Miller, G.A. *et al.*, 1993).

tube} → {tube} → {body structure} → {body part} → {part} → {entity}. El símbol "→" s'ha de llegir com a indicador de la relació "*is-a*" o "*is-a-kind-of*", que permet anar des dels ítems específics cap als genèrics. En aquest sentit, un determinat synset hereta tota la informació dels seus hiperònims. Dels exemples esmentats anteriorment, es pot afirmar que un 'bronchus' és un 'body part' i que un 'station wagon' és un tipus de 'car'.

El desenvolupament de les diverses branques del coneixement no és homogeni i, per tant, la mida i la granularitat de les jerarquies és molt diferent. A més, en el projecte WN hi ha 25 nodes superiors o *tops* mentre que s'arriba gairebé al doble en algunes segones versions d'EWN[7].

Els noms formen i constitueixen la jerarquia més desenvolupada tant per a WN com per a EWN. Tanmateix, també hi ha algunes organitzacions per a d'altres categories: els adjectius, els verbs i els adverbis. La taula següent mostra algunes de les relacions definides a EWN que es poden establir entre synsets pertanyents a la mateixa categoria o a categories diferents (vegeu Taula 3-1):

Relation	POS	Example
XPOS_NEAR_SYNONYM	noun – verb verb – adjective	{injection} – {to inject} {to alive} – {alive}
XPOS_NEAR_HYPERONYM	noun – adjective noun – verbs	{age} – {old} {election} – {to vote}
CAUSES	noun – noun	{microorganism} – {health problem}
PERTAINS_TO	adjective – noun	{pulmonary, pneumonic} – {lung}

Taula 3-1. Mostra de noves relacions a EWN.

La descripció que acabem de fer es pot aplicar de manera general a WN i a EWN però hi ha algunes diferències essencials que detallem a continuació. En primer lloc, cal dir que EWN és multilingüe. Aquest multilingüisme s'aconsegueix afegint una relació d'equivalència entre els synsets de cada llengua que participa en el projecte.

[7] EWN per a l'espanyol i el català reflecteixen perfectament l'organització de WN i mantenen el mateix nombre de *tops*.

Aquesta equivalència s'obté per mitjà de l'índex interlingual (ILI), que es pot definir com una llista no estructurada de significats l'objectiu únic de la qual és unir synsets de tots els wordnets existents. Cada synset en cada llengua té al menys un vincle directe o indirecte amb un registre en la llista ILI. Aquesta estructuració permet que cada llengua es pugui organitzar de manera autònoma però, al seu torn, manté la possibilitat d'estar relacionada amb qualsevol altra organització d'altres llengües o subàmbits temàtics.

Existeixen dues estructuracions independents de la llengua en la quan s'hagi creat l'ontologia específica:

- la *Top Concept Ontology*: és una jerarquia de conceptes independents de la llengua que reflecteix les principals distincions entre categories semàntiques, per exemple, substàncies i objectes, objectes naturals i artefactes, etc.

- *una jerarquia amb indicacions de domini*: són unes etiquetes que conformen una estructura de coneixement segons diversos àmbits temàtics, per exemple, el trànsit, els esports, la medicina, etc.

La Figura 3-4 mostra l'arquitectura global de la base de dades d'EWN:

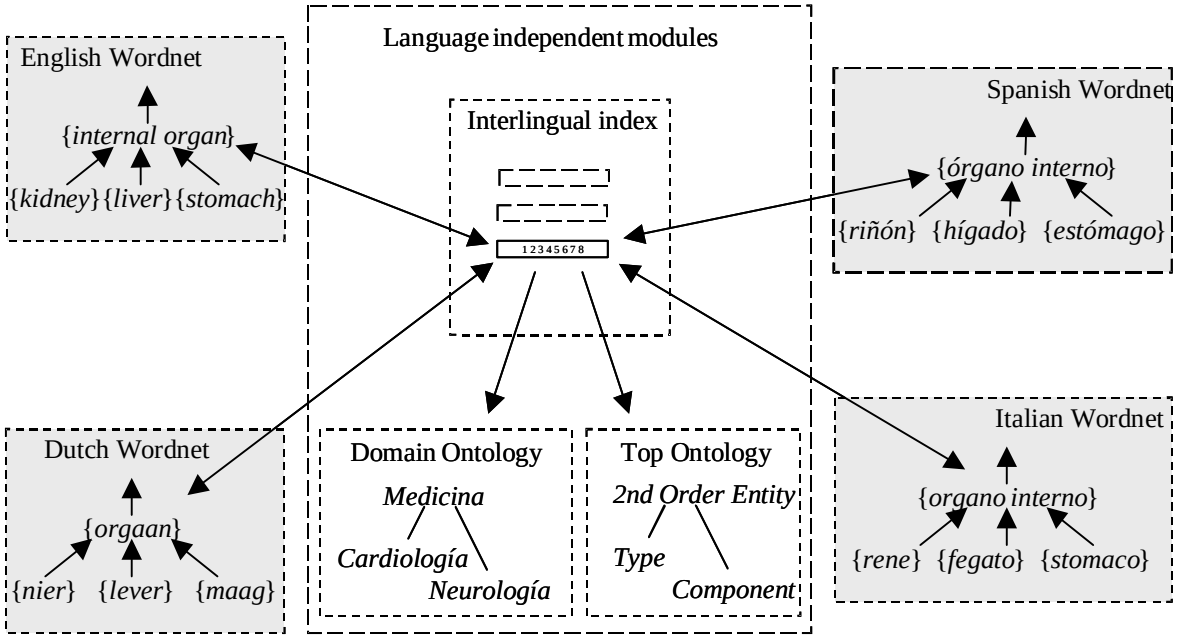


Figura 3-4. Organització general de la base de dades EWN.

Pel que fa als adjectius, tant EWN com WN (seguint Miller *et al.*, 1993) els classifiquen en dues categories majors: descriptius i relacionals. Un adjectiu descriptiu fa referència (o dóna valor a) un atribut del nom al qual modifica. En canvi, els adjectius relacionals no qualifiquen el nom sinó que estableixen una connexió externa amb entitats o àmbits i classifiquen els noms (Soler, 1997). Evidentment, podem trobar adjectius que funcionin com a descriptius i com a relacionals, segons el context, fet que representa un problema de difícil solució. En aquests casos, s'ha optat per donar prioritat als adjectius relacionals.

Per exemple, si tenim la seqüència *asma infantil*, l'adjectiu *infantil* està actuant com una propietat específica d'aquest tipus d'asma: l'edat del pacient, en aquest cas significa que afecta gent jove[8]. Es tracta doncs d'un cas on l'adjectiu *infantil* esdevé un adjectiu descriptiu. Però si considerem el terme *asma bacteriana*, l'adjectiu *bacteriana* informa que el pacient té la malaltia de l'asma produïda per una infecció

[8] La definició que dóna el *Diccionari Enciclopèdic de Medicina* (DEM), 1994 és: "*asma infantil: La que, en forma d'accessos, es presenta en els infants. Les crisis no són tan ben delimitades com a l'edat adulta, van sovint acompanyades de febre, tenen tendència a la insuflació pulmonar i s'associen a la presentació de raneres humides. Sempre hi ha una predisposició constitucional;*

bacterial, informació que relaciona la malaltia al seu origen. Es tracta, doncs, d'un adjectiu relacional.

Els adjectius descriptius són representats a EWN (i a WN) com una estructura bipolar en què cada element de l'estructura té un nucli i alguns synsets satèl·lits. Fixem-nos en la propietat *edat* ('age'), els valors que pot adoptar es poden considerar com dos valors principals i oposats: *jove* ('young') i *vell* ('old'). Ambdós valors es poden considerar com el nucli de cada cara de l'estructura. Aquesta relació d'antonímia es codifica com a 'near_antonym'. I d'altres adjectius relacionats amb el nuclis es vinculen amb la relació 'near_synonym'. En altres paraules, la llista gradual dels valors associats a la propietat està estructurada per mitjà de vincles apropiats. La relació "xpos_near_hyperonym"[9] s'utilitza per unir adjectius nuclears a la seva corresponent propietat. L'estructura resultants es representa a la Figura 3-5.

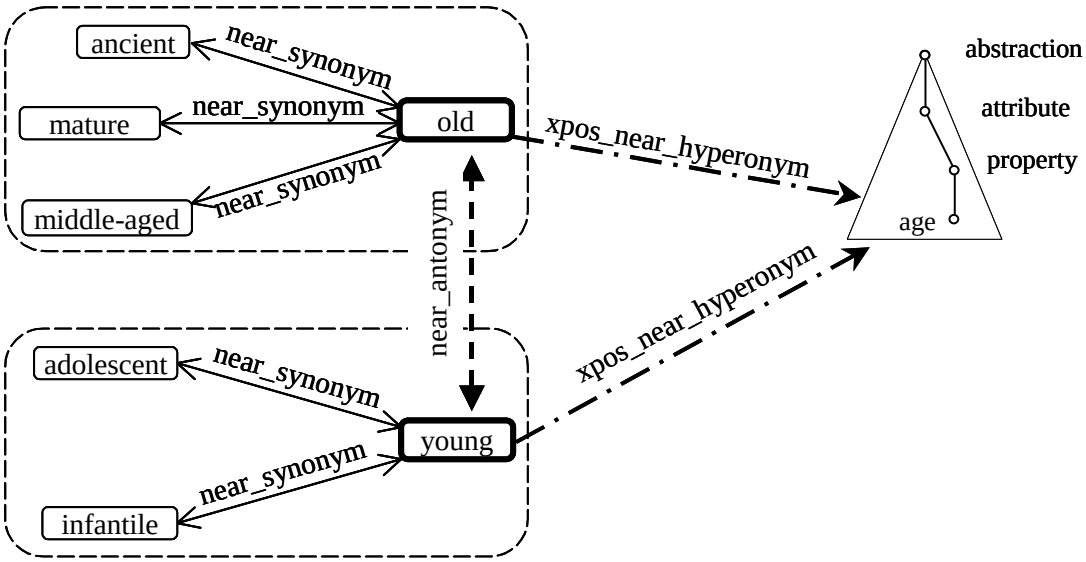


Figura 3-5. Adjectius descriptius: representació a EWN.

L'organització dels adjectius relacionals és més senzilla que la des descriptius. Només cal relacionar l'adjectiu amb el seu nom corresponent. Per exemple, l'adjectiu

diversos factors poden estar implicats en la seva aparició: rinofaringitis, causes psíquiques, al·lèrgens, etc. En general l'asma infantil desapareix en arribar l'infant a la pubertad"

[9] La relació "xpos_near_hyperonym" pot ser parafrasejada com a: X (adjectiu) és un valor atributiu per a la propietat Y (nom).

bronquial s'ha de relacionar amb el nom *bronqui*. La Figura 3-6 mostra gràficament el tractament d'aquest tipus d'adjectius.

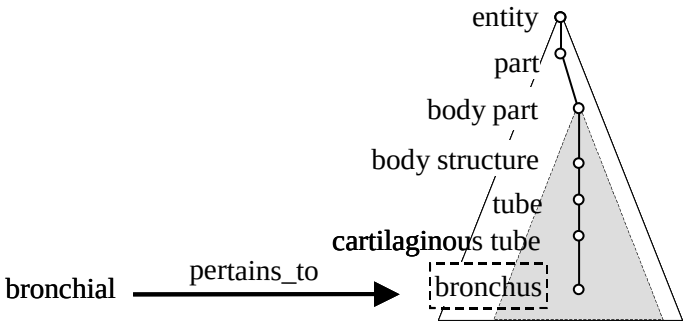


Figura 3-6. Adjectius relacionals: representació a EWN.

Finalment, en aquest apartat de descripció cal fer notar que no tot el sistema ha estat definit i implementat al mateix nivell. Així, les jerarquies nominals i verbals estan força més desenvolupades que les dels adjectius i verbs[10]. I també cal esmentar que s'han definit molts vincles, moltes relacions, però que aquestes no s'han aplicat efectivament al conjunt de tot el recurs.

Mostra

La base de dades lèxica EWN proporciona tres entrades per al concepte 'cell'. La Figura 3-7 només mostra la informació relativa al sentit pertinent per a l'àmbit del genoma.

[10] Els esforços fets per a dur a terme l'aplicació de cada POS depenen en gran mesura de l'organització responsable de cada llengua.

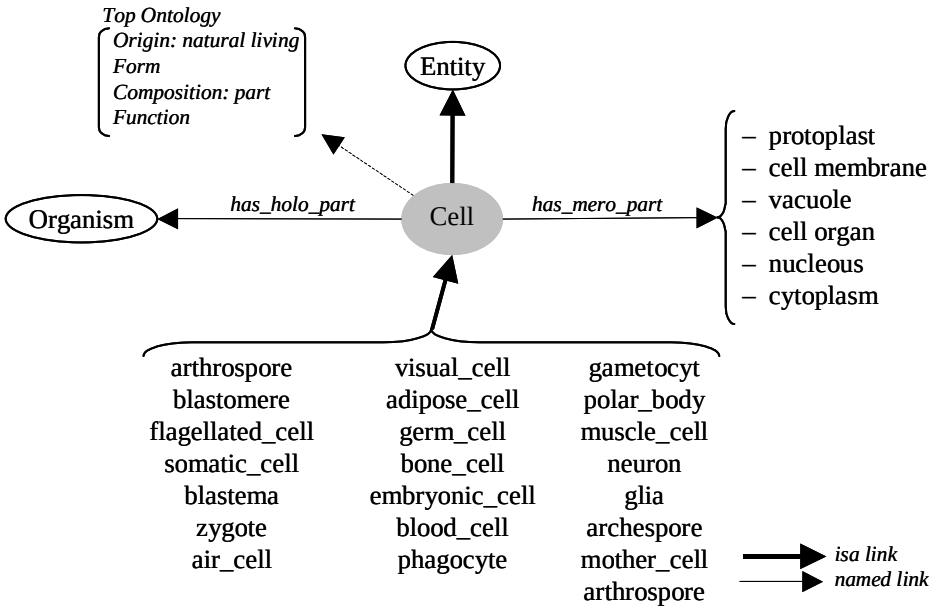


Figura 3-7. Representació del concepte 'cell' a EWN.

Anàlisi

EWN, i el seu antecessor WN, és una ontologia basada en el lèxic. Per tant, les paraules són la base de l'organització conceptual. Per aquest motiu, sovint no es considera com una ontologia real. Malgrat això, amb freqüència és utilitzat en un gran nombre de projectes principalment perquè es tracta d'un recurs realment públic que té una cobertura força gran en les llengües europees.

Un dels aspectes més criticats de WN (i en conseqüència d'EWN) és la seva granularitat asimètrica. Això significa que diferents àrees de l'ontologia, que reflecteixen àmbits diferents, estan desenvolupades de manera força desigual.

Finalment, cal esmentar també el fet que, atès que es tracta d'un recurs de coneixement general, algunes informacions es troben, a primer cop d'ull força barrejades. Això és així perquè un mateix synset inclou variants especialitzades i no especialitzades. Per exemple, el synset que inclou la variant *hematoma* (en principi una unitat especialitzada) també inclou les variants *blau* o *morat* que com es pot observar s'utilitzen en la llengua en un altre nivell d'especialització.

3.3.3 μKosmos

Descripció

El projecte μKosmos, conjuntament desenvolupat per investigadors de la New Mexico State University, la Carnegie Mellon University i diverses agències governamentals nord-americanes, presenta un estudi teòric de la varietat de microteories sobre lingüística computacional que han esdevingut centrals per donar suport a un sistema de traducció automàtica que utilitza una base de coneixements.

L'objectiu final és definir una metodologia útil per representar el significat del textos en llenguatge natural en un format interlingual que sigui neutral pel que fa a la llengua, una mena d'interlingua que anomenen *Text Meaning Representation* (TMR). El TMR és el resultat de l'anàlisi d'alguns textos en algunes de les llengües involucrades en el projecte per al procés de traducció automàtica. El resultat serveix com a *input* per al posterior procés de generació. El significat dels textos es deriva de l'anàlisi de la seva informació lèxica, sintàctica, semàntica i pragmàtica. Aquesta informació es representada en el TMR com a elements que s'han de veure com un model motivat independentment del món, és a dir, l'ontologia. El lèxic proporciona el lligam entre l'ontologia i el TMR perquè el significat dels ítems lèxics es defineixen i s'estableixen segons el seu mapeig amb els conceptes de l'ontologia i la relació amb l'estructura que presenta el TMR.

Pel que fa a les llengües, el projecte treballa amb documents sobre fusions d'empreses en anglès i japonès, tot i que afirmen que l'anàlisi també s'ha estès a l'espanyol i al francès utilitzant textos similars.

En el projecte μKosmos, una ontologia es defineix com a:

> «a computational entity and a resource containing knowledge about what "concepts" exist in the world and how they relate to one another. A concept is a primitive symbol for meaning representation with well defined attributes and relationships with other concepts. An ontology is a network of such concepts forming a symbol system where there are no

> uninterpreted symbols (except for numbers and a small number of known literals)[11].»

Des d'una aproximació a la traducció automàtica basada en el coneixement, la representació interlingual del significat s'extreu de les representacions dels significats de les paraules dels lexicons computacionals i pel coneixement capturat a l'ontologia. L'ontologia esdevé un recurs de coneixement independent de la llengua, i conté el conjunt de símbols i les possibles relacions entre aquests símbols.

Així, una ontologia creada per a propòsits del processament del llenguatge natural esdevé un conjunt de coneixements sobre el món (o sobre un àmbit especialitzat) que:

— agrupa símbols primitius per a la representació del coneixement;

— organitza aquests símbols de manera jeràrquica;

— interconnecta aquests símbols per mitjà de relacions semàntiques.

En el disseny de μKosmos, l'ontologia conté informació sobre:

— les categories (o conceptes) que existeixen en el món o en un àmbit determinat;

— les propietats d'aquestes categories;

— com es relacionen les unes amb les altres.

En un projecte orientat a la traducció automàtica, és important utilitzar l'ontologia per a:

— proporcionar les bases per al TMR;

— permetre que els lexicons de diferents llengües comparteixin el mateix coneixement;

— permetre que els analitzadors de la llengua de partida i els generadors en la llengua d'arribada comparteixin el mateix coneixement específic.

[11] Per a una descripció força més detallada dels criteris de disseny de μKosmos vegeu: http://crl.nmsu.edu/Research/Projects/mikro/ de Kevi Mahesh, 1996.

És interessant d'esmentar que en cada llengua, el lèxic està organitzat en superentrades que s'identifiquen utilitzant la forma de cada entrada de diccionari. Dins d'una superentrada, els lexemes individuals es representen en cada llengua mitjançant marcs o *frames*. Cada entrada lèxica disposa de diferents nivells d'informació, com ara:

- CAT (categoria sintàctica)
- ORTH (ortografia, abreviatures)
- PHON (fonologia)
- MOPRH (formes irregulars i informació de classe)
- SYN (característiques sintàctiques)
- SYN-STRUC (informació de dependència i subcategorització)
- SEM (semàntica lèxica i representació semàntica)
- LEXICAL-RELATIONS (col·locacions)
- PRAGM (informació pragmàtica)
- ANNOTATIONS (informació de l'usuari o lexicogràfica, exemples, etc.).

L'ontologia desenvolupada en el marc del projecte μKosmos arriba, en alguns casos, a 10 subnivells. El nivell superior o *top-level* (*all*) té tres primers nivells, *object*, *event*, i *property*. Aquest últim nivell inclou els subnivells *attribute* i *relation* tal com mostra la Figura 3-8.

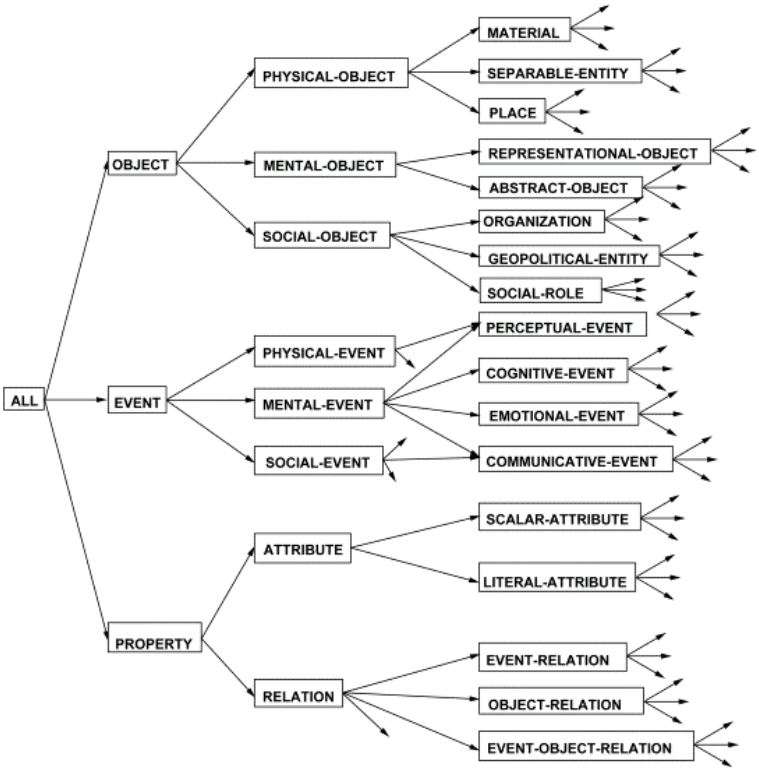


Figura 3-8. Nivells inicials de la jerarquia de μKosmos.

Tal com acabem d'esmentar, l'arquitectura per al processament del llenguatge natural de μKosmos conté quatre mòduls diferents orientats a l'anàlisi textual:

- el lèxic
- l'ontologia
- les representacions interlinguals
- les microteories

que s'interrelacionen tal com mostra la Figura 3-9.

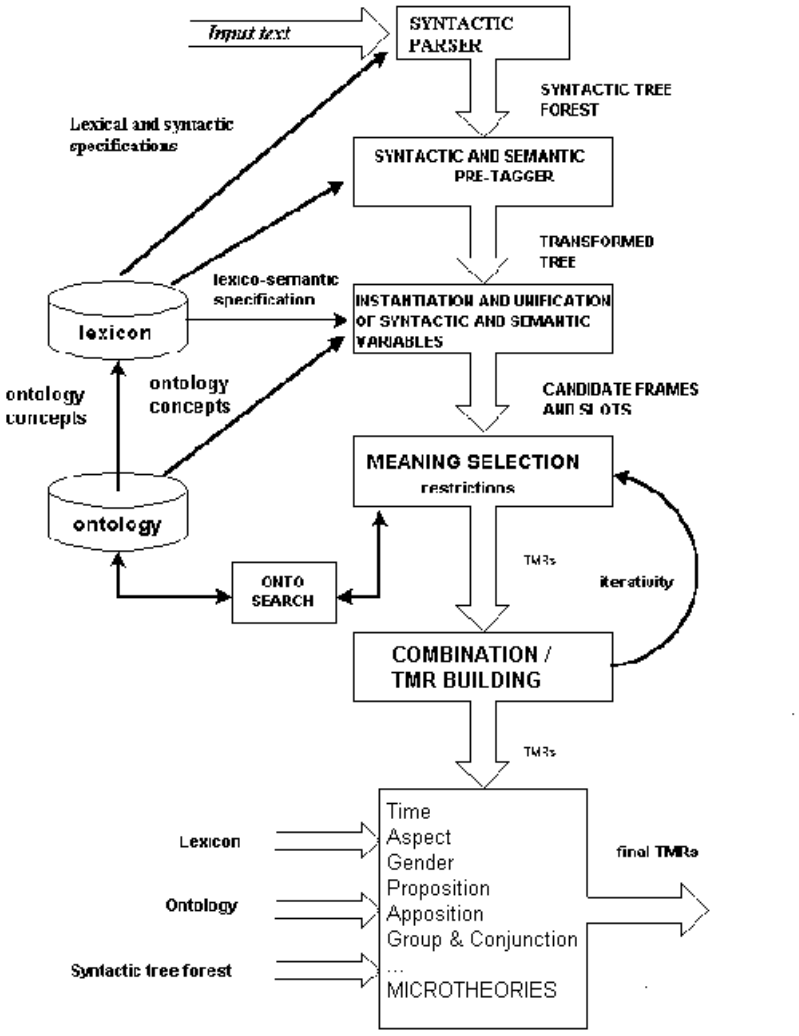


Figura 3-9. Arquitectura per al processament del llenguatge natural a μKosmos.

Mostra

A la Figura 3-10, mostrem la informació extreta de l'ontologia. Cal fer notar que aquesta figura mostra només els atributs, les relacions locals i les relacions adquirides per herència del concepte 'cell'[12].

[12] Aquesta informació ha estat obtinguda mitjançant OntoTerm, un sistema de gestió de terminologia basat en una ontologia creat per A. Moreno (Universidad de Málaga). Aquesta eina implementa l'ontologia de μKosmos mitjançant l'ús d'una base de dades relacional que es pot utilitzar en un ordinador personal. Per tant, no sabem si aquesta informació proporcionada per OntoTerm recull totes

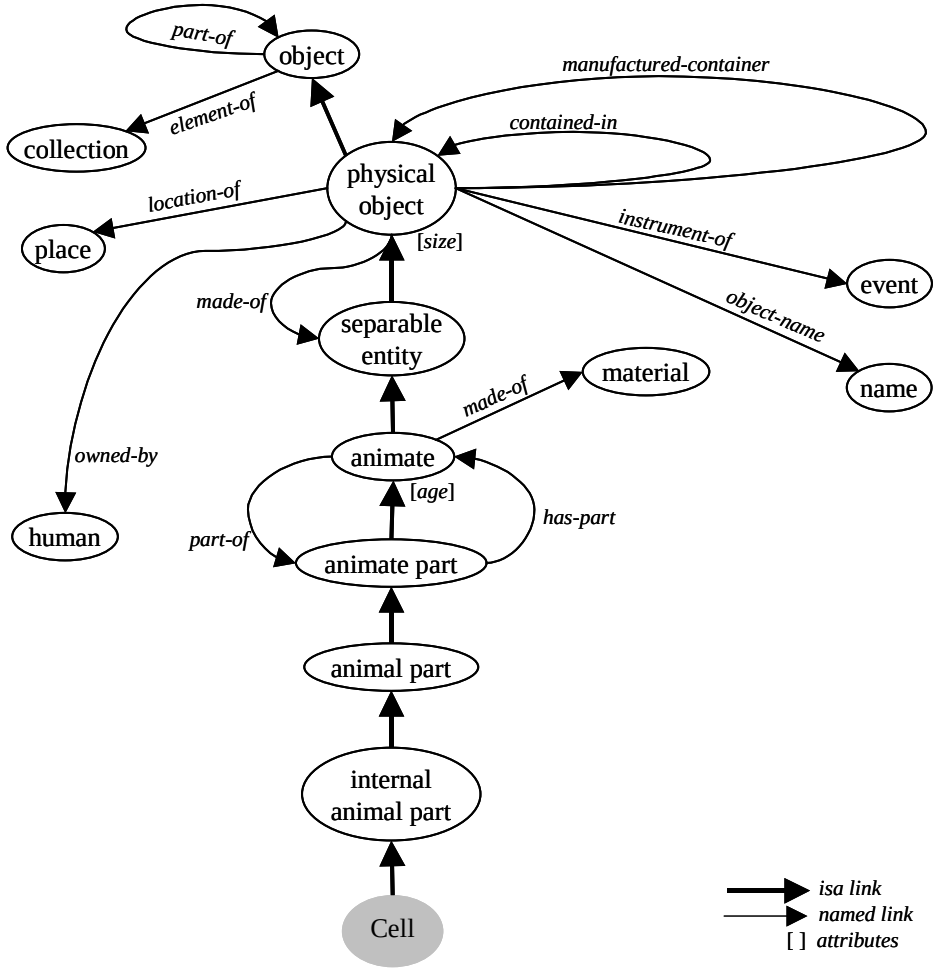


Figura 3-10. Representació del concepte 'cell' a μKosmos.

Anàlisi

Malgrat que es presenta com una ontologia general, no sabem quina és la seva cobertura pel que fa als àmbits que difereixen del seu interès de cara al sistema de traducció automàtica que ja hem esmentat que treballa, principalment, amb textos sobre fusions d'empreses.

Un dels seus principals avantatges és l'existència d'un recurs de suport, una eina de gestió de l'ontologia: OntoTerm®. Les característiques d'aquesta eina faciliten totes

---

les característiques originals del disseny de μKosmos. Tanmateix, no disposem d'informació directament de μKosmos, tot i la demanda feta directament per a una consulta amb propòsits

les operacions de maneig de l'ontologia. Cal destacar, també, que aquesta eina deixa que l'usuari afegeixi qualsevol tipus d'informació conceptual, com per exemple qualsevol classe de relació conceptual (no hi ha una llista prèviament tancada proporcionada pel sistema sinó que aquesta depèn de l'ontologia). Per tant, és necessari de fer-ne un ús molt sistematic i rigorós per evitar inconsistències i duplicacions, entre d'altres.

3.3.4 SIMPLE

Descripció

El projecte SIMPLE (*Semantic Information for Multifunctional Plurilingual Lexica*), que es pot considerar la continuació del projecte europeu PAROLE[13], intenta proporcionar un conjunt de recursos lingüístics per a les llengües de la Unió Europea. Cobreix dotze llengües diferents. Els investigadors treballen en la construcció d'un lèxic comú reutilitzable, que contingui informació morfològica, sintàctica i semàntica que ha de permetre que altres aplicacions es puguin desenvolupar amb un cost temporal i econòmic inferior.

L'objectiu principal és obtenir una base semàntica per als nivells morfològics i sintàctics que es van crear en el projecte PAROLE. Els lexicons semàntics es basen en corpus de llengua general i es vol arribar a cobrir fins a 10.000 significats (7.000 per als noms, 2.000 per als verbs i 1.000 per als adjectius). El format d'intercanvi per al lèxic és SGML i pel que fa a la informació morfològica i sintàctica, tots els lexicons semàntics comparteixen la mateixa DTD.

---

acadèmics.

[13] PAROLE és el primer projecte finançat per la Comunitat Europea. Aquest projecte tenia com a objectiu produir recursos de gran abast, genèrics i reutilitzables per a la llengua escrita en totes les llengües de la Unió Europea. Aquests recursos es van dissenyar per a cadascuna de les llengües seguint els mateixos principis, les mateixes especificacions lingüístiques i el mateix format de representació. Més concretament, es va crear un corpus per a 14 llengües i 12 lèxics en suport electrònic. Per a una descripció més detallada dels resultats d'aquest projecte vegeu: http://www.ub.es/gilcub/SIMPLE/simple.html.

Fins al moment actual, el projecte s'ha orientat a definir una arquitectura general que permeti codificar el significat lèxic. Per aquest motiu, els investigadors han treballat força en el desenvolupament d'una ontologia general (*top ontology*) que contingui els tipus semàntics i un conjunt d'eines d'anàlisi aplicables als ítems lèxics. Actualment, la majoria d'esforços s'adrecen a assegurar la consistència i facilitar la codificació del lèxic. En el marc del projecte, l'ontologia esdevé una eina semàntica utilitzada per representar els sentits principals d'una paraula i relacionar-los amb cada una de les entrades lèxiques per mitjà d'una plantilla.

Així, i tal com es diu en el document *SIMPLE Linguistic Specifications Paper*,

> «each word sense corresponds to a given semantic type. Each semantic type is actually a cluster of structured information. Semantic types differ in terms of how much information they convey. In other words, word senses differ in their degree of complexity, which is explicitly part of their semantic type.»
>
> (Lenci, A., *et al.*: 1999)

Per tal d'unificar la informació per a les 12 llengües involucrades en el projecte[14], SIMPLE treballa amb un patró d'informació que conté les dades següents:

— Tipus semàntic

— Plantilla

— SemU.

Hem de destacar que l'objectiu principal de SIMPLE no és la construcció d'una ontologia. Al contrari, és un intent de codificar la informació semàntica del lèxic d'un gran nombre de llengües. L'especificació formal per a la representació i codificació d'informació segueix aquest patró:

— Tipus semàntic

[14] Les llengües que participen en aquest projecte són: alemany, anglès, català, danès, espanyol, holandès, finès, francès, grec, italià, portuguès i suec.

- Informació d'àmbit temàtic
- Glossa
- Estructura argumental
- Restricció de selecció d'arguments
- Tipus eventiu, per als verbs
- Lligam dels arguments amb l'esquema de subcategorització sintàctica
- Informació sobre el tipus jeràrquic
- Informació de qualia
- Informació sobre l'alteració polisèmica regular
- Informació sobre relacions de *part-of-speech*
- Col·locacions extretes de corpus
- Relacions de sinonímia

El contingut d'aquest patró segueix les recomanacions del grup de treball EAGLES (EAGLES Lexicon/Semantics Working Group) i la seva organització es basa en extensions de la Teoria del Lèxic Generatiu (Pustejovsky, 1995).

L'ontologia de SIMPLE es divideix en una *core ontology* (formada pels tipus que s'han identificat com a central i comuns per a la construcció de diferents lexicons) i la *recommended ontology* (cosntituïda per tipus més específics que són nodes inferiors en la jerarquia). L'ontologia s'utilitza per proveir el cos conceptual que comparteixen tots els lexicons. És la plantilla (*template*) la que proporciona la interfície entre l'ontologia i el lexicó. La següent graella mostra una representació esquemàtica d'una plantilla, que inclou la posició de la SemU (unitat semàntica) en l'ontologia:

SemU:	Identifier of a SemU
SynU:	Identifier of the SynU to which the SemU is linked
BC Number:	Number of the corresponding Base Concept in EuroWordNet
Template_Type:	Semantic type of the SemU
Template_Supertype:	Semantic type which dominates the Template_Type of the SemU in the type-hierarchy
Unification_path:	Unification history of a template (for unified top-types)
Domain:	Domain information from LexiQuest domain list
Semantic Class:	One of the classes provided by LexiQuest
Gloss:	Lexicographic definition
Event Type:	Event sort (for event SemUs only)
Predicative Representation:	Predicate associated with the SemU, and its argument structure
Selectional Restr.:	Selectional restrictions on the arguments
Derivation:	Derivational relations between SemUs
Formal:	Formal relation between SemUs
Agentive:	Agentive relations between SemUs
Constitutive:	Constitutive relations between SemUs Constitutive semantic features
Telic:	Telic relations between SemUs
Synonymy:	Synonyms of the SemU
Collocates:	Collocate information
Complex:	Polysemous class of the SemU

Després d'haver observat totes les especificacions lingüístiques de SIMPLE, hem d'emfasitzar els esforços que s'han dut a terme per codificar la informació semàntica, encara que l'ontologia no estigui prou desenvolupada per poder ser utilitzada en d'altres aplicacions per al processament del llenguatge natural.

Mostra

SIMPLE, igual que EWN, dóna tres entrades per a "cell", cadascuna de les quals segueix una plantilla, tal com es mostra seguidament:

```
<SemU id="celula1_Organicobject"
example="e.g., célula (Unidad fundamental de los seres vivos, con
cierta autonomía)"
naming="célula"
weightvalsemfeaturel="WVSFTemplateOrganicobjectPROT
WVSFTemplateSuperTypeConcreteentityPROT
TSVP_OBJECT_TS_classificateur_de_nom_C TSVP_PLUS_TS_PART_T">
<RWeightValSemU semr="SRIsapartof" target="cuerpo_Organicobject"
weight="ESSENTIAL">

<SemU id="celula1_Instrument"
example="e.g., se ha disparado la célula fotoeléctrica (dispositivo
que transforma las variaciones de intensidad luminosa en variaciones
de de intensidad de una corriete)"
naming="célula"
weightvalsemfeaturel="WVSFTemplateInstrumentPROT
WVSFUnificationPathConcreteentity-ArtifactAgentive-TelicPROT
TSVP_APPARATUS_TS_classificateur_de_nom_C">
<RWeightValSemU semr="SRCreatedby" target="hacer_X"
weight="PROTOTYPICAL">
<RWeightValSemU semr="SRUsedfor" target="hacer_X"
weight="PROTOTYPICAL">

<SemU id="celula1_HumanGroup"
example="e.g., la célula del partido (Unidad o grupo separado de una
organización)"
naming="célula"
weightvalsemfeaturel="WVSFTemplateHumanGroupPROT
TSVP_PLUS_TS_HUMAN_T TSVP_PLUS_TS_COLLECTIVE_T
WVSFTemplateSuperTypeGroupPROT
TSVP_GROUP_NAMES_TS_classificateur_de_nom_C">
<RWeightValSemU semr="SRHasasmember" target="persona_Human"
weight="PROTOTYPICAL">
```

Figura 3-11. Representació del concepte ‘cell’ a SIMPLE.

Anàlisi

La informació anterior mostra com s’organitza la informació en una entrada lèxica a SIMPLE. Com podem veure, la informació lingüística s’estructura en termes de SemU, que contenen la identificació de l’entrada lèxica. Cada entrada lèxica inclou un exemple d’ús i altres entrades relacionades per valors semàntics. De fet, l’ontologia només s’utilitza com una manera de controlar la informació inclosa en la plantilla però no de manera explícita per a organitzar el lexicó.

Més concretament, les SemU inclouen atributs i relacions (a part de les predicatives). Els atributs poden contenir informació sobre l’ontologia i qualsevol altra informació rellevant (per exemple, ‘Connotation’, ‘Plus-Edible’, ‘Plus-Fictive’, ‘Plus-Human’, ‘Plus-Gradable’, etc.). Les relacions tenen com a objectiu descriure les SemU en

termes d'estructura de qualia ('tellic', 'agentive', etc.). Finalment, la codificació semàntica inclou especificació enciclopèdica i lingüística amb l'objectiu d'explicitar l'ús lingüístic de la unitat. Per tant, en el cas dels verbs, i per a tots els predicats, s'intenta reflectir l'estructura argumental i les restriccions de selecció que se li poden aplicar. Tota aquesta informació descriu la semàntica lèxica d'una unitat però no la seva posició en l'ontologia.

3.3.5 UMLS

Descripció

D'entre els àmbits temàtics que són objecte de recursos específics, cal esmentar la medicina. La majoria de recursos s'han desenvolupat per a l'anglès. L'únic recurs que inclou algun tipus d'informació per a l'espanyol és el projecte UMLS (*Unified Medical Language System*). Es tracta d'una recerca i d'un desenvolupament a llarg termini que té com a objectiu facilitar la recuperació i la integració de fonts d'informació de caràcter biomèdic que siguin llegibles per a una màquina (UMLS, 2001). Aquest projecte a llarg termini es va iniciar l'any 1986 i va rebre el suport de la NLM (*National Library of Medicine*)[15].

Els recursos d'UMLS s'utilitzen actualment en diferents aplicacions[16] (principalment relacionades amb la recuperació d'informació i la integració d'eines diferents) o en activitats de recerca com són els sistemes d'extracció de termes descrits a Maynard (1999).

Actualment, els recursos d'UMLS representen parcialment la informació relacionada amb el Projecte Genoma Humà. De tota manera, diversos investigadors (Yu *et al.*, 1999) proposen diferents tipus de modificacions o ampliacions per adaptar aquest recurs a aquest projecte mundial.

---

[15] Més informació sobre aquest projecte es troba disponible a [NLM, 1998], o bé en la següent adreça electrònica: http://www.nlm.nih.gov/research/umls/umlsmain.html.

[16] A la pàgina web http://www.nlm.nih.gov/research/umls/umlsapps.html és possible de consultar una llista d'aplicacions i serveis que utilitzen aquest recurs.

UMLS és gratuït als Estats Units i a usuaris internacionals que ho demanin, tot i que es requereixen acords addicionals amb els productors dels vocabularis que conté l'ontologia.

UMLS és un conjunt de tres grans fonts d'informació: a) UMLS Metathesaurus, b) UMLS Semantic Network, i c) SPECIALIST Lexicon. Els apartats següents descriuen cadascuna de les tres fonts esmentades.

a) UMLS Metathesaurus

El Metathesaurus conté informació sobre conceptes biomèdics i termes obtinguts d'un conjunt de vocabularis i sistemes de classificació controlats. Conté els noms, els significats, els contextos jeràrquics, els atributs, les relacions entre termes presents en els vocabularis font, i afegeix una determinada informació bàsica a cada concepte. A més, estableix noves relacions entre termes provinents de diferents vocabularis. El seu abast es veu determinat per l'abast dels vocabularis que fan servir com a font. La seva estructura permet afegir llengües que no siguin l'anglès, això és, algunes llengües europees. El Metathesaurus d'UMLS incloïa l'any 2001:

— 800.000 conceptes

— 1.400.000 "termes" (Eye, Eyes, eye=1)

— 1.900.000 cadenes (o noms de conceptes) (Eye, Eyes, eye = 3)Més de 50 vocabularis font.Per gestionar aquestes dades, la distribució d'UMLS inclou MetamorphoSys, que és una eina que permet la creació d'una versió adaptada segons les necessitats del Metathesaurus[17]. En aquest sentit, els usuaris poden excloure els vocabularis que no necessitin, alterar l'ordre de les unitats, excloure vocabularis que necessitin llicència d'ús, etc.

[17] MetamorphoSys té un requisits informàtics molt elevats: un mínims de 256 MB de memòria física i 8 GB recomanats d'espai en disc. Funciona sobre Unix, Linux i Windows.

b) UMLS Semantic Network

El Semantic Network (xarxa semàntica) proporciona una categorització consistent de tots els conceptes representats al Metathesaurus. A la versió de l'any 2001, hi ha 134 tipus semàntics i 54 relacions entre aquests. Els tipus semàntics són els nodes de la xarxa, i les relacions entre aquests són els vincles. Cada node té una etiqueta denominativa simple. Existeixen grans grups de tipus semàntics per als organismes, estructures anatòmiques, funcions biològiques, productes químics, esdeveniments, objectes físics, i nocions o idees. Tots els tipus semàntics es divideixen en entitats o esdeveniments. Cada concepte del Metathesaurus s'associa amb un o més tipus semàntics.

El vincle essencial és la relació "*is-a*", la qual estableix la jerarquia de tipus dins la xarxa i s'utilitza per decidir quin és el tipus semàntic específic que està disponible per assignar-lo a un concepte al Metathesaurus (relació d'hiperonímia). A més, s'ha utilitzat un conjunt de relacions no jeràrquiques entre els tipus semàntics, que estan agrupades en cinc categories majors que, al seu torn, també són relacions: 'physically related to', 'spatially_related_to', 'temporally_related_to', 'functionally_related_to' i 'conceptually_related_to.'

Les relacions s'estableixen entre tipus semàntics i no necessàriament afecten totes les instàncies dels conceptes que s'han assignat a aquell determinat tipus semàntic. Això és, la relació pot o pot no aparèixer entre una parella determinada de conceptes. Per exemple, tot i que la relació 'evaluation_of' s'estableix entre els tipus semàntics 'Sign' i 'Organism Attribute', un determinat signe o un atribut en particular pot no estar lligat per aquesta relació. Així, signes com "overweight" i "fever" són avaluacions dels atributs d'organismes "body weight" i "body temperature", respectivament.

L'organització de la xarxa permet que un tipus semàntic rebi informació dels seus antecessors utilitzant un mecanisme d'herència. En alguns casos, hi haurà un conflicte entre la localització dels tipus i el vincle que cal heretar. En aquests casos, el mecanisme d'herència es bloqueja.

c) SPECIALIST Lexicon

El SPECIALIST lexicon (lexicó especialitzat) és un lexicó anglès amb molts termes biomèdics. S'ha dissenyat per ser utilitzat per un sistema de processament del llenguatge natural. La versió de l'any 2000 d'UMLS inclou uns 108.000 registres lèxics.

L'entrada al lexicó per a cada paraula o terme recull informació sintàctica, morfològica i ortogràfica. Les entrades que comparteixen la seva forma de base i algunes variants s'entren com un sol registre lèxic[18].

La informació lèxica inclou la categoria sintàctica, la variació flexiva (p. ex., singular i plural per als noms, conjugacions verbals, els graus positiu, comparatiu i superlatiu dels adjectius i adverbis), i patrons de complementació (per exemple, els objectes i altres arguments que poden seleccionar els verbs, els noms i els adjectius). El lexicó reconeix onze categories: verbs, noms, adjectius, adverbis, auxiliars, modals, pronoms, preposicions, conjuncions, complementadors i determinants.

Els patrons oracionals bàsics d'una llengua estan determinats pel nombre i la naturalesa dels complements seleccionats pels verbs. El lexicó d'UMLS reconeix cinc grans patrons de complementació: intransitiu, transitiu, bitransitius, d'unió i transitiu complex. Les entrades verbals també codifiquen cada una de les formes flexives (parts principals del verb). Els verbs es classifiquen segons la seva flexió en unitats regulars, regulars grecollatines o irregulars. Les entrades nominals descriuen la flexió dels noms i les variacions gràfiques. Els patrons de complementació per als noms i la informació sobre nominalització també s'inclouen quan es consideren rellevants. A més, els adjectius porten incorporat un codi de posició per indicar la posició sintàctica en què poden aparèixer. Els adjectius poden ser qualitatius, classificadors o adjectius indicadors de color. Els adverbis es codifiquen segons les seves propietats modificadores. El lexicó reconeix adverbis oracionals, verbals i

[18] La forma bàsica és la no flexionada; el singular en el cas dels noms, l'infinitiu per als verbs i la forma positiva per als adjectius i adverbis.

intensificadors, i classifica els dos últims tipus en adverbis de manera, temporals i locatius.

Com hem esmentat anteriorment, tota la informació lingüística rellevant està disponible per a l'anglès però no per a les altres llengües.

Mostra

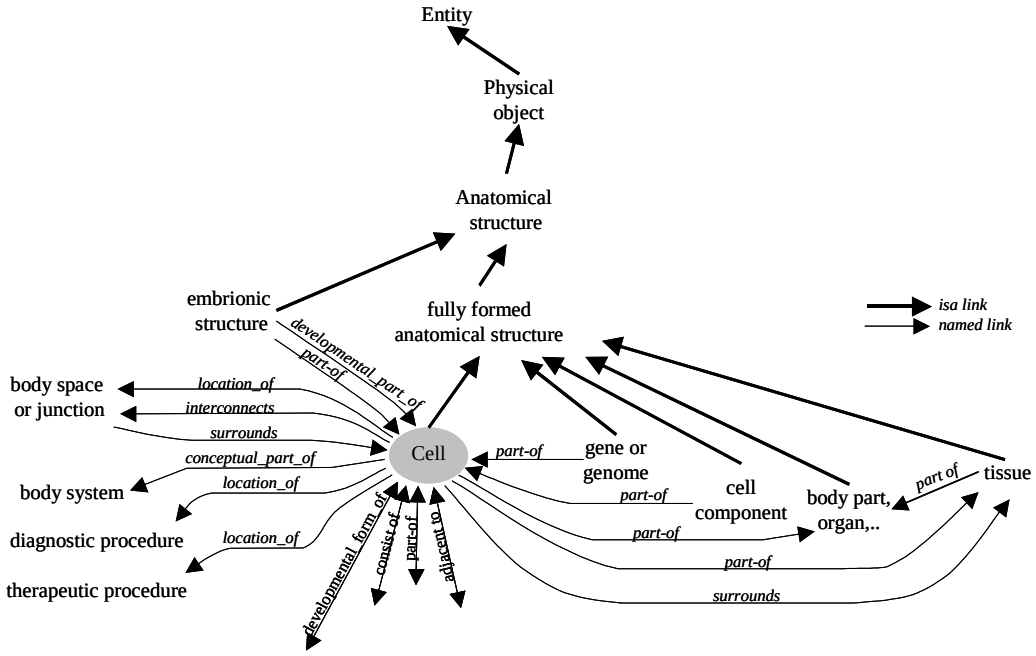


Figura 3-12. Representació del concepte 'cell' en l'UMLS Metathesaurus.

Anàlisi

D'una banda, l'UMLS Semantic Network forneix una organització molt detallada dels conceptes mèdics que s'ha comprovat que es pot extrapolar a altres subdominis mèdics de recent aparició[19].

[19] Vegeu, per exemple, la proposta de Yu *et al.* (1999). Tanmateix, sembla una proposta merament teòrica atès que els autors no proporcionen resultats del seu ús.

D'altra banda, tot i que l'estructura del Metathesaurus permet incloure llengües diferents de l'anglès, sorgeixen algunes limitacions quan s'utilitzen aquests tipus de recursos (com a mínim per al cas de l'espanyol en l'edició de 1998):

- La cobertura per aquestes llengües és força limitada[20].
- Les entrades per a l'espanyol no preveuen els caràcters del tipus á, É, ñ... (que es redueixen a a, E, n);
- Les paraules que participen d'un determinat concepte no estan lematitzades[21].

Com a conclusió final, volem indicar que la solució a alguns dels problemes esmentats és possible tot i que requeriria força despesa de temps i treball. De tota manera, l'ontologia té una gran validesa per a la consulta atès que es tracta d'un recurs dedicat a l'àmbit de la medicina.

Pel que fa als conceptes que inclou, alguns són força dubtosos o, com a mínim, curiosos. Fixem-nos, per concloure, en els exemples següents que són conceptes relacionats a UMLS: 'deportes', 'maltrato conyugal', "distensiones y esguinces', etc.

3.4 Ontologies: anàlisi comparativa

En aquesta secció, i després d'haver revisat les principals característiques de les cinc ontologies seleccionades, analitzarem alguns dels paràmetres clau que cal tenir en compte per avaluar una ontologia. Cal esmentar, abans però, que atès que es tracta d'ontologies força diferents, una comparació directa és una tasca força difícil. Com ja hem vist, les mostres de l'apartat anterior ens ensenyen les moltes diferències pel que fa al disseny i propòsit de cada ontologia.

[20] En l'edició de 1998, l'única eina per a l'espanyol incloïa gairebé 24 K conceptes (el nombre total de conceptes en aquell moment era d'uns 478 K en més de 40 recursos diferents). Versions posteriors incrementen aquesta xifra però els problemes de base persisteixen.

[21] Observem l'exemple *autoantígenos* enlloc d'*autoantígeno* o *complicaciones postoperatorias* enlloc de *complicación postoperatoria*. Aquest problema no es troba en els vocabularis per a l'anglès, on existeix un índex d'unitats lematitzades.

No obstant aquesta dificultat, algunes característiques són efectivament comparables. En aquest sentit, els elements revisats en l'anàlisi comparativa són: la disponibilitat, la facilitat de gestió, l'expressivitat, l'àmbit d'aplicació, el tipus d'ontologia i la mida, granularitat i completesa. Volem explicitar, en aquest punt que, a partir d'ara tota la informació que donarem sobre μKosmos ha estat extreta directament a partir de l'eina de gestió de descriurem detalladament en la secció 5 d'aquest capítol, OntoTerm®[22].

3.4.1 Disponibilitat

Entenem per disponibilitat l'accés per obtenir l'ontologia per part del seu creador via web o per mitjà d'acords formals a través de l'Institut. En el cas de Cyc, aquesta ontologia pertany a una empresa privada. Els seus creadors permeten l'accés a una part força reduïda d'informació que no és gaire transparent i molt difícil d'utilitzar. Pel que fa a EWN i SIMPLE (i específicament PAROLE), l'accés és menys restrictiu perquè l'usuari pot obtenir informació relativa a aquests recursos per mitjà d'ELRA (*European Language Resources Association*)[23].

Finalment, quant a μKosmos i UMLS, l'IULA ha tingut accés a totes dues ontologies. UMLS ha estat obtinguda gratuïtament i el disseny de l'ontologia de μKosmos ha estat consultat directament a la bibliografia disponible a la web i, posteriorment, hem pogut navegar per l'ontologia gràcies al ja esmentat sistema de gestió. La Taula 3-2 resumeix la informació que acabem de detallar.

[22] Vull agrair al Dr. Antonio Moreno Ortiz (Universidad de Málaga) totes les seves indicacions i el seu interès per facilitar-nos l'accés a OntoTerm®.

[23] Es pot obtenir més informació a la següent adreça web: http://www.icp.inpg.fr/ELRA/home.html.

Recurs	Disponibilitat
Cyc	Subconjunt disponible
EuroWordNet	Llicència (ELRA)
µKosmos	A través d'OntoTerm®
SIMPLE	Llicència (ELRA)
UMLS	Llicència (acord institucional)

Taula 3-2. Indicadors de disponibilitat de les ontologies.

3.4.2 Facilitats de gestió (ampliació i modificació)

Un aspecte essencial per al desenvolupament d'una ontologia és la disponibilitat d'eines que ajudin a mantenir la consistència en tot el sistema. Aquesta secció reflecteix les eines que es poden utilitzar per actualitzar, ampliar o modificar cada ontologia. Segons la informació de què disposem, les eines a l'abast de l'usuari són:

a) *Cyc*. No s'ha trobat cap indicació sobre l'existència d'eines de gestió.

b) *EuroWordNet*. Per a les versions espanyola i catalana d'EWN, hi ha algunes eines de gestió dissenyades principalment a ampliar l'ontologia. També és navegable per Internet[24].

c) *µKosmos*. L'eina utilitzada per a aquesta avaluació és OntoTerm®, una aplicació de gestió d'ontologies. Proporciona una interfície molt amigable per afegir conceptes, relacions i entrades lèxiques[25].

d) *SIMPLE*. A Bel *et al.* (2000) s'esmenta l'existència d'algunes eines per al català i l'anglès però aquestes semblen més aviat d'ús intern.

e) *UMLS*. L'única eina que s'inclou en la distribució d'UMLS és MetamorphoSys, un sistema que, com ja hem vist, permet adaptar i crear

[24] El navegador es pot obtenir a la següent url: http://nipadio.lsi.upc.es/cgi-bin/public/wei2.html.

[25] Una versió demo d'aquesta eina es troba disponible a: http://www.ontoterm.com.

subconjunts del UMLS Metathesaurus per cobrir més adequadament les necessitats dels usuaris[26].

3.4.3 Expressivitat

Totes les ontologies analitzades presenten diferents tipus de formalismes. Un dels principals paràmetres distintius per avaluar aquestes ontologies és el concepte i l'expressió de les relacions en cadascun dels formalismes. A continuació, indiquem algunes de les característiques sobre aquest aspecte per a cada ontologia:

a) *Cyc*: Utilitza CycL, un llenguatge de representació, que és essencialment un tipus de càlcul de predicats de primer ordre amb alguns trets addicionals: igualtat, afegitons per manca de raonament, lematització i un mínim càlcul de segon ordre (per exemple, la quantificació per sobre dels predicats es permet en algunes circumstàncies).

b) *EuroWordNet*: Descriu els conceptes (els synsets) com un conjunt de variants. Presenta un nombre finit de relacions i la seva eina de gestió és restrictiva pel que fa al tipus de relacions que es poden incloure (només les predefinides). Defineix un nivell superior (*top ontology*) d'acord amb els principis de semàntica lèxica més rellevants[27]. La informació s'hereta del seu antecessor excepte en els casos en què alguna part d'aquesta informació es redefinida o bloquejada.

c) *μKosmos*: Els conceptes es descriuen per la seva posició a l'ontologia i per la indicació de les seves propietats i valors[28]. Les relacions no es troben restringides pel que fa al nombre però cal definir, per a cada relació directa, el seu corresponent concepte de relació inversa. μKosmos permet l'herència múltiple que, mitjançant la seva eina de gestió, es pot visualitzar com a

[26] Algunes de les necessitats dels usuaris poden ser: excloure vocabularis que requereixin una llicència, excloure vocabularis que no els siguin útils, personalitzar el recurs segons la finalitat, etc.

[27] Segueix essencialment la teoria proposada per Pustejovsky (1995).

[28] Utilitzant l'eina de gestió OntoTerm® es pot visualitzar una definició en llenguatge natural de la majoria de conceptes.

herència exclusiva (només del pare, del concepte directament superior) o bé acumulativa (herència provinent de tot el camí d'hiperonímia).

d) *SIMPLE*: Cada unitat lèxica es descriu mitjançant un sistema de tipus organitzat sobre la base de l'herència ortogonal (d'acord amb l'estructura proposada a Pustejovsky, 1995). Tota la informació semàntica s'afegeix per refinar la informació lingüística (això és, tipus semàntics per a cada argument, relacions entre unitats semàntiques, etc.).

e) *UMLS*: Cada concepte té un lloc en la xarxa semàntica i es descriu per una etiqueta denominativa. Els conceptes es relacionen els uns amb els altres mitjançant un conjunt força ric i controlat de relacions específiques del camp mèdic (i també les relacions generals d'hiponímia i meronímia). UMLS presenta, *a priori*, un mecanisme d'herència simple però aquest procés es pot bloquejar quan sigui necessari.

De les ontologies analitzades, podem fer-ne dos grups. Un primer grup format per les ontologies que tenen jerarquies i informació associada a cada node de l'estructura jeràrquica (EWN, μKosmos i UMLS). Un segon grup constituït per les altres dues ontologies analitzades (Cyc i SIMPLE), en què la informació és representada i organitzada diferentment, seguint uns patrons no jeràrquics.

Tanmateix, totes cinc ontologies inclouen algun tipus de definició per als conceptes que recullen. L'expressió de les definicions en llenguatge natural es dóna de manera diferent: mitjançant definicions formals, una glossa, exemples, un context definitori, etc.

3.4.4 Àmbit d'aplicació

Com ja hem comentat, la majoria de les ontologies analitzades en aquest treball no són pròpies d'un determinat àmbit temàtic. Al marge d'UMLS, que està orientada a l'àmbit de la medicina, les altres quatre ontologies cobreixen informació general. De tota manera, aquestes ontologies tenen diferents branques del coneixement desenvolupades en major o menor mesura. Molt probablement, μKosmos ha desenvolupat considerablement les branques de l'ontologia que tenen a veure amb

l'àmbit de l'economia i la fusió d'empreses. En el cas d'EWN, aquest recurs ha desenvolupat de manera asimètrica diferents àmbits, segons els interessos dels usuaris que l'han utilitzat.

3.4.5 Tipus d'ontologia

Molt breument, i pel que fa al tipus d'ontologia, cal dir que tant EWN com SIMPLE han estat concebuts des del punt de vista del lèxic (ja hem dit que SIMPLE no és *per se* una ontologia sinó que utilitza una ontologia general com a recurs semàntic). Per tant aquestes dues ontologies són ontologies anomenades lèxiques. Contràriament, μKosmos, UMLS i Cyc poden classificar-se com a ontologies conceptuals. Excepte Cyc, la informació en aquests casos es representa mitjançant els conceptes que s'expressen utilitzant diferents etiquetes que contenen tota la informació requerida (vegeu l'apartat 3.4.3 sobre expressivitat) i permeten transmetre el seu significat.

3.4.6 Mida, granularitat i completesa

La mida dels recursos analitzats és molt diferent. La Taula 3-3 mostra les xifres globals per a cada recurs en les diferents llengües considerades.

Recursos	Mida de l'ontologia	Mida per llengües		
		Anglès	Espanyol	Català
Cyc	3.000	14.000	0	0
EWN		90.000	50.000	20.000
μKosmos[29]	4.800	0	0	0
SIMPLE		?	3.000	3.000
UMLS (Edició 2001)	134	800.000	30.000	0

Taula 3-3. Recursos analitzats: comparació de mida.

Al marge de la informació recollida a la Taula 3-3, cal tenir en compte que no totes les ontologies presenten el mateix nivell de granularitat en tots els àmbits[30]. La

Figura 3-13 mostra la informació relativa al concepte "body part" a μKosmos i EWN. Totes dues ontologies inclouen el concepte però el nombre d'hipònims és força divergent: 1.639 conceptes a EWN i 42 a μKosmos. Aquest desequilibri es deu molt probablement el fet que EWN és una ontologia general orientada al lèxic, que ha estat ampliada en l'àmbit mèdic mentre que μKosmos també és una ontologia general però està destinada a servir de suport a un sistema de traducció automàtica orientat a l'àmbit econòmic. A més, el criteri d'ampliació és força diferent en ambdues ontologies, mentre que en el cas d'EWN es fomenta el màxim de detall per als conceptes, en el cas de μKosmos, el criteri d'ampliació sembla restringir-se a afegir només els conceptes que siguin estrictament necessaris per al sistema de traducció. UMLS mereix un comentari al marge atès que cobreix l'àmbit mèdic amb un nombre determinat de conceptes i, per tant, la granularitat no és un paràmetre directament comparable[31].

[29] Existeixen diversos mòduls lèxics (anglès, japonès i espanyol) per a μKosmos, però en cap cas no s'indica el nombre d'entrades lèxiques. A la implementació d'OntoTerm, el sistema proporciona l'eina per incloure tota la informació lèxica per a diverses llengües (el sistema forneix una llista amb tots els codis de llengua previstos per la ISO). La informació lèxica, lligada al concepte, s'organitza d'acord amb les llengües implicades i utilitzant una plantilla dissenyada prèviament

[30] Fixeu-vos que, en alguns casos, aquesta informació no mereix cap indicació.

[31] Si agafem com a exemple la unitat lèxica *asthma*, EWN defineix la cadena d'hiperonímia següent: *asthma→respiratory disease→disease*. Ben al contrari, UMLS relaciona aquesta unitat lèxica amb el tipus semàntic *Disease or Syndrome*. Aquesta unitat lèxica no apareix a μKosmos.

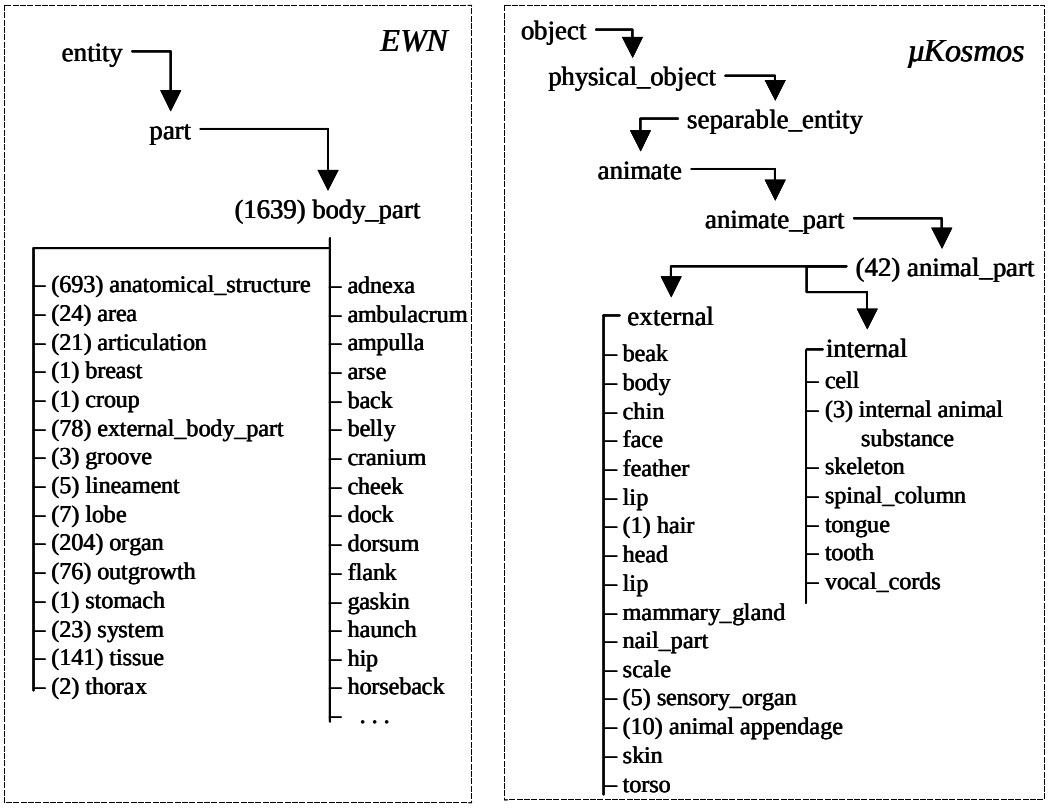


Figura 3-13. Completesa: comparació entre μKosmos i EWN[32].

3.5 OntoTerm: un sistema de gestió de terminologia basat en una ontologia

OntoTerm® és un sistema de gestió terminològica (SGT) creat per A. Moreno, professor de la Universidad de Málaga, amb la finalitat de superar alguns dels principals problemes de les bases de dades terminològiques existents en dos sentits. D'una banda, es basa en una modelització conceptual, això és, l'àmbit temàtic ha d'estar estructurat abans d'entrar-ne els corresponents termes. La construcció d'un ontologia esdevé, en aquest enfocament, el primer pas en la construcció de la base terminològica. D'altra banda, està orientat a l'intercanvi d'informació terminològica i per aconseguir-ho, implementa els estàndards per a l'intercanvi de terminologia: Martif (ISO 1220) i les categories de dades del CLS Framework (ISO 1620).

[32] En aquesta figura només hi ha representada la relació "*is-a*". El número al costat d'alguns conceptes indica la quantitat d'hipònims.

OntoTerm® no permet entrar termes en la base de dades terminològica si el seu corresponent concepte no ha estat prèviament explicitat, introduït, en l'ontologia. Aquest SGT basat en una ontologia presenta quatre mòduls[33]:

- Ontology Editor (l'editor d'ontologies)
- TermBase Editor (l'editor de la base de dades terminològiques)
- Ontology Navigator (el navegador de l'ontologia)
- HTML Report Generator (el generador de fitxes HTML).

L'Ontology Editor és el mòdul que permet a l'usuari crear una nova ontologia[34] o obrir-ne una d'existent i és la primera pantalla que apareix quan engeguem el programa. L'ontologia es visualitza de la manera següent:

[33] Per a una informació més detallada d'OntoTerm vegeu la pàgina oficial del programa a http://www.ontoterm.com i, per a una informació complementària en català sobre el funcionament de l'eina podeu consultar el *Manual d'ús d'OntoTerm*, redactat per Judit Feliu i Martí Quixal. Barcelona: Institut Universitari de Lingüística Aplicada, 2002.

[34] Quan creem una nova ontologia, el sistema proporciona un subconjunt de l'ontologia de μKosmos com si fos una miniontologia de base (sense aquests conceptes el programa no funciona).

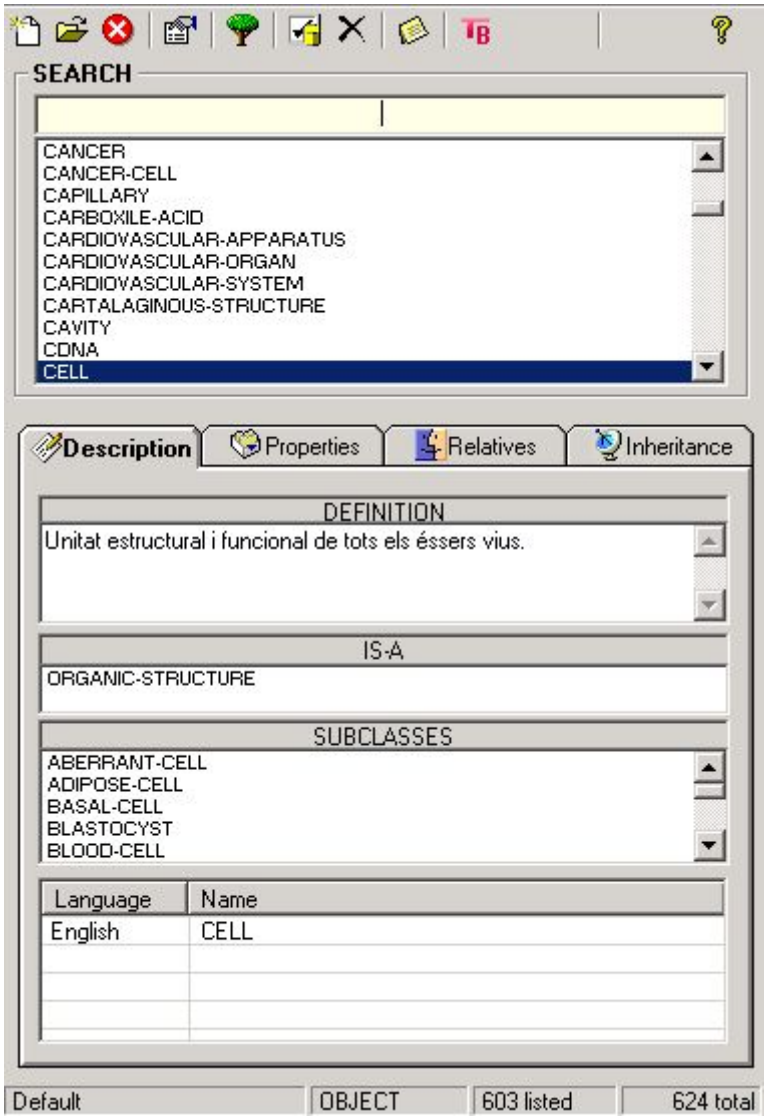


Figura 3-14. Pantalla principal d'OntoTerm®.

L'Ontology Editor permet crear o modificar ontologies; afegir, modificar o esborrar conceptes seleccionats d'una ontologia; veure l'arbre de l'ontologia de manera total o parcial i incloure anotacions relatives al concepte.

Per a cada concepte, és possible d'incloure una descripció (definició del concepte, relacions d'hiponímia i hiperonímia, indicació per llengua), les seves propietats, els conceptes relacionats i la informació sobre l'herència. El sistema organitza la informació a partir dels conceptes, els atributs i les relacions. Els atributs i les relacions es poden assignar localment o bé heretar-se. Pel que fa a l'herència, cal esmentar que el sistema presenta herència exclusiva (totes les relacions i atributs que

corresponen al concepte i les que corresponen només al seu concepte immediatament superior), i herència acumulativa (que inclou l'herència exclusiva més totes les relacions i atributs que rep el concepte de tot el camí d'hiperonímia).

El segon mòdul principal d'OntoTerm® és l'Ontology Navigator. Permet veure —permet navegar mitjançant hipervincles de les pàgines generades— tota la informació extreta de l'ontologia, això és, tots els conceptes organitzats en un arbre jeràrquic, la seva definició i les relacions entre els conceptes.

A més, des d'un punt de vista lingüístic i terminològic, cal destacar el TermBase Editor. És el mòdul destinat a la creació, edició i navegació de les unitats terminològiques corresponents als conceptes de l'ontologia. Utilitza una llista de categories estàndards (seguint la norma ISO) per assignar informació lingüística a cada concepte de l'ontologia. El seu format de visualització és:

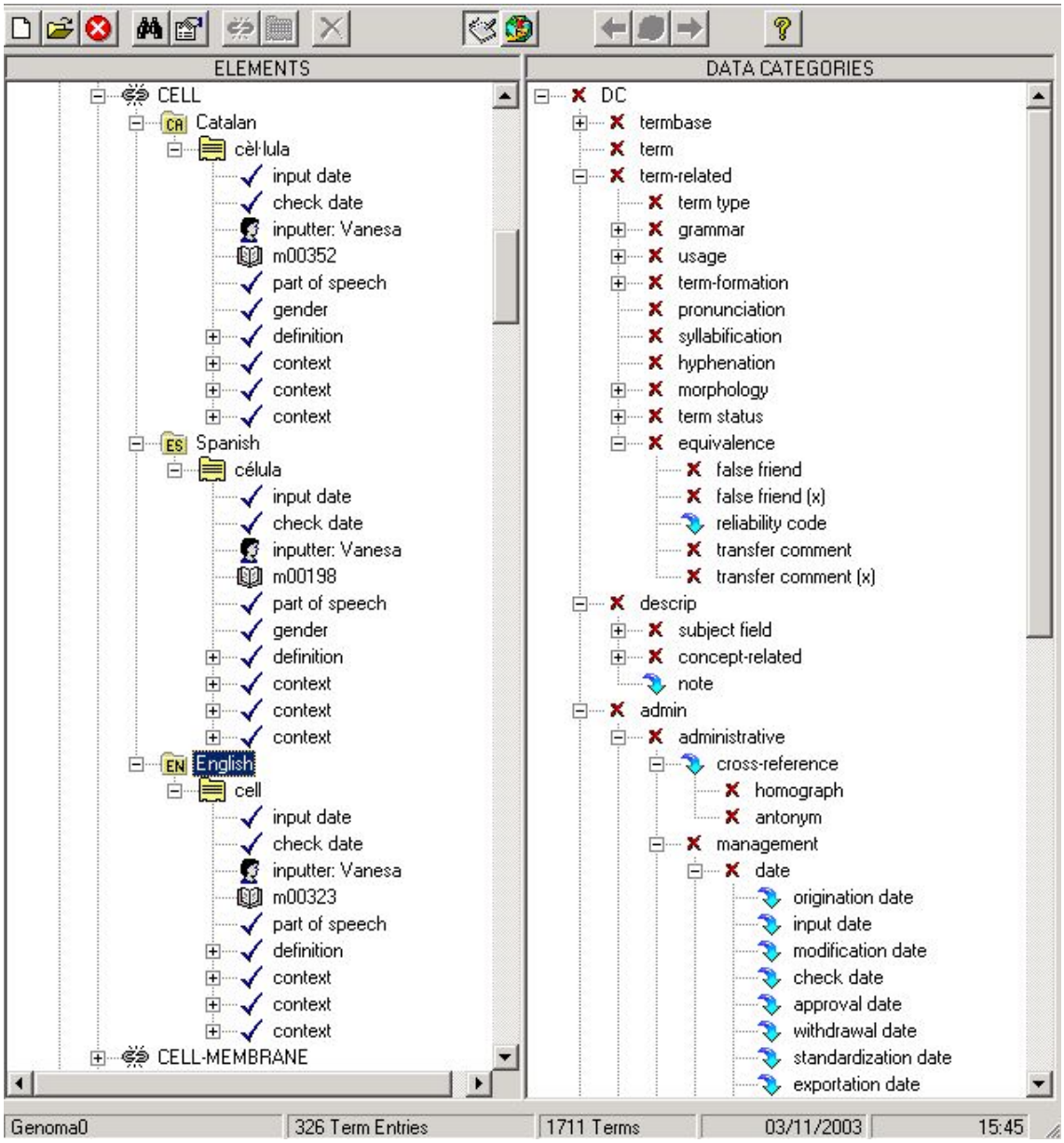


Figura 3-15. TermBase Editor d'OntoTerm®.

Finalment, l'HTML Report Generator és l'eina que permet a l'usuari de seleccionar els conceptes que desitja per a la creació de pàgines web. Es tracta del format de sortida per a tota la informació inclosa en els mòduls de l'eina OntoTerm® descrits anteriorment.

Abans de passar a l'apartat final d'aquest capítol, on presentarem el tractament de les relacions conceptuals en l'ontologia creada per al projecte GENOMA en el marc de l'Institut, volem dir que l'eina de gestió que hem utilitzat ha estat OntoTerm® i és per

aquest motiu que hem cregut essencial aquest breu comentari sobre la seva constitució i les grans línies de funcionament d'aquesta eina de gestió de terminologia basada en una ontologia.

3.6 L'ontologia del projecte GENOMA: tractament de les relacions conceptuals

En els apartats precedents, hem observat el diferent tractament de les relacions conceptuals en les cinc ontologies analitzades. Si ens fixem, a tall de resum, en les tres ontologies de les quals podem extreure informació per a l'ontologia del genoma humà, vegem que el nombre i diversitat de relacions recollides és molt gran. Com afirma Barrière (2002: 93):

> «Semantic networks are composed of concepts and relations. While it is generally agreed that the set of concepts is unlimited and corresponds to the open-class set of words, researchers do not agree about the size and nature of the set of semantic relations needed to capture all knowledge that can be expressed in natural language. (...) The sets in use do overlap; and often their different numbers, ranging from 25 to over 100, do not reflect disagreement over which relations are important but arise primarily from representing knowledge at different levels of precision (also referred to as granularity or detail).»

Així, EWN estableix quatre tipus de relacions semàntiques bàsiques i, seguidament, segons les diferents versions de la base de dades lèxiques, s'introdueixen les relacions necessàries per donar compte dels lligams entre synsets del tipus *causes* o *pertains_to*. Pel que fa a UMLS, s'han definit 54 relacions entre conceptes, d'entre les quals podem esmentar *physically_related_to*, *developmental_part_of*, *etc.* Finalment, per al cas de μKosmos, les relacions no jeràrquiques, que se sumen a la relació d'hiponímia-hiperonímia, són molt abundants i es troben estructurades de la manera següent:

- EVENT-OBJECT-RELATION (enactment, ingredient, origin, received-by...)

- EVENT-RELATION (condition, replacement-for)
- EVENT-STATE-RELATION (agent, beneficiary, depend, opposite...)
- INVERSE-EVENT-OBJECT-RELATION (enactment-of, ingredient-of, behind...)
- INVERSE-EVENT-RELATION (condition-of, replaced-by)
- INVERSE-EVENT-STATE-RELATION (agent-of, depended-on, path-of...)
- INVERSE-OBJECT-RELATION (has-example, language-of, accessible-to...)
- OBJECT-RELATION (approximately, equal-to, connects, painted-by...).

Podem afirmar, doncs, que el tractament de les relacions conceptuals varia en funció de l'eina que s'utilitza per a gestionar el coneixement especialitzat i de les necessitats de representació d'aquest coneixement en els diferents àmbits (lèxic general i lèxic especialitzat).

En el marc del nostre treball, hem introduït la tipologia de relacions conceptuals presentada en el capítol precedent en l'eina de gestió d'ontologies que acabem de descriure: OntoTerm®. Val a dir que el sistema obliga a predefinir, per a cada tipus de relació conceptual, la seva corresponent relació conceptual inversa. Per tant, si el concepte *a* té relació amb *b*, quan seleccionem el concepte *b* per a la seva visualització, el sistema ens mostrarà també quina és la relació inversa que manté amb el concepte *a*. Els set grans tipus de relació conceptual, i els seus subtipus apareixen de la manera següent en el sistema de gestió:

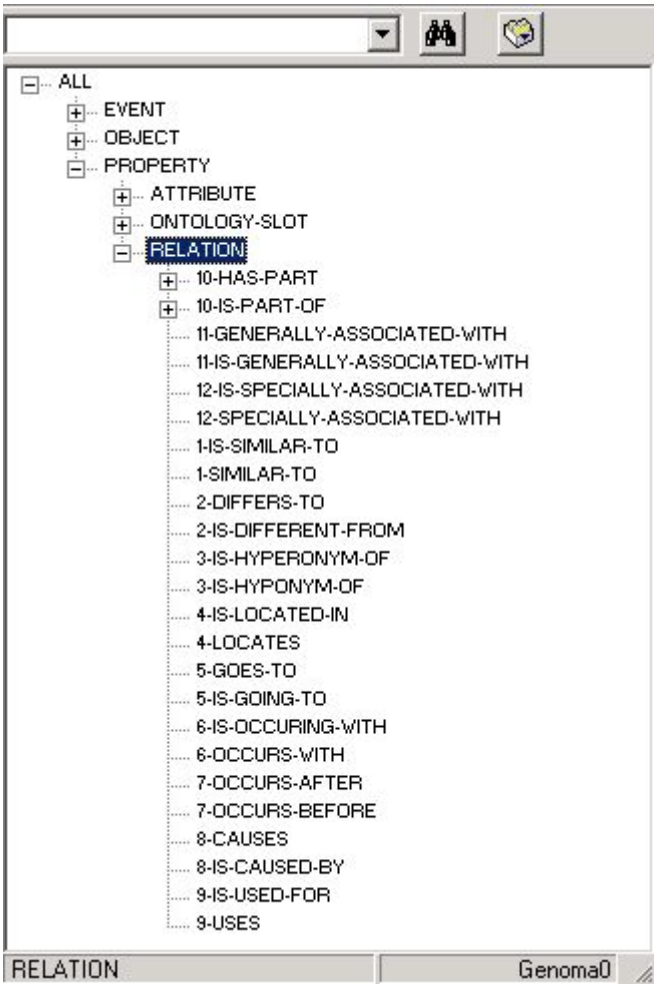


Figura 3-16. Principals tipus de relacions conceptuals incloses a OntoTerm®.

I, pel que fa a les característiques de les unitats, en el cas de la relació meronímica, la representació arbòria de les relacions conceptuals pren la forma següent:

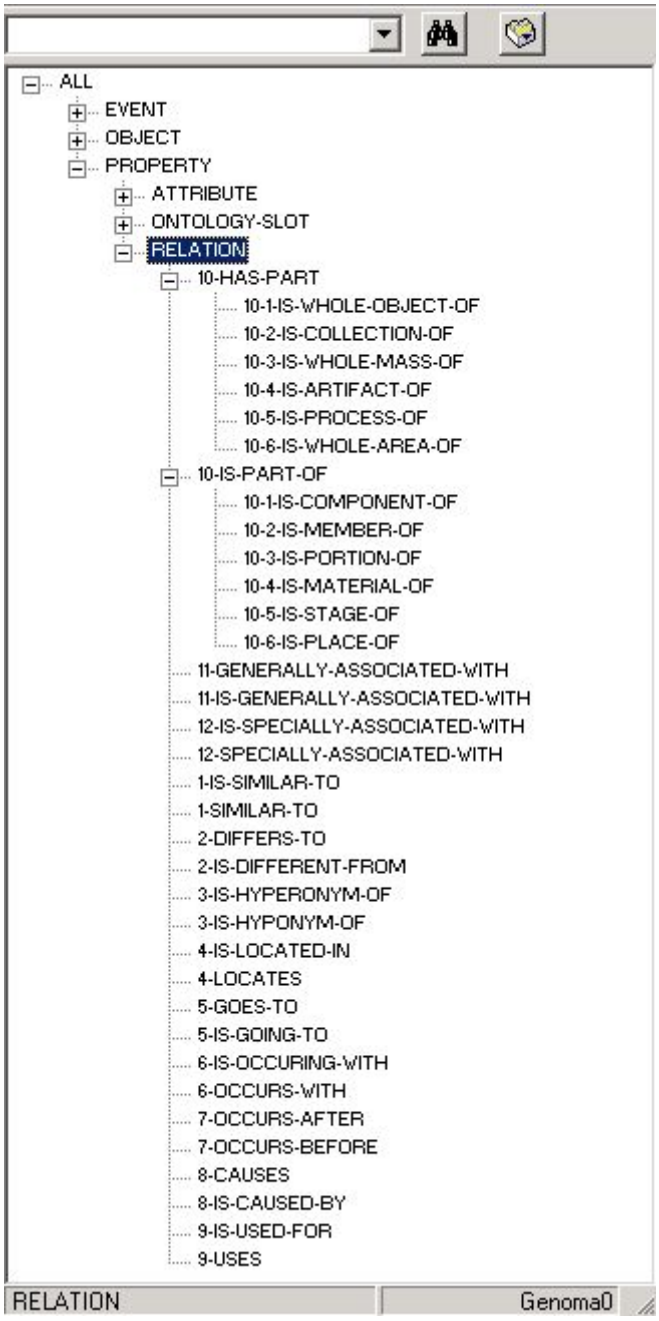


Figura 3-17. Tipus i subtipus de relacions conceptuals introduïdes a OntoTerm®.

La figura següent mostra l'aplicació d'algunes d'aquestes relacions en el marc de l'ontologia. Hem de dir que en l'ontologia, tots els conceptes s'organitzen sobre la base de la relació jeràrquica d'inclusió, és a dir, que entre tots els conceptes subordinats hi ha un lligam *is_a* amb el seu o els seus corresponents superordinats. Tanmateix, algunes altres relacions conceptuals han estat ja aplicades i permeten vincular els conceptes de l'ontologia. Vegem l'exemple del concepte 'cell':

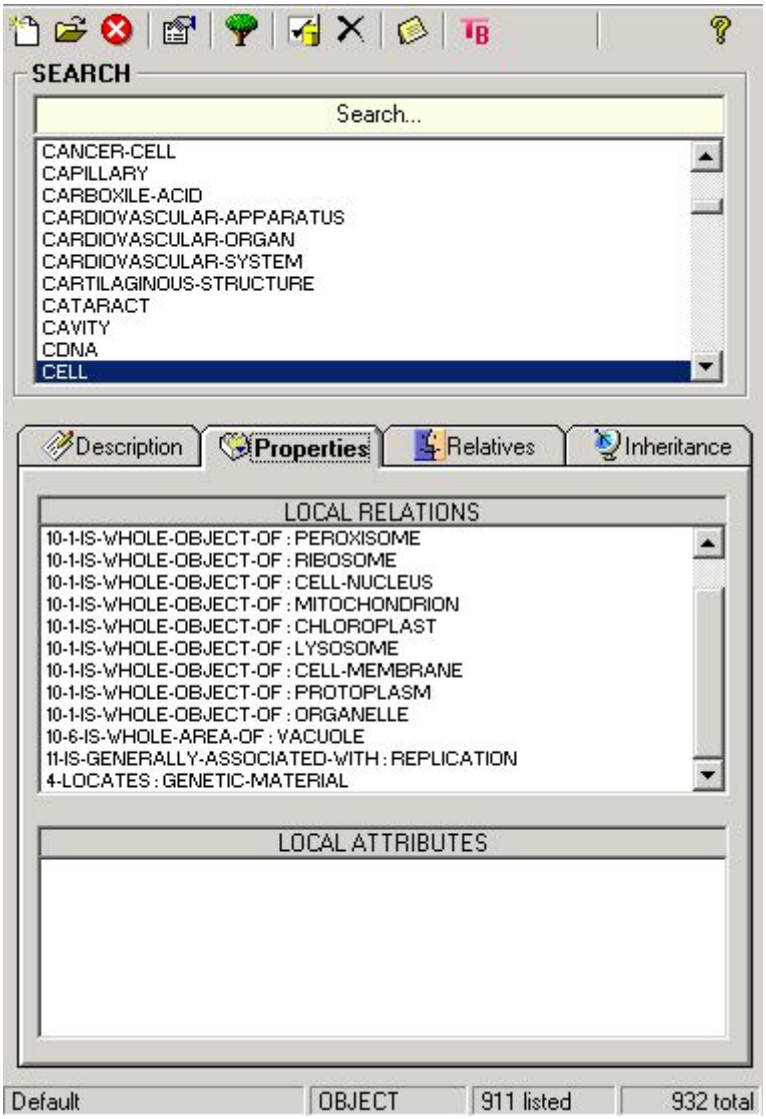


Figura 3-18. Indicació de les relacions conceptuals per al concepte ‘cell’.

On s’ha establert la relació de meronímia (component-objecte) entre els conceptes ‘cell’ i ‘golgi-apparatus’, ‘endoplasmic-reticulum’, ‘peroxisome’, ‘ribosome’, ‘cell-nucleus’, ‘mitochondrion’, ‘chloroplast’, ‘lysosome’, ‘cell-membrane’, ‘protoplasm’ i ‘organelle’; la relació de meronímia (lloc-àrea) entre els conceptes ‘cell’ i ‘vacuole; la relació associativa general entre els conceptes ‘cell’ i ‘replicació’; i la relació seqüencial locativa entre els conceptes ‘cell’ i ‘genetic-material’, com podem observar seleccionant la pestanya "properties" sota el quadre de diàleg on s’indiquen les "local relations".

A més, cadascun dels conceptes que apareixen relacionats estan vinculats amb el concepte 'cell' per mitjà de les respectives relacions inverses que, el sistema, adjudica de manera automàtica. Ocorre de manera anàloga amb el concepte 'transcription' la representació conceptual del qual es mostra de la manera següent:

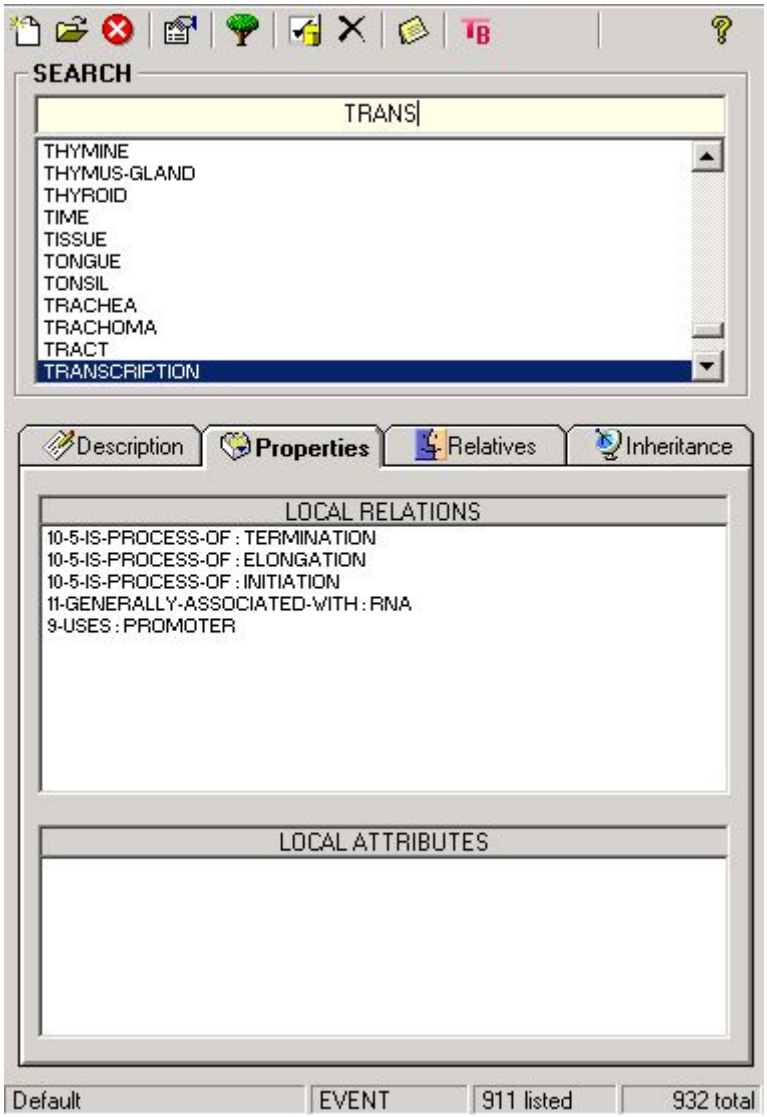


Figura 3-19. Representació de les relacions conceptuals per al concepte 'transcription'.

En la figura 3-19 disposem de 5 relacions conceptuals. Més concretament, s'observa que els conceptes 'termination', 'elongation' i 'initiation' estan en relació meronímica de procés-resultat amb el concepte 'transcription'. Així mateix, el procés

de ‘transcription’ està associat al concepte ‘RNA’ i, finalment, el procés de ‘transcription’ utilitza, com a instrument d’una funció, el concepte ‘promoter’.

No voldríem acabar aquest capítol sense indicar que l’ontologia conté, en el moment de la recerca, 932 conceptes (entre els quals es comptabilitzen les relacions conceptuals que es consideren també conceptes). Aquesta ontologia presenta, de manera estretament lligada, un mòdul lèxic en el qual s’està treballant per disposar de totes les unitats terminològiques corresponents a cadascun dels conceptes. És en una part d’aquest mòdul lèxic, separat tanmateix de les unitats terminològiques, on podríem introduir, com a informació complementària i per a cada tipus de relació conceptual, tots els marcadors lingüístics de relació conceptual verbals que en el capítol següent es presentin com a rellevants. Per tant, l’ontologia i el mòdul lèxic completat, d’una banda, amb les unitats terminològiques i, de l’altra, amb les unitats verbals que vehiculen relacions conceptuals, serviran com a informació de base per al sistema d’extracció semiautomàtica de relacions conceptuals, les característiques principals del qual es presenten en el capítol 6 d’aquesta tesi.

Capítol 4

Anàlisi de les dades

Capítol IV

4 Anàlisi de les dades

4.1 Introducció

Aquest capítol està dedicat íntegrament a la descripció del corpus d'anàlisi que hem utilitzat per validar els marcadors lingüístics explícits que vam establir mitjançant una anàlisi manual en el treball de recerca i en treballs posteriors (Feliu, 2000; 2001). Així, començarem descrivint les característiques del corpus d'anàlisi, principalment, pel que fa a la grandària i a la temàtica. Seguidament, indicarem per a cada tipus de relació conceptual quins han estat els marcadors que han servit de punt de partida per a l'explotació del corpus i, encara per a cada tipus de relació conceptual, indicarem el grau de precisió de cadascun dels marcadors analitzats.

4.2 El corpus d'anàlisi

Per tal de constituir un corpus d'anàlisi que asseguri la representativitat pel que fa al fenomen estudiat (Biber, 1993), hem procedit, en primer lloc, a decidir que la nostra llengua de treball és exclusivament el català. Un cop presa aquesta decisió, hem acordat que el corpus ha de tenir al voltant de cent mil paraules i que està format per documents de diferents subàrees temàtiques dins de l'àmbit temàtic del genoma humà.

Més concretament, hem constituït un corpus d'anàlisi de 109.816 paraules repartides en 18 documents sobre el genoma humà classificats en les subàrees temàtiques

següents: estructura interna (5), enginyeria genètica (2), biotecnologia (1), malalties (3), immunologia (1), neurociència (1), filogènia (3), diferenciació (1) i recerca genètica (1). Aquests són els documents de què disposàvem en el Corpus Tècnic de l'IULA en el moment de la recerca i després d'haver dut a terme el procés d'introducció de les unitats neològiques que apareixien en els textos[1].

La indicació explícita de la subàrea temàtica a què pertanyen aquests documents respon a la voluntat d'ampliar el tipus de documents en diferents subàrees i, molt probablement, en nivells d'especialització diversos. En el nostre estudi, però, no farem diferències en el tipus de relació que es dóna en cada subàrea temàtica atès que creiem que les relacions conceptuals s'estableixen entre determinats tipus de conceptes però cadascuna d'elles és susceptible d'aparèixer en textos de diferents nivells d'especialització[2].

Per fer les cerques sobre els marcadors verbals de relació conceptual al corpus, hem utilitzat l'eina d'explotació *BwanaNet*, desenvolupada per J. Vivaldi i disponible en el marc de l'Institut. Aquesta eina ens ha permès de fer cerques força precises sobre el conjunt de documents seleccionats. Vegeu la figura 4-1 per a la interfície de consulta utilitzada per a les cerques:

[1] El corpus amb els marcadors objecte d'anàlisi en aquest capítol es troba disponible i accessible en la seva totalitat en el CD-ROM que adjuntem com a annex al final d'aquest volum.

[2] Per a una anàlisi sobre l'aparició de les relacions conceptuals contrastades en textos de dos nivells diferents d'especialització vegeu Feliu 2000.

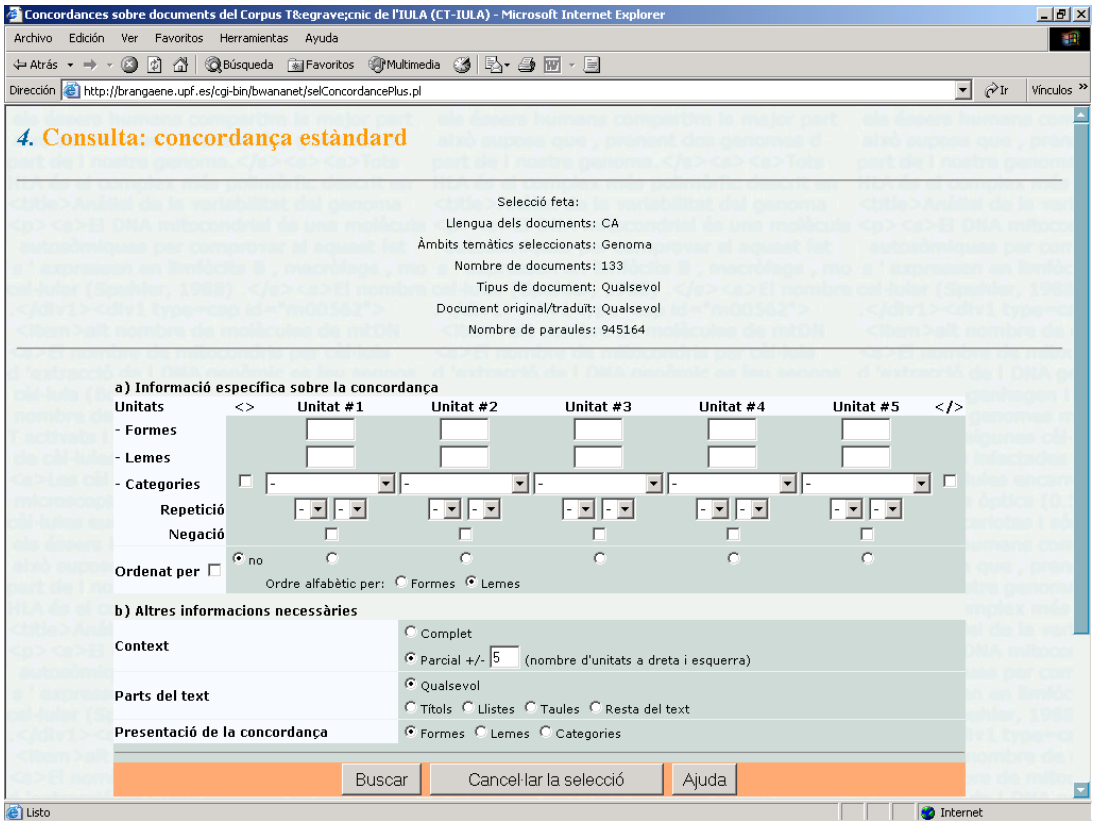


Figura 4-1. Formulari de consulta estàndard a *BwanaNet*.

Així, hem restringit la cerca als contextos de la frase (marcada estructuralment com a <s> i </s>) però també als elements d'una llista (<item>), les cel·les d'una taula (<cell>) i, finalment, als títols (<head>) que, en molts casos esdevenen fragments on la relació i les unitats terminològiques adquireixen un pes molt elevat atès que condensen força informació en una unitat estructural molt ben delimitada.

Del que acabem de dir es desprèn que en el procés d'anàlisi s'han acotat els fragments d'estudi a l'àmbit de la frase (i els seus derivats estructurals) i, en cap cas, no ens hem proposat d'analitzar les relacions que es van encadenant amb l'ajuda de marcadors discursius i connectors, al llarg de fragments discursius més grans.

4.3 Marcadors lingüístics de relacions conceptuals

En aquest apartat, presentem els resultats de l'anàlisi del corpus pel que fa als set grans tipus de marcadors lingüístics de relacions conceptuals (MLRC) corresponents

a la relació de semblança, inclusió, seqüencialitat, causalitat, instrumental, meronímia i associació amb els subtipus de semblança positiva o negativa i seqüencialitat espacial o temporal.

Així, per a cada tipus de relació, indiquem quins han estat els MLRC buscats al corpus d'anàlisi i quins resultats quantitatius n'hem obtingut. Presentem dos tipus de resultats. En primer lloc, el nombre d'ocurrències d'un determinat MLRC, és a dir, la freqüència absoluta sobre el total de paraules que constitueixen el corpus. En segon lloc, la freqüència relativa d'encert, és a dir, el número de vegades en dades percentuals en què aquest MLRC vehicula la relació conceptual objecte d'anàlisi.

4.3.1 Relació de semblança

Recordem que la relació de semblança és la que s'estableix per equivalència o oposició entre dos o més elements. Es tracta d'una relació que presenta dos subtipus, una semblança positiva, en què s'inclouria l'equivalència total i l'equivalència parcial; i una semblança negativa, en què tractarem l'oposició i el contrast entre els elements vehiculats.

4.3.1.1 Relació de semblança positiva

Pel que fa a la relació de semblança positiva hem analitzat els següents MLRC:

1) *Això és (ser)*

2) *Assemblar-se (a)*

3) *Ser com*

4) *És a dir (ser)*

Pel que fa al primer MLRC, *això és*, hem trobat 16 ocurrències i, en tots els casos, aquest marcador actua com a connector discursiu. En cap de les ocurrències trobades tenim un lligam explícit entre dues unitats terminològiques a banda i banda del MLRC. Per aquest motiu, i encara que en algun cas aïllat puguem trobar aquesta unitat expressant una relació de semblança positiva, hem de deixar de banda aquest

marcador per al mòdul lèxic del nostre sistema de detecció de relacions atès l'elevat grau de soroll que comporta.

Quant al MLRC *assemblar-se (a*) hem trobat tres ocurrències. En totes tres ocurrències es vehicula la noció de semblança (100%) però l'aparició de pronoms personals i pronoms febles s'erigeix com un dels aspectes que caldrà tractar en el capítol següent abans de definir els patrons sintacticosemàntics que haurà de tenir en compte el sistema de detecció semiautomàtica. Vegem-ne un exemple:

a. És a dir, (els parlants d'aquestes famílies lingüístiques)$_a$ ***s'assemblarien*** tant menys ***a***(ls habitants putatius de l'Orient mitjà)$_b$ com més lluny visquessin.

Tractem ara l'estructura *ser com*, la qual apareix en 8 ocasions en el corpus d'anàlisi. D'aquestes 8 ocurrències, el 37,5 % dels casos expressa semblança positiva. Un dels possibles criteris per tal de reduir el soroll d'aquest marcador és bloquejar l'etiquetatge com a MLRC de semblança positiva quan apareix com a element inicial de la frase per tal com en aquest cas es fa una comparació a nivell macrotextual, però no entre unitats terminològiques que apareixen en el marc de l'oració o d'un element estructural equivalent. Fixem-nos que, en l'exemple *b* tenim efectivament expressada la relació objecte d'anàlisi, i en l'exemple *c*, una possible ocurrència que caldria bloquejar en el sistema.

b. (Els mecanismes genètics bàsics d'aquests bacteris)$_a$ ja ***eren com*** (els dels éssers vius actuals)$_b$ descrits al capítol 2.

c. ***És com*** si tots dos fossin dominants; per això el nom de condominància.

Finalment, en el cas del MLRC *(ser) és a dir*, hem detectat 72 ocurrències. En les dues terceres parts dels casos (48 de 72 que conformen el 66,6% de precisió), s'estableix l'equivalència entre un concepte *a* i una paràfrasi o explicació *b*, però no entre dues unitats terminològiques estrictament parlant. Només en tretze casos entren

en joc dues unitats terminològiques. Per tant, ens trobem davant d'un MLRC que efectivament expressa equivalència[3] però que en la majoria de casos estableix un procés d'expansió o de reducció (segons la direcció en què es presenta l'equivalència) d'una determinada unitat com per exemple a *d*:

d. A més d'aquests gens, hi ha altres (loci dels quals no es té constància de la seva expressió, com per exemple HLA-H, HLA-J i HLA-X, la seqüència dels quals indica que són còpies degenerades d'altres gens de classe I)$_a$, ***és a dir***, que són (pseudogens)$_b$ o (gens truncats)$_c$.

e. (Presenten un fenomen d'empremta («imprinting») genètica)$_a$, ***és a dir***, (l'efecte fenotípic és diferent segons el progenitor que transmet la malaltia)$_b$.

Tanmateix, creiem interessant de recollir també aquests casos en què trobem una paràfrasi o explicació en alguna de les dues bandes: P_R_UT o bé UT_R_P. En aquests casos, creiem que en el fons s'estableix una relació de semblança entre les dues unitats (una amb valor especialitzat més alt i, l'altra, generalment com a descripció o definició). De fet, hem decidit recollir aquesta informació per tal com representa una informació especialitzada molt valuosa per definir i atribuir propietats a una determinada UT.

Ara bé, vegem l'exemple *f* per a una equivalència que considerem prototípica entre dues unitats terminològiques:

f. Per comprendre (l'expressió de la informació genètica)$_a$, ***és a dir***, (el pas de DNA a proteïna)$_b$, cal tenir present que els gens estan formats per la combinació lineal de 4 nucleòtids diferents, mentre que les proteïnes estan formades per la combinació lineal de 20 aminoàcids diferents.

[3] Vegeu Bach, 2001 per a una explicació més detallada d'aquest marcador.

Ens interessa de destacar, tot i que en un 66,6% dels contextos les unitats no són en ambdós casos unitats terminològiques, que ens trobem davant d'un MLRC d'un gran valor i utilitat perquè el tipus d'informació que forneix es pot utilitzar per extreure contextos definitoris que es podrien incloure gairebé directament en una base de dades terminològiques. A més, la unitat *b* quan hi ha una expansió i la unitat *a* quan hi ha una reducció solen contenir informació de caràcter atributiu sobre la unitat amb la qual es vincula. Aquesta informació de caràcter atributiu es podria convertir en parelles d'atribut i valor lligades a les unitats terminològiques de la base de dades i, al seu torn, als conceptes de l'ontologia.

Gràficament, la distribució que reben els marcadors d'aquest subtipus de relació conceptual és la que mostra la Figura 4-2:

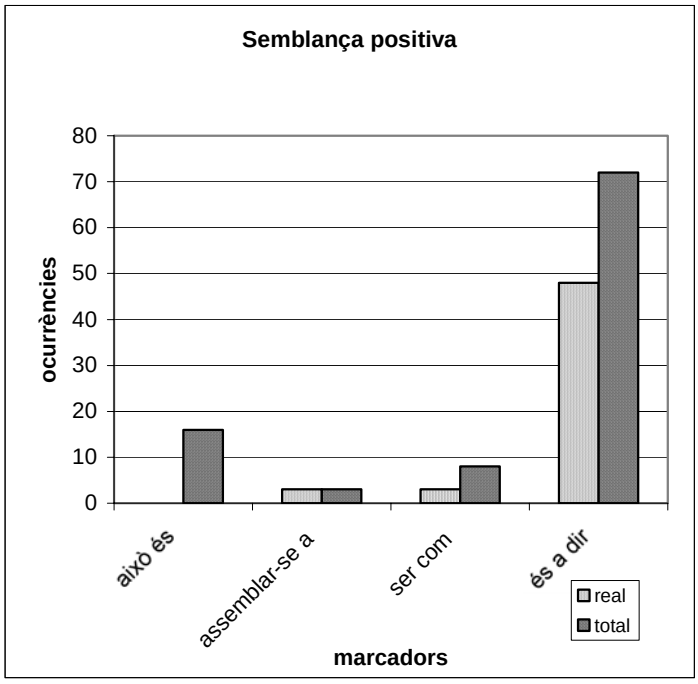


Figura 4-2. Gràfic dels marcadors de semblança positiva.

4.3.1.2 Relació de semblança negativa

Passant a la relació de semblança negativa, que fa referència a l'oposició i el contrast que es pot establir entre dos o més elements, hem analitzat els següents MLRC:

a) *Diferenciar / diferenciar de*

b) *Distingir / distingir de*

c) *Oposar*

d) *Ser el contrari*

Pel que fa als resultats obtinguts amb aquests MLRC, comencem dient que els dos últims, *oposar* i *ser el contrari* no han donat cap ocurrència sobre el total de paraules del corpus. Per aquest motiu, creiem que es tracta de dos marcadors que poden tenir, en algun cas, la característica de vehicular una relació conceptual d'oposició o contrast però que la seva productivitat és tan minsa que requereixen un esforç massa elevat pel benefici que comportaria la seva detecció semiautomàtica.

Passant al cas de *diferenciar* i *diferenciar de*, hem fet dues cerques diferents. La cerca de *diferenciar* engloba les ocurrències amb preposició que són 4 d'un total de 37. D'aquestes 4 ocurrències, 2 expressen una relació de semblança negativa (50%). Vegem-ne un exemple:

> g. Essencialment es tracta de treballar de manera que (els afectes de la manipulació genètica)$_a$ puguin ésser ***diferenciats de*** (les interaccions)$_b$.

De la resta de contextos, trobem 4 casos més en què el verb *diferenciar* expressa semblança negativa en casos en què el segueix la preposició *de*, xifra que representa un 10,81% de precisió. Algun exemple d'aquesta relació és:

> h. (Les diverses parelles de cromosomes)$_a$ ***es diferencien*** (en la seva grandària)$_b$ i (en la posició relativa del centròmer)$_c$, que pot estar situat al mig del cromosoma o desplaçat cap a un dels extrems, anomenats telòmers.

Sembla important de destacar que aquest percentatge tan baix d'encert es deu, d'una banda, als nombrosos casos en què apareix la unitat *diferencies*, que ha estat etiquetada en el corpus com a verb atès que li falta un accent i el sistema no és capaç de desambiguar-la correctament. D'altra banda, creiem molt més important d'esmentar que la unitat *diferenciar* i en nombroses ocasions la seva forma

pronominal *diferenciar-se* té un significat propi en l'àrea temàtica del genoma humà. Així, el DEM defineix la corresponent unitat deverbal com a:

> **Diferenciació** *f*
>
> Procés mitjançant el qual les cèl·lules, els teixits i els òrgans canvien d'estructura i de forma per tal d'efectuar funcions específiques. El concepte és aplicat especialment a la diferenciació cel·lular, puix que la diferenciació de teixits prové de mecanismes de regulació a diferent nivell. La diferenciació té lloc durant el desenvolupament embrionari o en processos de regeneració.

Val a dir, doncs, que ens trobem davant d'un cas en que un mateix marcador pot vehicular dues relacions conceptuals diferents. D'una banda, aquest marcador vehicula en alguns pocs casos una relació de semblança negativa d'oposició o contrast. D'altra banda, en nombroses ocasions s'utilitza aquesta forma verbal per tal d'indicar una relació conceptual associativa especialitzada mitjançant l'estructura *a* ***diferenciar(-se)*** *b*, com per exemple:

i. Recentement s'ha descrit que (aquestes cèl·lules)$_a$ *poden* ***diferenciar-se*** *cap a* (un fenotip)$_b$ que no té capacitat citolítica i que secreta mediadors i pot col·laborar amb les cèl·lules B (vegeu conseqüències de l'activació de linfòcits T).

j. (La situació d'aïllament d'una població)$_a$, per exemple, farà que una comunitat desenvolupi una llengua i una cultura peculiars, però també en ***diferenciarà*** (els gens)$_b$.

Finalment, pel que fa a la unitat *distingir* / *distingir de* trobem 2 casos en què aquest verb apareix seguit de la preposició *de* i en totes dues ocasions expressa una relació de semblança negativa:

k. El terme prió va ser elegit per destacar la hipòtesi que l'agent causal de les malalties neurodegeneratives eren (proteïnes infeccioses)$_a$ que

es distingien de (qualsevol partícula vírica)$_b$ per l'afinitat mostrada a l'àcid nucleic.

Dels 18 casos en què apareix el verb *distingir*, tenim un 100% d'encert quan aquest va seguit de la preposició *de*. Ara bé, quan aquesta preposició no apareix, es donen dues estructures sintàctiques diferents que creiem mereixen un esment. En primer lloc, tenim 6 casos en què l'estructura oracional és *distingir* a *de* b, c, n, on s'indica la semblança negativa d'oposició (33,3% de precisió), l'exemple més bàsic del qual seria:

l. Així doncs, hom pot ***distingir*** (una formiga)$_a$ d'(un peresós)$_b$, una pomera del vesc o un dofí d'un llobarro.

Tenim, però, un altre patró que s'expressaria com a *distingir* a, b, … i n. En aquest cas, el marcador lingüístics representa un patró de la relació d'inclusió de classe (hiponímia-hiperonímia), per exemple:

m. Hom pot ***distingir*** (dos grans grups de tècniques)$_a$ segons a quin nivell es faci la tipificació: (tècniques a nivell de proteïna)$_b$ i (tècniques a nivell de DNA)$_c$.

n. Els loci de classe II es localitzen a la zona més centr4omèrica del complexe HLA i podem ***distingir****-hi* (cinc subregions o famílies principals de loci)$_a$: (HLA-DR)$_b$, (HLA-DQ)$_c$, (HLA-DO/DN)$_d$ i (HLA-DM)$_e$.

Caldrà veure, doncs, en el capítol següent de quina manera es pot proposar una estratègia que permeti diferenciar els usos del marcador lingüístic *distingir*, com a vehiculador d'una relació de semblança o d'inclusió de classe. El quadre següent indica la proporció dels marcadors analitzats en aquest apartat de manera esquemàtica.

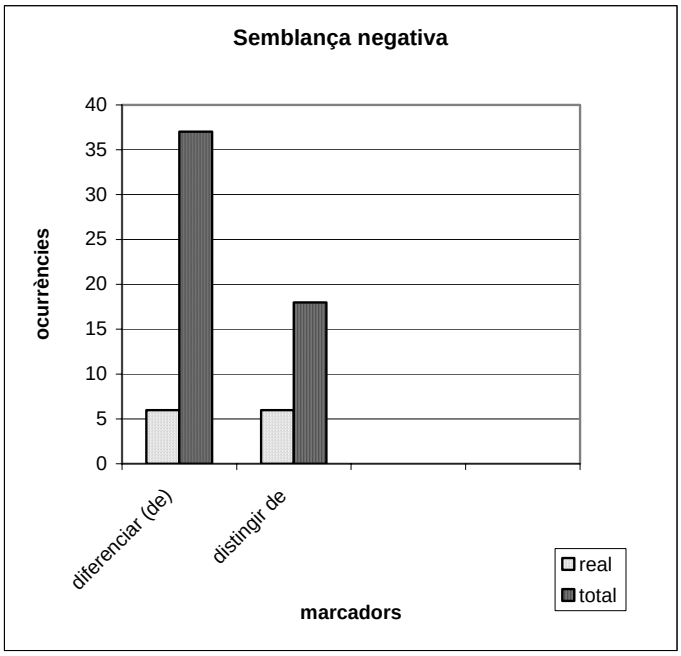


Figura 4-3. Gràfic dels marcadors de semblança negativa.

4.3.2 Relació d'inclusió de classe

Per a la relació d'inclusió de classe hem fet tres consultes essencials:

a) :

b) Com

c) *Ser + det. indef. (un, una, uns, unes).*

Deixem de banda el marcador *:* per tal com no es considera una unitat lingüística explícita verbal que componen el conjunt de marcadors seleccionats per a l'anàlisi i la consegüent proposta de sistema de detecció semiautomàtica. El resultat d'aquesta cerca ha donat 351 contextos en què apareix aquesta marca tipogràfica. D'aquests, 110, gairebé una tercera part (31,3%) donen lloc a una relació d'inclusió de classe però nosaltres ens limitarem a les expressions verbals explícites i, per tant, l'anàlisi d'aquests tipus d'estructures queden per a treballs posteriors en què analitzem molt més intensivament aquesta relació.

Anàlogament, el cas de la conjunció *com* ha donat 4 casos, el 50% dels quals expressa una relació d'hiponímia hiperonímia però, en cap cas, no apareix cap unitat verbal al costat i, per tant, deixem al marge aquests tipus d'estructures.

Ens centrem, doncs, en el marcador lingüístic més important de l'expressió de la relació conceptual d'inclusió de classe: *ser_un*. Es tracta del patró lingüístic més estudiat al llarg de tota la bibliografia i el qual rebrà un tractament en profunditat en el capítol següent que esperem ens permeti de reduir el soroll trobat en la cerca d'aquest MLRC. S'han detectat 168 ocurrències, 73 de les quals indiquen la relació d'inclusió de classe. Aquesta xifra representa un 43,45% dels casos. Podem deduir, doncs, que sense aplicar encara cap estratègia semàntica, gairebé la meitat de les ocurrències de *ser_un* expressen la relació d'inclusió. Veurem en el capítol cinquè com reduir el soroll que presenta aquesta relació, alguns dels exemples més paradigmàtics de la qual són:

o. (El mtDNA humà)$_a$ ***és una*** (molècula circular de doble cadena)$_b$ (cadena lleugera o L, light strand; i cadena pesada o H, heavy strand) de 16.569 parelles de bases que codifica 13 proteïnes (tres subunitats de la citocrom oxidasa, set subunitat de la NADH deshidrogenasa, dues subunitats de l'ATPasa i el ditocrom b), 22 RNAs de transferència i 2 RNAs ribosòmics (12 S i 16 S).

p. (El DNA mitocondrial)$_a$ ***és una*** (molècula circular de doble cadena)$_b$ que representa una petitíssima part del nostre genoma.

q. En bacteris, (la divisió cel·lular)$_a$ ***és un*** (procés)$_b$ relativament senzill.

r. Des del punt de vista molecular, (un gen)$_a$ ***és un*** (tros de DNA)$_b$ que realitza una funció concreta.

s. (L'hemoglobina)$_a$ ***és un*** (pigment respiratori)$_b$ encarregat de transportar oxigen per la sang, des dels pulmons fins als diversos teixits.

Vegem, de manera esquematitzada, la informació que hem proporcionat per a la relació d'inclusió de classe:

Figura 4-4. Gràfic dels marcadors d'inclusió de classe.

4.3.3 Relació de seqüencialitat

En aquest apartat, tractarem separadament els quatre subtipus de relació de seqüencialitat. Això és, dedicarem un subapartat específic per a la relació seqüencial espacial de localització o locativa, un subapartat a la relació seqüencial espacial de direcció i dos subapartats més a les relacions seqüencials temporals de simultaneïtat i d'anterioritat-posterioritat.

4.3.3.1 Relació de seqüencialitat espacial de localització

Els marcadors seleccionats per a aquest tipus de relació conceptual són:

a) *Aparèixer*

b) *Evidenciar*

c) *Iniciar en*

d) *Ocórrer*

e) *Originar*

f) *Produir + prep. (de, durant, en, per, gràcies a, mitjançant)*

g) *Quedar*

h) *Realitzar + prep. (amb,* a *partir de, de,* m*itjançant, per)*

i) *Situar + prep. (a, amb, en, dins, entre, per*)

j) *Tenir lloc*

k) *Veure (en)*

Començant pel marcador *aparèixer*, hem de dir que aquesta ocurrència es troba en el total del corpus en 42 ocasions. D'aquestes, 11 ocurrències, en tots els casos seguides de les preposicions *a* i *en*, indiquen una relació seqüencial locativa (26,19%). Vegem-ne dos exemples:

t. (Aquests homínids)$_a$ ***van aparèixer a*** (Kenya)$_b$ i (Sud-àfrica)$_c$ fa 2,5-1,5 milions d'anys, i ja empraven instrumentals tallats en pedra.

u. (La diferenciació dels bascos)$_a$ ***apareix en*** (la primera component principal)$_b$, sobresortint com la diferenciació genètica més important dins d'aquest context geogràfic.

De la resta dels 42 casos, hem d'esmentar que un 23,8% (10 ocurrències) expressen la relació de seqüencialitat temporal atès que el marcador *aparèixer* es presenta acompanyat de les unitats *durant, mentre, fa x (+ indicació temporal)*. Els altres 21 contextos d'aparició del marcador aparèixer presenten estructures on només es troba o bé el concepte *a* o el concepte *b* però no es manté el lligam com a mínim binari que hem descrit al capítol 2, requeriment necessari per considerar una relació conceptual. Vegem alguns exemples de la materialització temporal de la relació seqüencial i, seguidament, observarem algun cas en què no tenim dos conceptes sinó només un concepte vinculat:

v. (La teoria moderna de l'evolució)$_a$ ***aparegué durant*** (les dècades de 1930 i 1940)$_b$ com una síntesi dels coneixements genètics i del concepte de selecció natural.

w. Finalment, ***fa*** 600 milions d'anys, ***van aparèixer*** (els primers organismes pluricel·lulars)$_a$.

El marcador lingüístic *evidenciar* no indica, en cap dels 9 casos detectats en el corpus, una relació seqüencial locativa. Hem de dir que en 5 dels casos, es tracta d'una unitat que expressa una relació associativa (55,5%), i, per aquest motiu, cal analitzar si es tracta d'una relació associativa general o especialitzada. Els contextos que hem detectat indiquen que es tracta d'una relació associativa general i, per tant, aquest marcador haurà de passar a ser analitzat en el grup dels MLRC associativa. En els altres 4 casos, no apareix o el concepte *a* o el concepte *b*, o bé el verb apareix en la seva forma negativa i, per tant, aquests contextos no es comptabilitzen com a vehiculadors d'una relació conceptual.

Si passem al marcador *iniciar(-se) en*, només disposem de dos casos. Un d'ells expressa efectivament la relació locativa mentre que l'altre expressa una relació temporal de simultaneïtat. Vegem el context del cas que ens interessa i observarem que, en aquests casos, serà en el capítol següent d'estratègies de refinament on haurem d'indicar que cal tenir un mínim d'etiquetatge semàntic dels conceptes *a* i *b* per poder discernir si es tracta d'una seqüencialitat locativa o temporal:

x. (La separació de les cadenes de DNA)$_a$ ***s'inicia en*** (uns punts molt concrets dels cromosomes)$_b$ anomenats orígens de replicació. [Locativa]

y. En efecte (la cultura capsiana)$_a$ (7.000-5.000aC) ***s'inicia en*** (el paleolític)$_b$ i molts dels seus elements perviuen en el neolític tot i el canvi en el mode d'obtenció de l'aliment que suposà el neolític.

Pel que fa al marcador *ocórrer*, no hi ha cap de les 4 ocurrències que expressin localització. Ben al contrari, aquesta unitat apareix com a participi modificant un

substantiu que actua com a unitat de coneixement en el context però en cap cas no lliga dues unitats diferents.

Dels 27 casos del marcador *originar*, hem de dir que quan l'estructura sintàctica és simplement *a* originar *b*, aquest actua en un 44,4% dels casos (12 ocurrències) com a transmissor de la relació causal. Vegem-ne un exemple:

z. (Aquesta acumulació progressiva de mutacions)$_a$ ***origina*** en primer lloc (una hiperplàsia o creixement desordenat)$_b$ i després (invasivitat cel·lular)$_c$, (vascularització o angiogènesi)$_d$ i finalment (la metàstasi)$_e$.

En canvi, quan tenim una indicació temporal (*fa x [+ temporal])* o bé una indicació locativa *(en el lloc, [+ locatiu]*), de les quals només disposem d'un context de mostra de cadascun dels dos tipus, aleshores s'indica una localització o temporalització que depèn sempre de la semàntica dels conceptes involucrats en l'expressió de la relació conceptual. Vegem-ne un exemple:

aa. (La variabilitat d'aquest grup de seqüències)$_a$ ***podria haver estat originada fa*** (37.000 a 107.000 anys)$_b$ depenent de les estimes de mutació utilitzades per aquesta regió del mtDNA.

Quant al marcador *produir(-se) + preposició*, hem detectat 35 casos. La relació conceptual que es materialitza per a cada context depèn en gran mesura de la preposició que apareix, així com també de la naturalesa semàntica dels conceptes implicats. Així, hem trobat 19 casos (54,2%) en què la relació que apareix és una relació causal que té de manera subjacent la noció de seqüencialitat i que s'expressa gràcies a les preposicions *per* i *mitjançant*. En 5 casos (14,28%), quan la preposició que apareix és *a* o *en*, la relació conceptual vehiculada és efectivament la seqüencial locativa mentre que, en dos casos, l'aparició de la unitat *durant* fa que la noció de la relació expressada sigui la relació seqüencial temporal. Vegem-ne un exemple de cada:

bb. Quan es va emprar per primer cop aquest assaig, a la dècada de 1970, es va veure que el 80% de les substàncies que produeixen càncer són

mutagèniques, cosa que indica que (molts càncers)$_b$ ***són produïts per*** (mutacions)$_a$. [Causalitat]

cc. En canvi, si (la mateixa mutació)$_a$ ***es produeix en*** (una cèl·lula germinal)$_b$ (un òvul o un espermatozoide), afectarà totes les cèl·lules del descendent que provingui d'aquell gàmeta, i tota la seva nissaga, en forma de malaltia hereditària. [Localització]

dd. Se sap que (l'expansió)$_a$ és prezigòtica i ***es produeix durant*** (la gametogènesi)$_b$. [Temporal]

El marcador verbal *quedar* sol actuar com a verb de suport. Així, hem detectat força contextos d'aparició d'aquest marcador, concretament 39, però només en 4 casos (10,25% de precisió) es tracta d'una relació locativa. En la resta de casos, com acabem d'esmentar, aquest verb apareix seguit d'altres verbs en forma de participi i, per tant, no té un significat propi que vehiculi la relació seqüencial locativa. Vegem un exemple en què efectivament, tot i l'aparició d'un verb de suport, s'expressa aquesta relació de localització gràcies a l'aparició de la preposició *a* o *en*:

ee. Com hem esmentat, un cop es produeix un error en un gen, (aquest error)$_a$ *pot* ***quedar reflectit en*** (la proteïna que codifica)$_b$ en forma de canvis més o menys greus d'aminoàcids.

De les 21 ocurrències del marcador *realitzar + preposició*, tenim que la relació conceptual que es materialitza depèn essencialment de la preposició que acompanya l'element verbal. Així, quan la preposició és *a* i *durant* més un concepte vehiculador d'informació temporal, tenim una relació seqüencial temporal (28,57% dels casos); quan trobem la preposició *en* disposem d'una relació seqüencial locativa (només un 9,5% dels casos) i, quan l'element preposicional és *amb* o *a partir de*, s'expressa una relació d'instrumentalitat (19%). Com es pot observar, aquests percentatges deixen un marge d'error d'un 43% aproximadament on no es dóna cap relació conceptual atès que apareix una negació davant del marcador lingüístic o bé el context no ens proporciona el concepte *a* o *b*. Vegem un exemple de cadascun dels tipus de relació conceptual depenent de la preposició emprada:

ff. (Els primers experiments de transferència de material genètic en cèl·lules humanes)$_a$ ***es van realitzar durant*** (els anys vuitanta)$_b$. [Temporal]

gg. Recentment, (els avenços)$_a$ ***realitzats en*** (el camp de la genètica)$_b$ i (la biologia del desenvolupament)$_c$ han proporcionat dades molt interessants que ajuden a entendre els mecanismes evolutius. [Locativa]

hh. (L'anàlisi d'aquests polimorfismes)$_b$ ***es realitza mitjançant*** (anticossos comercials)$_a$ que aglutinaran els eritròcits segons posseeixin l'antigen corresponent. [Instrumental]

De manera semblant, el marcador *situar* expressa localització o temporalitat segons la preposició que l'acompanya. Així, dels 24 contextos de què disposem, en un percentatge força elevat dels casos, un 87,5%, s'expressa una relació seqüencial locativa mentre que en 3 ocurrències, que conformen el 12,5% restant, ens trobem davant d'una relació temporal. Cal analitzar, doncs, el contingut semàntic dels conceptes relacionats per poder determinar, a més de la preposició, en quins casos ens trobem davant d'una localització o d'una indicació temporal

ii. (Ambdues síndromes)$_a$ ***estan situades a*** (la mateixa regió del cromosoma 15)$_b$ i estan causades per la pèrdua d'aquesta petita regió, algunes vegades visible per citogenètica i d'altres només a nivell molecular, perquè són molt petites o bé perquè es deuen a disomies uniparentals. [Locativa]

jj. Per aconseguir-ho, (els cromosomes duplicats units encara pel centròmer)$_a$ ***se situen al*** (centre de la cèl·lula en un pla)$_b$. [Locativa]

kk. A més, anàlisis recents de seqüenciació d'una part del DNA mitocondrial assenyalen l'existència d'una (expansió de la població)$_a$ ***situada a l'entorn de*** *fa* (20.000 anys)$_b$. [Temporal]

En relació al marcador lingüístics *tenir lloc*, hem de dir que hem aïllat 11 contextos d'aparició 4 dels quals expressen una relació locativa i 4 expressen una relació de causalitat, que configura un 36,3% de precisió en cada cas. Per tant, cal tenir en compte quina és la preposició que apareix en cada cas per distingir una relació d'una altra. En el cas de la localització, les preposicions que indiquen aquesta noció són *a*, *en* i *dins*. En canvi, en els altres casos, o bé no apareix cap preposició o sorgeix l'estructura *a* tenir lloc entre *b* i *c*. Vegem-ho:

ll. Els encebadors hibriden cada un amb una de les dues cadenes de la regió del DNA que es vol amplificar i estan orientats de manera que (la síntesi del nou DNA)$_a$ ***té lloc*** *només* ***dins de*** (la regió que ens encebadors flaquegen)$_b$. [Locativa]

mm. Major exposició del mtDNA a danys oxidatius, deguda als (processos de fosforilació oxidativa)$_a$ que ***tenen lloc al*** (mitocondri)$_b$. [Locativa]

nn. S'han de tenir en compte (les recombinacions)$_a$ que poden ***tenir lloc entre*** (el gen mutat)$_b$ i (els marcadors analitzats)$_c$. [Causalitat]

Finalment, volem indicar que el marcador que era molt present en el corpus de cardiopaties *veure('s en)* en el corpus que ara analitzem no té cap ocurrència on efectivament s'indiqui una localització i, per tant, atès l'elevat grau de soroll que proporciona, serà un marcador deixat de banda per a la proposta de sistema de detecció semiautomàtica.

La figura següent presenta de manera succinta la informació sobre aquest subtipus de relació conceptual que acabem de detallar.

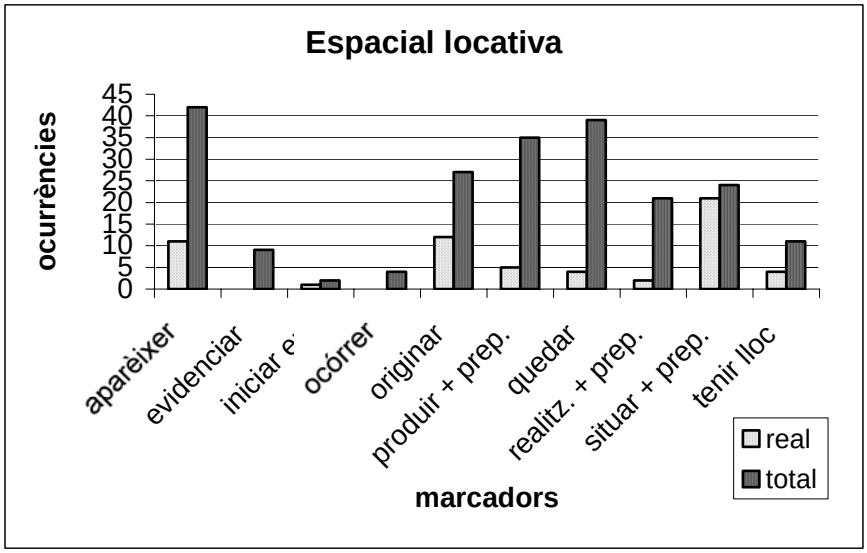


Figura 4-5. Gràfic dels marcadors de seqüencialitat espacial locativa.

4.3.3.2 Relació de seqüencialitat espacial de direcció

Els marcadors que en un primer moment hem cregut que vehiculaven la relació seqüencial espacial de direcció i que, com veurem, hauran de ser reduïts per la seva alta dependència al primer corpus que havíem analitzat però que no poden ser generalitzats com a marcadors universals d'aquest tipus de relació són:

a) *Allunyar*

b) *Apropar*

c) *Arribar*

d) *Continuar*

e) *Mesurar*

f) *Propagar*

D'aquests, hem de descartar de la llista alguns marcadors que han donat zero contextos vàlids i, per tant, no es poden considerar representatius. Aquestes unitats són: *apropar*, *mesurar* i *propagar*.

Dels 3 marcadors que resten, i per ordre alfabètic, hem de comentar que *allunyar* presenta 8 ocurrències, 2 de les quals indiquen direcció quan apareix una preposició.

Tanmateix, 4 ocurrències presenten *allunyar* com a modificador del concepte *a* o *b* i, per tant, són informacions sobre el concepte però no expressió d'una relació conceptual. De tota manera, tot i que només disposem d'un 25% d'encert, creiem que per la semàntica mateixa del marcador, aquest indica una direcció i caldrà veure com reduir el soroll en els casos en que la unitat actua de modificadora d'una altra unitat de coneixement especialitzat. Vegem un exemple de relació seqüencial espacial de direcció:

oo. En les successives generacions, (aquestes fluctuacions)$_a$ es poden anar acumulant***, allunyant-se de*** (l'estat inicial)$_b$.

Pel que fa al marcador *arribar*, val a dir que es tracta d'una unitat força freqüent en el corpus. Així, n'hem trobat 53 ocurrències, 16 de les quals (30,1%) indiquen direcció. En la resta de casos, o bé el concepte *a* i *b* no són unitats de coneixement especialitzades o bé apareix una oració negativa on, aleshores, es bloqueja la possibilitat de materialització de la relació seqüencial espacial de direcció. Un parell d'exemples de direcció són:

pp. (El trencament entre el món antic i medieval)$_a$ ***arriba al*** (Magreb)$_b$ amb les primeres expedicions musulmanes (l'any 643), inicialment confinades a Egipte.

qq. Sabem que (les girafes tenen el coll llarg)$_a$ *per poder* ***arribar a*** (la part alta dels arbres)$_b$ i menjar-se'n les fulles.

Finalment, *continuar* ens dóna un percentatge molt baix d'eficiència (només un 12%) en la transmissió de la relació de direcció. Tanmateix, creiem que la semàntica dels conceptes *a* i *b* ens servirà per reduir aquesta alta taxa de soroll en el capítol següent. De moment, observem un exemple en què es materialitza la relació de direcció:

rr. Quan una comunitat creix i s'expandeix a noves regions, els grups que la conformen es van separant de la seva regió d'origen i s'estableixen en llocs nous, dels quals sorgeixen (altres grups)$_a$, que de nou en expandir-se, ***continuen*** (el seu camí cap a llocs més distants)$_b$.

L'esquema d'aparició dels marcadors de la relació seqüencial espacial de direcció és el següent:

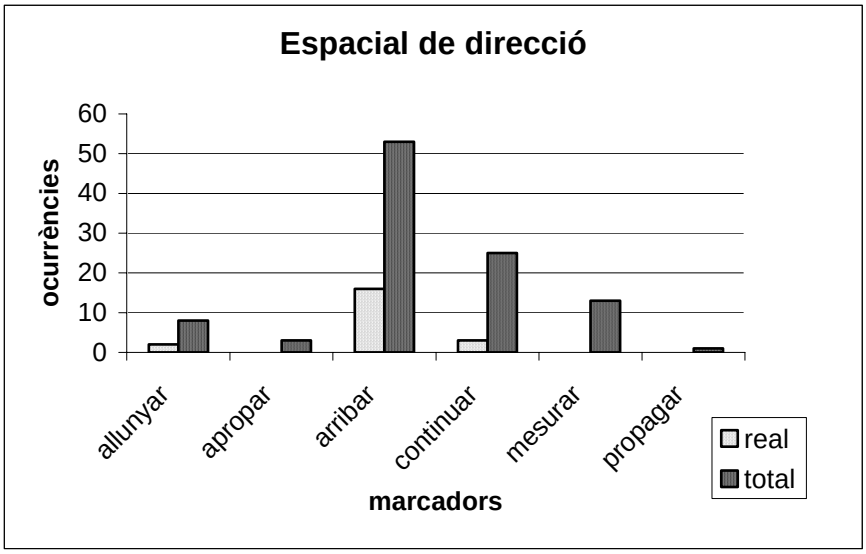


Figura 4-6. Gràfic dels marcadors de seqüencialitat espacial de direcció.

4.3.3.3 Relació de seqüencialitat temporal de simultaneïtat

Els marcadors lingüístics verbals seleccionats per a l'anàlisi d'aquesta relació són:

a) *Manifestar*

b) *Mostrar*

c) *Presentar*

Hem de dir que es tracta d'una relació força difícil de detectar només amb unitats verbals. Hem cercat locucions del tipus *alhora*, *al mateix temps* i *simultàniament* i tot i que el nombre d'ocurrències és menor, el soroll es redueix considerablement atès que contribueixen plenament a la semàntica del marcador. Tanmateix, la nostra anàlisi se centra en marcadors verbals i l'anàlisi d'unitats de tipus no verbals queda per a futures recerques. Així, per als marcadors llistats, hem trobat que, *manifestar* esdevé una unitat amb un significat concret en medicina i genètica diferent del d'altres àrees temàtiques. Per aquest motiu, podem afirmar que es tracta d'un marcador que expressa una relació associativa especialitzada que caldrà analitzar amb més detall al capítol següent. Vegem l'únic cas de simultaneïtat (gràcies a

l'adverbi *simultàniament*) i associació especialitzada alhora, i un exemple de relació associativa especialitzada:

ss. Altres vegades, (els dos al·lels)$_a$ ***manifesten simultàniament*** (la seva informació)$_b$. [Simultaneïtat/Associació especialitzada]

tt. En tots els exemples que hem discutit fins ara hem suposat que (dos individus que tinguin el mateix genotip)$_a$ ***manifestaran*** (el mateix fenotip)$_b$, és a dir, tindran la mateixa aparença. [Associació especialitzada]

El verb *mostrar* apareix en 100 ocasions en el corpus d'anàlisi. D'aquestes, s'estableix efectivament una relació conceptual en 68 casos. En els 32 restants o bé no apareix el concepte *a* o *b*, o bé es produeix una negació i, per tant, la relació queda invalidada. Dels 68 casos vehiculadors de relació conceptual, 13 ens donen informació metaexplicativa sobre un determinat concepte. Aquest fenomen, força usual en aquest tipus de textos, ens deixa amb 55 casos (un 80,88%) en què es dóna una relació conceptual que, lluny de ser una indicació de direcció, es vehicula com una relació associativa general que tractarem en el capítol següent. Vegem-ne un parell d'exemples:

uu. (Aquests ratolins transgènics)$_a$ ***mostraven*** (mort prematura)$_b$, sovint al voltant dels cent dies d'edat, (dèficits importants en tasques d'aprenentatge espacial)$_c$, (conducta neofòbica)$_d$ i (disminució del consum de glucosa en algunes àrees corticals)$_e$, encara que no s'observà la neuropatologia pròpia de la malaltia d'Alzheimer.

vv. Cal assenyalar també que (les distàncies genètiques calculades)$_a$ ***mostren*** (una clara separació entre les poblacions ibèriques i les poblacions nordafricanes)$_b$.

Finalment, el verb *presentar* tampoc no és un bon indicador de relació temporal de direcció. Ben al contrari, es tracta d'una unitat que apareix 131 vegades en el nostre corpus amb la següent distribució: 28 indicacions d'atributs d'un concepte; 12 indicacions de meronímia (9,16%) (quan el significat de la unitat és equiparable a

tenir) i 46 indicacions de relació associativa (35,11%). Caldrà veure en les estratègies de detecció com es pot diferenciar entre meronímia i associació atès que, com hem vist, no es tracta d'un marcador que expressi direcció.

Per a la relació temporal de simultaneïtat, l'esquema resultant de l'aplicació de les dades numèriques és el següent:

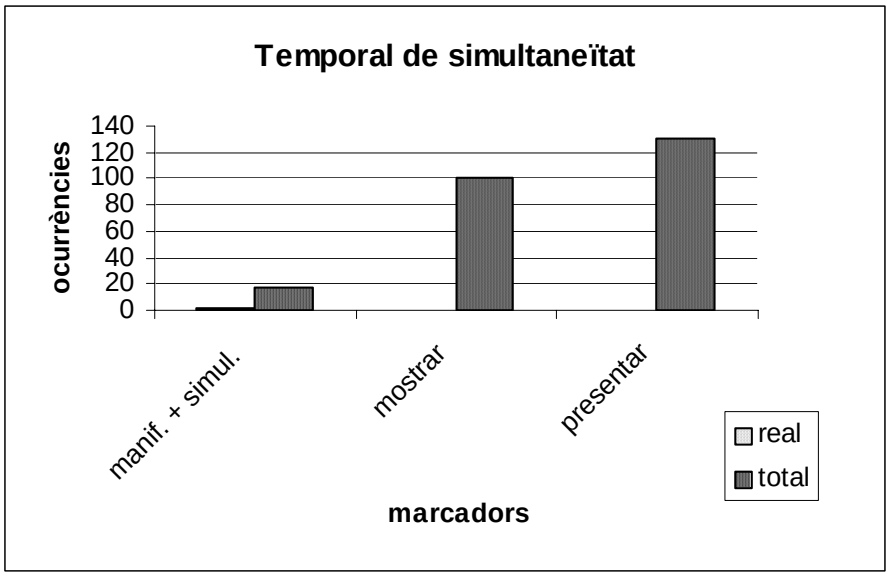


Figura 4-7. Gràfic dels marcadors de seqüencialitat temporal de simultaneïtat.

4.3.3.4 Relació de seqüencialitat temporal d'anterioritat-posterioritat

Els marcadors per a expressar aquesta relació temporal que hem triat a partir de les dades de l'exploració prèvia que hem presentat al capítol segon són:

a) *Correspondre*

b) *Localitzar*

c) *Seguir de / Ser seguit*

d) *Transcórrer*

Començarem indicant que el verb *transcórrer* no apareix en cap cas com a tal en els nostres contextos sinó que apareix en la seva forma no personal de participi modificant una unitat de coneixement especialitzat. Per tant, no expressa la relació temporal que ens ocupa en aquest subapartat.

Pel que fa al verb *correspondre*, disposem de 31 contextos en què actua com a MLRC però lluny d'indicar temporalitat, indica una associació de caràcter general entre un concepte *a* i un concepte *b* (en un 74,2% dels casos). La resta indica una negació o bé el verb *correspondre* apareix sota la forma discursiva *segons correspongui*. Fixem-nos en els dos exemples següents:

ww. (Les primeres evidències d'ocupació humana a la zona)$_a$ daten fins a uns 700.000 anys i ***corresponen a*** (restes classificades com a Homo erectus)$_b$, el qual probablement travessà la barrera sahariana per la vall del Nil.

xx. En aquest codi, (cada tres nucleòtids (un triplet))$_a$ ***es corresponen amb*** (un aminoàcid específic de la proteïna en qüestió)$_b$.

Si passem ara al marcador *localitzar*, vegem que aquest té una freqüència d'aparició força alta (39 ocurrències). En 27 dels casos (xifra que representa gairebé un 70%) expressa la relació seqüencial locativa (i per tant, caldrà analitzar-lo a partir d'ara com a vehiculador d'aquesta relació). En la resta de casos (el 30,76%), es produeix un ús específic del verb *localitzar* que té un significat propi en genètica i que caldrà tenir en compte en el grup de relacions associatives especialitzades que detallarem en el capítol següent. Vegem un exemple de cadascun dels usos que acabem de descriure:

yy. Un cop a l'interior, (el DNA)$_a$ ***es localitza dins*** (els endosomes)$_b$, on la presència de les molècules catiòniques dificulta la seva degradació. [Locativa]

zz. El primer pas per ***localitzar*** (un gen)$_b$ és conèixer el cromosoma on es troba. [Associativa]

Finalment, volem indicar l'exemple paradigmàtic de marcador de relació seqüencial temporal d'anterioritat-posterioritat. Es tracta del marcador *(ser) seguir de* que apareix en 9 ocasions i en el 100% dels casos es presenta com a indicador d'anterioritat–posterioritat. Vegem-ne dos fragments a tall d'exemple:

aaa. Es realitzà (una electroforesi en agarosa)$_a$ que ***fou seguida d'*** (una transferència dels fragments digerits i separats a un filtre per Southern blot)$_b$.

bbb. Aquesta tècnica consisteix en (l'amplificació d'un fragment d'un locus HLA)$_a$ ***seguida d'***(una fixació en una membrana)$_b$ i (la posterior detecció amb diferents sondes marcades)$_c$.

El quadre següent mostra com *(ser) seguit de* és l'únic marcador que té una certa rellevància en el marc de la relació seqüencial temporal d'anterioritat-posterioritat:

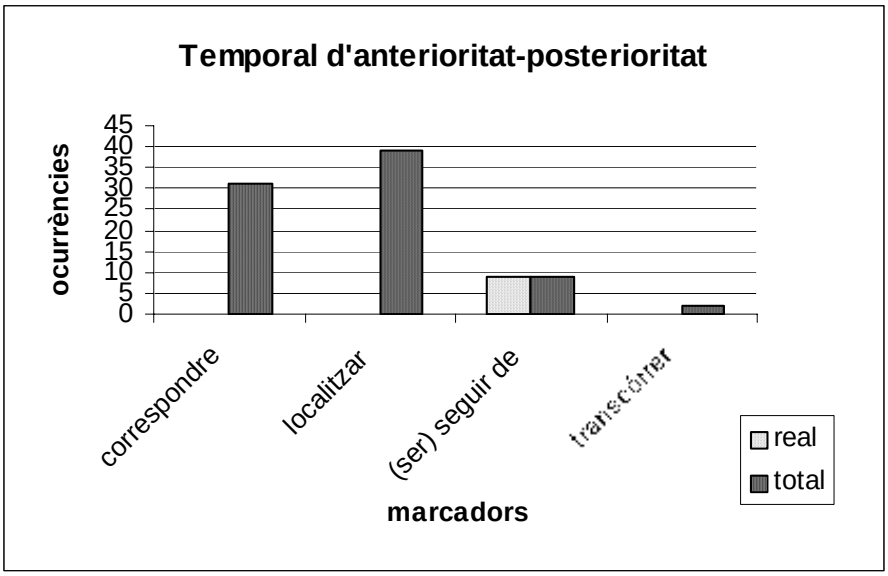


Figura 4-8. Gràfic dels marcadors de seqüencialitat temporal d'anterioritat-posterioritat.

4.3.4 Relació de causalitat

Pel que fa al quart gran tipus de relació, la relació de causalitat, hem iniciat la nostra cerca sobre el corpus a partir dels marcadors següents:

a) *Aparèixer*

b) *Causar / ser la causa de*

c) *Contribuir*

d) *Dependre de*

e) *Deure a*

f) *Donar lloc a*

g) *Implicar*

h) *Produir*

i) *Provocar*

j) *Reforçar*

k) *Trobar*

D'aquests marcadors, hem de dir que els dos últims no expressen causalitat. En el cas de *reforçar*, trobem només 2 ocurrències i, en cap cas, no s'indica la noció de causalitat. Per aquest motiu, creiem que es tracta d'un marcador no rellevant per a l'anàlisi de les dades. Pel que fa a *trobar*, hem aïllat 201 contextos en què apareix aquesta unitat. En cap cas no expressa la noció de causalitat i només podem mantenir aquesta unitat com a MLRC quan es materialitza sota l'estructura *a* es troba a/en *b*. En aquest cas, la relació conceptual que s'expressa és la relació seqüencial espacial locativa que apareix en una quarta part de les ocurrències (56 contextos que configuren un 27,86% dels casos). Vegem un exemple clar:

ccc. Per tant, la primera cosa que podem dir és que (el material hereditari)$_a$ ***es troba a*** (l'interior de les cèl·lules)$_b$.

Passem, ara sí, als marcadors que, en major o menor grau, indiquen causalitat. Seguint l'ordre alfabètic amb què els hem presentat analitzarem, en primer lloc, la unitat verbal *aparèixer*. Es tracta d'una unitat que hem recollit 42 vegades. D'aquestes, només 4 (9,5%) expressen causalitat mentre que 10 expressen una relació espacial locativa (23,8%) i 9 expressen temporalitat (21,42%). La noció de relació que expressa depèn de l'estructura sintacticosemàntica que manifesta cada context. Així, quan tenim una relació de causalitat, tenim una condició que matisa que quan passa *a*, apareix *b*. Per a indicar localització, s'utilitza la preposició *a* o *en* seguida d'un concepte la categoria semàntica del qual és locatiu o entitat i, finalment,

per a indicar la temporalitat, s'utilitza també la preposició *en* més una indicació de lloc o bé l'oració *aparèixer fa* x + indicació temps. Vegem un exemple de cada estructura:

ddd. (<Quan> l'adenoma creix més d'un centímetre de diàmetre)$_a$ ***apareixen*** (mutacions en K-ras)$_b$ i, posteriorment, s'observen delecions en diferents cromosomes, generalment el 18 i el 17, que inclouen el gen DCC i xxx, respectivament. [Causalitat]

eee. L'emergència i reemergència de zoonosi és un tema complex, perquè abraça (agents zoonòtics coneguts)$_a$ que ***apareixen en*** (llocs prèviament no observats)$_b$, i agents que desapareixen i tornen a aparèixer més tard i més lluny, en forma epidèmica.

fff. Finalment, ***fa*** (600 milions d'anys)$_b$ ***van aparèixer*** (els primers organismes pluricel·lulars)$_a$.

El marcador menys polisèmic i que comporta menys soroll dins d'aquesta relació de causalitat és *causar*, i també *ser la causa de* tot i que aquest apareix amb menys freqüència. De les 52 ocurrències de *causar*, 47 expressen causalitat. I de les 4 de *ser la causa de*, 3 també indiquen aquesta noció. Els 5 casos de soroll (9,6%) es deuen al fet que la unitat està mal etiquetada o bé apareix en una frase negativa o interrogativa o, encara, que un dels dos conceptes és indeterminat (per exemple, *això*). Vegem tres exemples clars de causalitat:

ggg.(La majoria de malalties neurològiques)$_a$ ***estan causades per*** (mutacions puntuals)$_b$ o ***per*** (delecions o duplicacions tan petites que la citogenètica no pot evidenciar l'anomalia)$_c$.

hhh.(Les mutacions)$_b$ ***poden ser causades per*** (factors intrínsecs als organismes i als seu funcionament)$_a$ i (per agents externs)$_{a'}$.

iii. Avui, (la sida)$_a$ ***és la causa principal de*** (mort en adults en una ciutat com Abidjan)$_b$.

El marcador *contribuir* no dóna uns resultats gaire precisos (només 2 relacions causals en 9 contextos d'aparició) perquè en la majoria de casos els conceptes *a* i *b* no són unitats de coneixement especialitzat mínimament lexicalitzades sinó que són oracions senceres. Caldrà veure en el capítol següent si existeix algun mecanisme per aïllar els contextos adequats i, en cas contrari, caldrà plantejar la no vitalitat d'aquests marcador. Vegem-ne, tanmateix, un exemple:

jjj. (El neolític)$_a$ entrà al nord d'Àfrica des de l'est, on sens dubte, ***contribuí al*** (sorgiment del Regne d'Egipte)$_b$, i s'estengué, lentament, al llarg de la costa mediterrània cap al Magreb.

En el cas de *dependre*, tenim 25 ocurrències 16 de les quals (64%) ens indiquen causalitat tot i que la detecció de l'expressió de la causa no és tan clara com amb el marcador *causar*. Més aviat es tracta d'una indicació del factor que fa que pugui passar *b* o que aquest *b* resulti d'una manera o d'una altra. Tanmateix, mantenim que expressa una relació de causalitat i en els 7 casos en què no és així, es deu al fet que apareix sota l'estructura *depenent de x, a* MLRC *b*, casos per als quals caldrà proposar un mecanisme de bloqueig en el capítol següent. Una mostra de la relació de causalitat és:

kkk. (La variabilitat clínica abans esmentada)$_b$ ***depèn de*** (el nombre de triplets)$_a$.

Quant al marcador lingüístic *deure('s) a*, tenim 49 contextos de 76 (gairebé un 65%) on s'expressa causalitat, seguint l'estructura *a* és degut/es deu a *b*. Els casos en què no funciona els trobem quan *deure a* apareix com a locució inicial del tipus *degut a x, a* MLRC *b*, en que es materialitza com una condició per a un altre tipus de relació. Això ocorre en 15 dels 76 contextos analitzats. Finalment, la resta de casos no recollits responen al fet que es tracta d'oracions negatives o sense algun dels dos conceptes que haurien d'estar implicats en aquest lligam binari. Vegem dos exemples de causalitat:

lll. També hi ha (anomalies cromosòmiques)$_b$ ***degudes a*** (la manca d'un dels dos cromosomes homòlegs)$_a$.

mmm. (La degradació de la matriu extracel·lular)$_b$ ***és deguda a*** (un augment de l'activitat proteolítica)$_a$ com a conseqüència de l'expressió i secreció de proteases.

Passem a un marcadors que ens proporciona 100% de precisió. Es tracta de *donar lloc a*, detectat en 25 ocasions la totalitat de les quals expressen la relació de causalitat. Podem afirmar que es tracta d'un marcador perfecte per tal com no proporciona gens de soroll i totes les ocurrències expressen la noció de causalitat seguint l'esquema de base *a causa b*. Fixem-nos en dos exemples de l'ús d'aquest marcador que tindrà un paper essencial en el sistema de detecció semiautomàtica:

nnn.(Interaccions més curtes)$_a$ ***donen lloc a*** (la inactivació funcional de la cèl·lula)$_b$ (fenomen anomenat anèrgia).

ooo. (L'activació antigènica dels limfòcits T)$_a$ ***dóna lloc a*** (un augment molt ràpid de la síntesi proteica)$_b$ (set o deu vegades) en el període de tres a deu hores que segueixen a l'activació.

En relació al marcador *implicar*, aquest sembla tenir una doble funció. Dels 64 contextos de què disposem, 11 (17,18%) segueixen l'estructura causal *a* implica *b*. Ara bé, 36 d'aquest contextos es materialitzen en la forma *a* està implicat en *b* (56,25%). En el primer cas, observem l'expressió de la relació causal mentre que, en el segon cas, ens trobem davant d'una relació associativa de caràcter general que caldrà veure amb més detall més endavant. Fixem-nos en el que acabem d'afirmar a través dels dos exemples següents:

ppp. (Una desregulació d'aquest sistema)$_a$ ***implica*** (malaltia)$_b$, ja que la manca de reconeixement de substàncies estranyes implica la prevalència de la infecció, mentre que respostes contra el que és propi donen lloc a malalties autoimmunitàries. [Causalitat]

qqq. (Les proteïnes codificades per aquests gens)$_a$ ***estan implicades en*** (els processos de resposta immune)$_b$ i, per tant, han estat extensament estudiades per llurs repercussions clíniques. [Associativa]

El marcador *produir* esdevé una unitat d'aparició força elevada en el nostre corpus. Disposem per a l'anàlisi de 151 contextos d'aparició d'aquest marcador. En una tercera part dels contextos 56 de 151 (37%) indica causalitat. Ara bé, segons quina sigui la preposició que duu al darrere, també pot indicar seqüencialitat espacial locativa (11 ocurrències que representa un 7,2% de precisió), temporal (6 ocurrències que són un 3,9%) o una relació d'instrument funció (4 ocurrències que conformen un 2,6%). Ens limitem ara a exemplificar la relació de causalitat tot i que som conscients que en el capítol cinquè caldrà presentar una anàlisi del tipus de preposicions que pot requerir per expressar una determinada relació conceptual o una altra:

rrr. (Tota mutació)$_a$ ***produeix*** (un canvi en un caràcter hereditari)$_b$ (produeix un al·lel nou), canvi que pot produir una modificació morfològica que la selecció natural s'encarrega de seleccionar, triant la més apta en cada cas.

sss. (Les mucoses)$_a$, per la seva banda, ***produeixen*** (secrecions)$_b$ i (enzims hidrolítics)$_c$ i (proteolítics)$_d$.

Acabarem aquest repàs als marcadors de la relació de causalitat amb un indicador força precís: *provocar*. Disposem de 45 casos, 38 dels quals indiquen causalitat (gairebé un 85%). El 15% restant de soroll es deu principalment al fet que no disposem del concepte *a* o bé aquest és un element anafòric del tipus *això, fet que,* etc. Parem l'atenció en dos exemples que ens mostren la bona efectivitat d'aquest marcador com a vehiculador de la noció de causalitat:

ttt. (La mutació)$_a$ ***provoca*** (l'aparició de nous al·lels)$_b$.

uuu. D'altra banda, (els adenovirus de primera generació)$_a$ ***provoquen*** (una forta inflamació no específica)$_b$ i (una important resposta immunitària)$_c$.

L'esquema següent ens resumeix la informació numèrica aportada en aquest apartat sobre la relació de causalitat detectada en el corpus d'anàlisi:

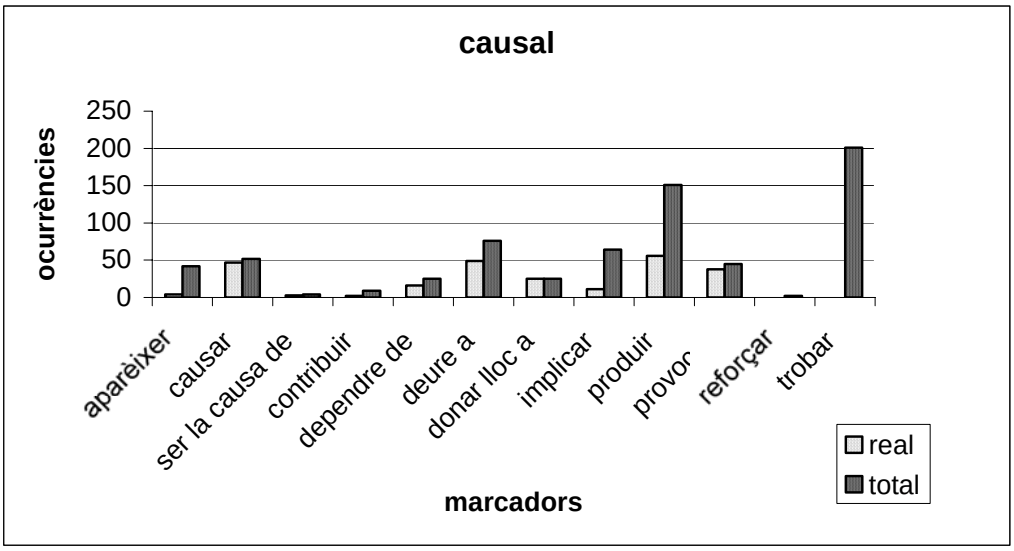


Figura 4-9. Gràfic dels marcadors de causalitat.

4.3.5 Relació instrumental

Els cinc marcadors verbals que hem triat i que seran objecte d'anàlisi com a vehiculadors o no de la relació instrumental són:

a) *Fer + preposició*

b) *Realitzar + preposició*

c) *Servir*

d) *Usar*

e) *Utilitzar*

De les 40 ocurrències del verb *fer* seguit de preposició, indiquen una relació instrumental 9 aparicions (22,5%) seguides de les preposicions *a partir de*, *amb*, *gràcies a* i *mitjançant*. La resta de casos no indiquen la relació d'intrument-funció sinó que un simple esdeveniment ocorre en el marc del context de frase, títol, ítem o cel·la. Vegem dos exemples de la relació d'instrument que caldrà refinar per tal d'eliminar una xifra tan gran de soroll:

vvv.(Una altra anàlisi de la filogènia de les seqüències)$_b$ ***es*** *pot* ***fer gràcies a*** (el càlcul de distàncies genètiques)$_a$ i (llur posterior representació en arbres filogenètics)$_{a'}$.

www. (L'amplificació de la regió I del mtDNA)$_b$ ***es feu mitjançant*** (la tècnica de la reacció en cadena de la polimerasa)$_a$ (Polymerase Chain Reaction, o PCR) en uns volums de reacció de 25 a 50 xl.

El mateix passa amb el marcador *realitzar*, el qual necessita les preposicions *a partir de*, *amb* i *mitjançant* com a elements de suport per expressar la relació instrumental. Només en 4 dels 31 (un 12,9%) casos de què disposem apareixen aquestes preposicions i, per tant, s'expressa efectivament la relació d'instrument-funció:

xxx.(L'anàlisi d'aquests polimorfismes)$_b$ ***es realitza mitjançant*** (anticossos comercials)$_a$ que aglutinaran els eritròcits segons posseeixin l'antigen corresponent.

En el cas de *servir*, la noció d'instrumentalitat se sol expressar amb el verb *fer* + *servir* en infinitiu. Disposem de 23 contextos, 11 dels quals (33,3%) expressen la relació conceptual d'instrument-funció. En aquest cas no és sempre l'aparició de la preposició la que indica l'instrumentalitat i la funció sinó el suport del verb *fer* a *servir*. Vegem-ne tres casos, un amb aquest element de suport, l'altre amb la preposició *per* + *infinitiu* i un tercer amb *servir* + *de*:

yyy.Diversitat de seqüències: vàrem fer (una estima)$_b$ ***fent servir*** (un paràmetre de diversitat similar al de diversitat nucleotídica)$_a$.

zzz. Així com (el DNA)$_a$ ***serveix per*** (emmagatzemar)$_b$ i (expressar la informació genètica)$_c$, les proteïnes són la clau de les funcions cel·lulars.

aaaa. És a dir, (les cadenes "velles")$_a$ ***serveixen de*** (motlle)$_b$ per fer les "noves", seguint la norma d'aparellament esmentada.

Tenim només 2 casos per al marcador verbal *usar*, un dels quals vehicula la relació d'instrumentalitat i l'altre no atès que manca el concepte *a*. Vegem el context en què apareix aquesta relació, conscients de la poca representativitat de la unitat:

bbbb. Aquest fet origina la primera dificultat en la interpretació dels resultats, perquè s'ha vist en diversos casos que les alteracions conductuals que s'observen en els animals depèn de quina sigui (la soca normal)$_a$ ***usada per a*** (els encreuaments)$_b$.

Acabem aquest apartat amb el marcador *utilitzar*, força més emprat que *usar* i que dóna millors resultats. En un 49,2% dels casos (62 sobre 126 ocurrències) s'expressa la relació d'instrumentalitat. En nombroses ocasions, l'estructura del context d'aparició de la unitat és *a* s'utilitza per a *b* mitjançant *c* o bé, per aconseguir *c*, *a* utilitza *b*. Amb això volem indicar que, aquesta relació dóna peu, sovint, a l'aparició d'una estructura terciària, i no binària, com és habitual. L'element *c* sol ser una oració en infinitiu o bé una nominalització d'aquest infinitiu. Vegem alguns casos del que acabem d'esmentar:

cccc. (Els oligonucleòtids)$_a$ ***utilitzats com a*** (sonda)$_b$ eren d'una llargada de 18 nuleòtids i els marcàrem amb una molècula de digoxigenina a l'extrem 3' utilitzant DIG Oligonucleotide 3'-End labeling Kit seguint les instruccions del proveïdor.

dddd. Posteriorment, estimàrem (una matriu de distàncies genètiques entre les poblacions)$_b$ ***utilitzant*** (la distància de pairwise)$_a$.

eeee. Més recentment, ***s'estan utilitzant*** (antagonistes)$_a$ (*per bloquejar* completament la resposta hormonal)$_b$.

ffff. (El procediment de split and mix)$_a$ ***s'utilitza per a*** (la síntesi de la primera quimioteca de pèptids soluble descrita pel grup d'Houghten)$_b$.

gggg. <Per a aconseguir> (la propagació del virus)$_b$ ***s'utilitzen*** (línies cel·lulars)$_a$ en les quals s'ha inserit la regió E1 i l'expressen; la més utilitzada és la línia de cèl·lules epitelials de ronyó.

Abans de tancar aquest apartat, fixem-nos amb l'esquema que resumeix l'índex d'aparició de cadascuna de les unitats analitzades:

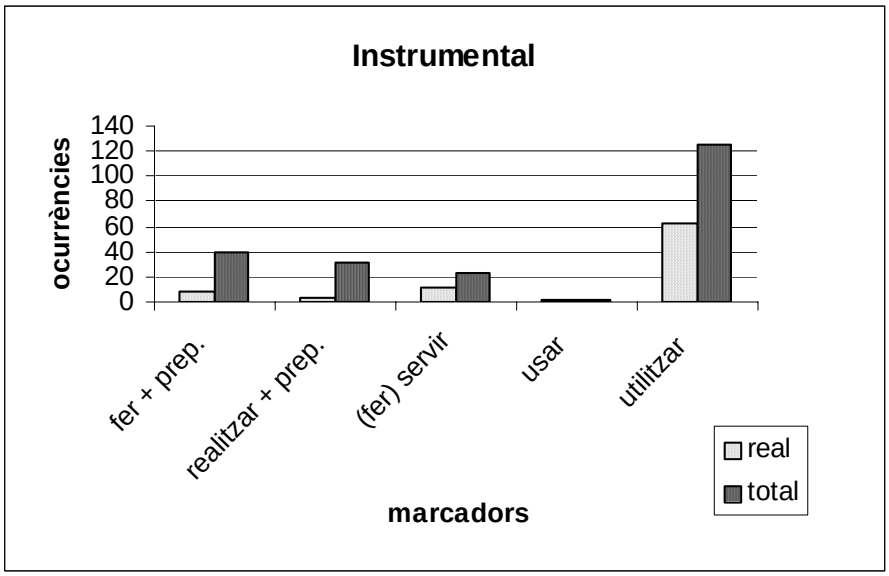


Figura 4-10. Gràfic dels marcadors d'instrumentalitat.

4.3.6 Relació de meronímia

Per a la relació de meronímia disposem d'un nombre força elevat de marcadors lingüístics. Les unitats objecte d'estudi en aquest cas són:

a) *Agrupar*

b) *Aplegar*

c) *Caracteritzar*

d) *Compondre*

e) *Constar de*

f) *Constituir*

g) *Definir*

h) *Englobar*

i) *Formar*

j) *Incloure*

k) *Integrar*

l) *Presentar*

m) *Reunir*

n) *Tenir*

D'aquests, indiquem inicialment que *aplegar*, *reunir* i *caracteritzar* no indiquen en cap dels casos una relació meronímica. Només disposem d'una ocurrència d'*aplegar*, que no vehicula la relació meronímica, i el mateix passa amb les quatre ocurrències de *reunir*. En canvi, *caracteritzar* apareix 43 vegades en el nostre corpus però els contextos que proporciona no expressen meronímia sinó que contenen una alta quantitat d'informació sobre les propietats d'una determinada unitat de coneixement especialitzat. És a dir, podrien reconvertir-se en parelles d'atribut-valor pròpies dels conceptes als quals estan caracteritzant.

Si resseguim ara cadascun dels marcadors lingüístics, veiem que *agrupar* té poca presència en el nostre corpus. Només apareix 4 vegades i tan sols en una ocasió indica meronímia. Es tracta, doncs, d'un marcador poc rellevant que caldrà considerar si es manté o no per al seu tractament semiautomàtic. L'exemple de què disposem és:

> hhhh. Però Joseph Greenberg (1987), que ja havia abordat abans la classificació lingüística d'Àfrica i de Nova Guinea amb la seva metodologia innovadora i controvertida, va proposar que (les llengües del continent americà)$_b$ ***es*** *podien* ***agrupar en*** (tres grans famílies)$_a$: esquimal-aleutiana, nadené (que comprèn trenta-quatre idiomes

parlats en dos territoris: el NW del Canadà i el SW dels EUA) i ameríndica, que engloba els 583 idiomes restants parlats pels nadius americans.

El marcador *compondre('s)* té una taxa d'aparició de 20 ocurrències en el nostre corpus. En gairebé la meitat dels casos, en 9 contextos que representen concretament un 45%, expressa una relació de meronímia de la qual podem observar els exemples següents:

iiii. El principi comú a aquestes tècniques es basa en generar, in vitro o in vivo, (col·leccions)$_a$ (biblioteques o també quimioteques) ***compostes per*** (un gran nombre de molècules)$_b$ (pèptids, oligonucleòtids, molècules orgàniques) o bé per (seqüències)$_c$ o (estructures aleatòries)$_d$, d'entre les quals una o diverses poden tenir les propietats biològiques desitjades.

jjjj. (Aquestes quimioteques)$_a$ ***estan compostes per*** (subquimioteques individuals)$_b$ en les quals una posició queda definida mentre que les altres s'ocupen amb mescles d'aminoàcids.

kkkk. Abans del neolític, (el paisatge lingüístic europeu)$_a$ *devia* ***estar compost per*** (moltes llengües d'àmbit reduït i parentiu llunyà)$_b$.

La totalitat de les ocurrències de *constar de* ens indiquen una relació meronímia. Així, en 4 contextos d'aparició d'aquest marcador es vehicula la relació de part-tot que és objecte d'anàlisi en aquest apartat (100%). Vegem-ne un parell d'exemples:

llll. (L'estructura gènica de tots els loci de classe I)$_a$ és força similar i ***consta d'***(uns nou exons separats per introns de diferent llargària)$_b$.

mmmm. (Els gens que codifiquen les cadenes X de classe II)$_a$ ***consten de*** (cinc exons)$_b$: el primer codifica el pèptid senyal, el segon i el tercer codifiquen els dominis 1 i 2, el quart exó codifica la regió transmembrana, la regió citoplasmàtica i part de la regió 3' no

traduïda i finalment, el cinquè exó codifica la regió no traduïda restant.

Una tercera part de les aparicions del marcador *constituir* expressa la relació meronímica. En aquest cas, 12 de les 36 ocurrències (33,3%) presenten aquesta noció que s'expressa mitjançant l'estructura *b* constitueix *a* o *a* està constituït per *b* en la majoria de casos. Vegem-ne dos exemples:

nnnn. La resta són (bacteris)$_b$, (fongs)$_c$, (protozous)$_d$, (cucs)$_e$ i fins i tot (paràsits)$_f$ ***que constitueixen*** (la nostra flora autòctona)$_a$ i viuen amb nosaltres en aparent equilibri.

oooo. (El seu genoma)$_a$ ***està constituït per*** (un DNA de doble cadena de 36kb)$_b$, amb una capacitat codificadors de 20 proteïnes, i que uneix covalentment a cada extrem una proteïna de 55kDa.

No obtenim tants bons resultats amb el marcador *definir*. Aquesta unitat apareix 51 vegades en el corpus però només en 7 ocasions expressa una relació meronímica (13,7%) que, a més, és força difícil de detectar semànticament. Vegem-ne un exemple.

pppp. Per contra, (els haplotips)$_a$ ***definits per*** (marcadors amb una major taxa de mutació)$_b$, com els microsatèl·lits, poden ser idèntics per descendència o per estat.

Ben al contrari, el marcador *englobar* proporciona uns resultats excel·lents. En el 100% de les seves ocurrències (12 contextos d'aparició) expressa una relació meronímica. Es tracta d'una dels marcadors que forneix millors resultats i que, sens dubte, serà repès en el disseny del sistema de detecció semiautomàtica de relacions conceptuals. A tall d'exemple, fixem-nos en els dos contextos següents:

qqqq. (L'últim període humit)$_a$ ***engloba*** (part del paleolític superior)$_b$ i (el neolític)$_c$ i es considera que finalitza al tercer mil·lenni abans de Crist.

rrrr. Cal assenyalar que (els haplotips H35, H36 i H38)$_a$ ***engloben*** (tots els cromosomes nordafricans que pertanyen a l'haplogrup 21)$_b$ i, en el seu conjunt, són equivalents a l'haplotip 4 descrit a Hammer et al. (1997; 1998).

El marcador *formar* és un dels que té una major presència en el nostre corpus d'anàlisi. Es tracta d'una unitat que apareix 149 vegades, 92 de les quals expressa la relació meronímica (64,78% d'encert). Aquesta expressió es vehicula per les estructures *b* forma *a, a* està format per *b* i *b* forma part d'*a*, principalment. Els casos en què es produeix soroll es deuen majoritàriament a una mala desambiguació del corpus (forma apareix com a verb quan en realitat és un nom) o bé als usos de la unitat *formar* com a sinònim de *crear*, aquests últims més difícils de solucionar. Vegem uns exemples de la relació meronímica expressada per aquest marcador lingüístic:

ssss. (La cadena pesant)$_a$ ***està formada per*** (un nucli polipeptídic de 40kD unit a un oligosacàrid a la regió extracel·lular)$_b$.

tttt. De fet, es considera que (els microsatèl·lits)$_b$ ***formen part de*** (l'anomenat DNA escombraria)$_a$, és a dir, sense funció coneguda.

uuuu. Això suposa determinar (la seqüència de 3000 milions de nucleòtids)$_a$ ***que formen*** (el nostre DNA)$_b$.

Analitzem, seguidament, el marcador *incloure* que apareix en el corpus un total de 69 vegades. En un 53,62% dels casos, una mica més de la meitat de les ocurrències, aquest marcador expressa meronímia (37 contextos). El soroll es produeix en algunes ocasions per un mal etiquetatge de la unitat ja que trobem nombroses ocurrències materialitzades en la forma adjectival *inclòs* i els seus derivats. En d'altres casos, observem que aquest marcador també expressa la relació d'inclusió de classe (14 contextos que representen un 20,28% dels casos). Per tant, caldrà proposar una estratègia de bloqueig per a aquests casos d'ús adjectival i recórrer a la semàntica de les unitat que acompanya per tal de reduir el soroll al mínim. Vegem, de moment,

dos exemples de relació meronímica i un d'inclusió de classe expressada pel marcador *incloure*:

vvvv. (La branca semítica al nord d'Àfrica)$_a$ ***inclou*** (diversos dialectes de l'àrab)$_b$. [Meronímia]

wwww. Cal assenyalar que (ambdós haplotips)$_a$ ***inclouen*** (cromosomes Y amb l'al·lel derivat per a la mutació 12f2)$_b$. [Meronímia]

xxxx. (Aquestes malalties)$_a$ es coneixen també com a (malalties complexes)$_{a'}$ i ***inclouen*** (la diabetis)$_b$, (la hipertensió)$_c$, (les malalties coronàries)$_d$, (l'asma)$_e$, (les malalties psiquiàtriques)$_f$ i (la tromboembòlica)$_g$, entre d'altres. [Inclusió]

De les 19 ocurrències d'*integrar*, hem detectat 11 casos (57,89%) en què es vehicula la relació meronímica. Hem de descartar els contextos que contenen una oració negativa, per tal com aleshores es bloqueja la possibilitat de l'aparició de la noció de meronímia i tots els casos en què la unitat verbal apareix en la seva forma pronominal *integrar-se*, on tampoc no es vehicula la noció de part-tot. Vegem algun cas de meronímia expressada per aquest marcador lingüístic verbal:

yyyy. Aquesta informació porta al (mapa)$_a$ que ***integra*** (la totalitat de la informació disponible d'un cromosoma determinat o del genoma)$_b$ incloent-hi seqüències codificadores, microsatèl·lits i mapes de clons contigus (contigs).

zzzz. Des del punt de vista del mapa físic ja s'ha aconseguit un (mapa total del genoma)$_a$ ***integrant*** (la informació del mapa genètic)$_b$, (el mapa d'híbrids de radiació)$_c$ i (el mapa de YACs)$_d$, amb un marcador STS cada 100 a 500 kb.

Només un 16,8% de les ocurrències del verb *presentar* expressen meronímia. Aquest percentatge representa, en dades absolutes, 22 dels 131 casos que hem analitzat. En canvi, hem detectat que aquest marcador indica en un gran nombre de casos, 87, una relació d'associació general (66,4%). En la resta d'ocasions, es tracta de contextos

mancats d'algun tipus d'informació, sigui el concepte *a* o el concepte *b*, o bé es tracta d'informació atributiva sobre un d'aquests dos conceptes, informació que haurà de ser reconduïda d'alguna manera cap a la base de dades terminològica en forma d'atributs i valors d'aquests atributs per a un concepte determinat. Vegem un parell d'exemples de meronímia i un d'associació:

aaaaa. (Les molècules de classe I i II del sistema HLA)$_a$ són uns dímers glicoproteics de membrana que ***presenten*** (quatre regions)$_b$: una regió extracel·lular aminoterminal a la qual s'uneix l'antigen, una regió extracel·lular d'estructura similar a les immunoglobulines, una regió transmembrana i una regió carboxi-terminal intracel·lular. [Meronímia]

bbbbb. En tenim un exemple en els grups sanguinis ABO, (el gen del qual)$_a$ ***presenta*** (tres al·lels (A, B i O))$_b$; la combinació de les parelles d'aquest al·lels forma els quatre grups sanguinis coneguts. [Meronímia]

ccccc. Des dels anys 40, hom sabia que (els bascos)$_a$ ***presentaven*** (la freqüència de Rh negatiu més alta del món)$_b$; també posseeixen baixes freqüències del grup B i altes del O. [Associació]

Arribem a l'últim marcador analitzat per a la relació meronímica, el verb *tenir*, considerat tradicionalment com el vehiculador prototípic d'aquest tipus de relació. Per al verb *tenir*, hem analitzat 371 contextos, 115 dels quals transmeten la noció de meronímia (31%). Els casos en què no es dóna aquesta relació es deuen principalment a una oració negativa, a relacions que semblen indicar un lligam entre dos elements però que és una associació però no una indicació de part i tot, i un nombre força elevat d'atribucions o especificacions sobre algun dels dos conceptes. Vegem, doncs, que ocorre un fenomen semblant al de *presentar* i que caldrà estudiar en paral·lel en el capítol següent. Per ara, fixem-nos en els dos exemples següents de meronímia:

ddddd. (El domini 3 de la cadena pesant)$_a$ ***té*** (uns 90 aminoàcids)$_b$ i se situa entre el domini 2 i la regió transmembrana.

eeeee. (Tot ésser viu)$_a$ ***té*** (el seu propi material hereditari)$_b$, des dels bacteris més senzills fins als humans, passant pels protozous, llevats, fangs, plantes i la resta d'animals.

Per a concloure aquest apartat, suggerim la taula següent com a resum del comportament dels 14 marcadors analitzats com a possibles vehiculadors de la relació de meronímia:

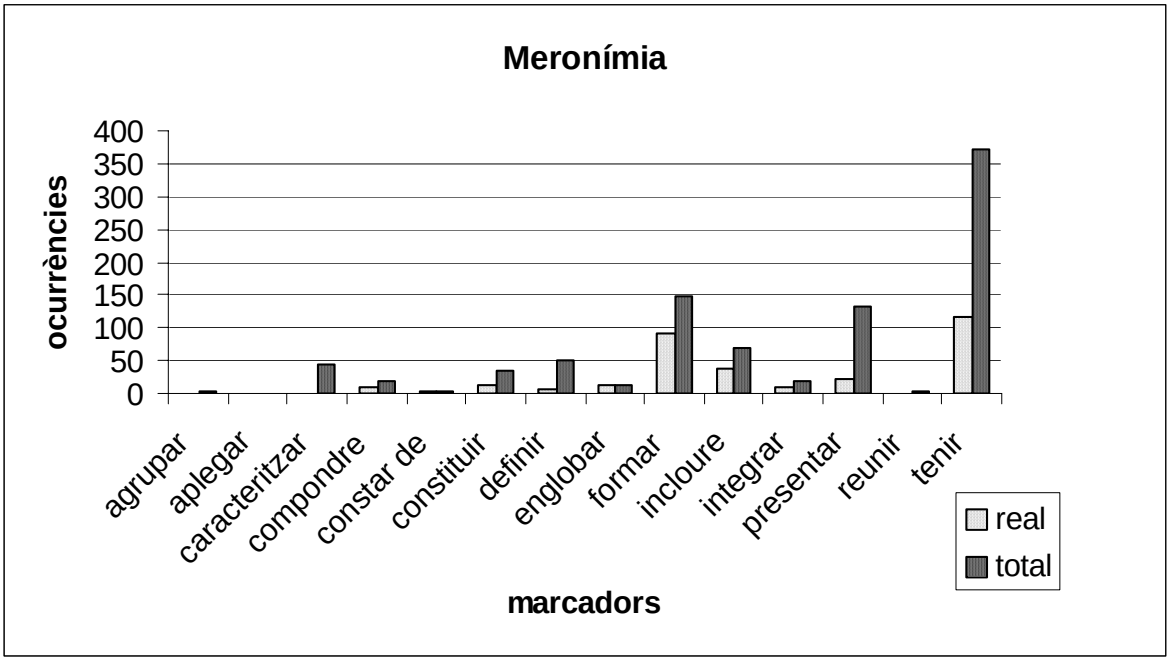


Figura 4-11. Gràfic dels marcadors de meronímia.

4.3.7 Relació d'associació

Per a l'última de les relacions conceptuals amb què treballem hem establert els següents marcadors lingüístics que caldrà validar ara en la primera anàlisi d'aquestes unitats:

a) *Confondre*

b) *Considerar*

c) *Correlacionar*

d) *Correspondre*

e) *Determinar*

f) *Indicar*

g) *Intervenir*

h) *Manifestar*

i) *Reflectir*

j) *Representar*

k) *Simular*

l) *Suggerir*

m) *Veure*

Començarem, com en les altres relacions, deixant de banda els marcadors que no expressen una relació associativa. Aquests són *suggerir*, amb 23 ocurrències en el corpus i *veure*, amb 5. Es tracta de dos verbs utilitzats per donar informació metaexplicativa, això és, per referir-se a graelles, taules i resultats però no indiquen una associació directa entre dos conceptes especialitzats.

Passem ara a la resta de marcadors, als quals haurem de sumar en la nostra síntesi final aquells marcadors d'altres relacions conceptuals que, en algunes ocasions, també indiquen associació, sigui general o especialitzada.

El marcador *confondre* apareix en 3 ocasions en el nostre corpus. Aquest marcador expressa una associació en la totalitat del casos (100%) tot i que, de rerefons, trobem que també vehicula un noció que equipara les característiques dels conceptes *a* i *b* i que, com a noció última, podria tenir una certa tendència a assimilar-se a la relació de semblança. Tanmateix, mantenim que es tracta d'un marcador associatiu tal com mostren els exemples següents:

fffff. Observen que (la papallona)$_a$ ***es confon amb*** (l'escorça del bedoll clar)$_b$, però destaca molt sobre la del bedoll ennegrit, i que passa a l'inrevés amb la papallona fosca.

ggggg. A Anglaterra, hi viu una (papallona nocturna de color clar (Biston betularia))$_a$ que de dia reposa sobre els bedolls, ja que gràcies al seu color ***es confon*** *fàcilment* ***amb*** (l'escorça d'aquests arbres)$_b$.

Pel que fa al marcador *considerar*, aquest apareix 69 vegades al nostre corpus. En un 23,18% dels casos, expressa una relació associativa (16 ocurrències). En canvi, aquesta relació no es manté en estructures del tipus *es considera* + *adjectiu* o quan *considerar* és sinònim de *es creu que*. Vegem un exemple de relació associativa:

hhhhh. I pel que sabem concorda també amb les dades lingüístiques: (el basc)$_a$ ***es considera*** (un isolat lingüístic)$_b$, sense cap parentiu clar amb cap altra llengua actual.

Tots dos exemples del marcador lingüístic *correlacionar* expressen la relació associativa. Es tracta, per tant, d'un marcador poc freqüent però amb un índex de precisió del 100%. Vegem-ne una mostra:

iiiii. La forma més severa, la DMD, és deguda a la manca de distrofina i la forma més lleu (BMD)$_a$ ***es correlaciona amb*** (la presència d'una distrofina anòmala)$_b$.

En el cas de *correspondre*, partim de 31 contextos en els quals trobem 23 casos de relació associativa (74,19%). Quan es materialitza una relació associativa, aquesta sol respondre a l'estructura *a* es correspon amb *b* i *a* correspon a *b*. Vegem-ne alguns casos clars:

jjjjj. (Les primeres evidències d'ocupació humana a la zona)$_a$ daten fins a uns 700.000 anys i ***corresponen a*** (restes classificades com a Homo erectus)$_b$, el qual probablement travessà la barrera sahariana per la vall del Nil.

kkkkk. En aquest codi, (cada tres nucleòtids (un triplet))$_a$ ***es corresponen amb*** (un aminoàcide específic de la proteïna en qüestió)$_b$.

Quant al marcador lingüístic *determinar*, trobem un nombre força elevat d'ocurrències, 181, de les quals hem etiquetat com a marcadores de relació associativa un 39,77% dels casos (72 en termes absoluts). Volem notar aquí que, en nombroses ocasions la relació d'associació no es manté perquè apareix una unitat que sembla tenir un cert caire de fraseològica especialitzada, ens referim a *determinar genèticament*, que sembla funcionar de manera travada i amb un significat específic en l'àmbit temàtic que estem analitzant. Vegem alguns contextos d'aparició de *determinar* que expressen associació general i alguns casos en què expressa la relació associativa especialitzada:

llllll. És (aquesta fracció)$_a$ la que ***determina*** (la diferència entre els individus)$_b$. [Assoaciació general]

mmmmm. (Les relacions entre haplogrups)$_a$ ***foren determinades mitjançant*** (un criteri de màxima parsimònia)$_b$, que és especialment fàcil d'aplicar en una regió no recombinant i a polimorfismes amb una baixa taxa de mutació. [Assoaciació general]

nnnnn. El cromosoma Y conté (els gens)$_a$ que ***determinen*** (la masculinitat)$_b$, s'hereta uniparentalment, per via paterna, i en la major part de la seva longitud no presenta recombinació. [Associació especialitzada]

ooooo. En l'exemple anterior, (un al·lel)$_a$ ***determinarà*** (pèsols de color verd)$_b$ i l'altre, pèsols de color groc. [Associació especialitzada]

El següent marcador objecte d'estudi es mostra com a una unitat polisèmica segons la seva estructura sintàctica. Quan *a* indica *b* tenim una relació associativa però en els casos en què trobem *a* indica on/el lloc *b.*.. ens trobem davant d'una relació seqüencial locativa que caldrà tenir en compte més endavant. I, també, trobem casos en què l'estructura que apareix és *a* indica amb/mitjançant *b* i, aleshores, ens trobem

davant d'una relació d'instrumentalitat. Aquí ens centrarem en l'exemplificació d'un tipus de cada tot i que, pel que fa a la relació associativa, trobem 17 contextos dels 69 totals que hem aïllat on s'expressa aquest tipus de relació (24,63%, és a dir, gairebé un 25% dels casos).

ppppp. El valor òptim es troba al voltant de 1,8, (un valor inferior)$_a$ ***indica*** (una contaminació per excés de proteïnes o restes de fenol)$_b$. [Associació]

qqqqq. (Esquema de la molècula del mtDNA)$_a$ ***on s'indica*** (la posició de la regió control amb els dos segments hipervariables)$_b$. [Localització]

rrrrr. (La posició de les quatre bases nucleotídiques (C, A, T i G))$_a$ ***s'indica amb*** (els ponts d'hidrogen que les mantenen unides)$_b$. [Instrumental]

El marcador *intervenir* té una presència discreta en el nostre corpus. Del total de 19 ocurrències, 12 expressen la relació d'associació (63,15%). Es tracta d'una unitat que produeix força bons resultats i de la qual volem mostrar alguns exemples on es veu clarament que, tot i que es tracta d'una relació associativa, hi ha una noció de localització de fons:

sssss. (El coneixement dels mecanismes)$_a$ que ***intervenen en*** (l'aprenentatge de la memòria)$_b$ és un objectiu prioritari en la comprensió del comportament.

ttttt. ***En*** aquest procés de còpia, anomenat, (replicació)$_b$ ***intervenen*** (diversos enzims)$_a$ que controlen el procés i tenen cura que la còpia s'hagi fe3t correctament, sense mutacions.

El següent marcador, *manifestar*, pot rebre un tractament equivalent al de *determinar* per tal com trobem alguns contextos en què apareix com a vehiculador d'una relació associativa general però, en d'altres casos, es realitza com una associació de caràcter especialitzat. Disposem de 17 casos, 12 dels quals presenten una associació (70,58%

de precisió). Vegem exemples de caràcter general i de caràcter especialitzat depenent del tipus semàntic de les unitats que vehicula aquest marcador lingüístic de relació conceptual:

uuuuu. El problema principal és que (animals amb manipulació genètica procedents de cèl·lules 129)$_a$ ***manifesten*** (alteracions conductuals molts semblants a les de la soca 129)$_b$, tot i que la soca receptora sigui una altra, la qual cosa evidencia la influència del context genètic en la interpretació de les dades conductuals. [Associació general]

vvvvv. Altres vegades, (els dos al·lels)$_a$ ***manifesten*** *simultàniament* (la seva informació)$_b$. [Associació especialitzada]

wwwww. En tots els exemples que hem discutit fins ara hem suposat que (dos individus que tinguin el mateix genotip)$_a$ ***manifestaran*** (el mateix fenotip)$_b$, és a dir, tindran la mateixa aparença. [Associació especialitzada]

El marcador *reflectir* apareix 7 vegades en el corpus d'anàlisi i, d'aquestes set ocurrències, 5 indiquen una relació associativa (71,4%). En alguns casos, aquesta unitat va acompanyada dels verbs de suport *quedar* i *veure's reflectit* per tal de vehicular la noció d'associació. Vegem-ne dos exemples clars:

xxxxx. (Aquests tipus de distribucions)$_a$ ***són capaces de reflectir*** (processos demogràfics del passat)$_b$.

yyyyy. (La complexitat de l'anàlisi del genoma humà)$_a$ ***queda reflectida*** en (l'enorme informació que conté)$_b$.

Passem ara al marcador lingüístic *representar*. Aquesta unitat està documentada en el corpus en 57 ocasions, 33 de les quals amb la noció de relació associativa i, en 5 casos i sempre quan li segueixen les preposicions *a*/*en*, indica la relació seqüencial espacial de localització. Per tant, en un 57,89% dels casos expressa associació i tan sols en un 8,7% dels casos, localització. Fixem-nos en un exemple de cada cas:

zzzzz. (Les seqüències emmarcades en negreta)$_a$ ***representen*** (l'encebador L15996)$_b$ i (el lloc d'unió de l'encebador H16401)$_c$, encebadors utilitzats en la seqüenciació del fragment analitzat. [Associació]

aaaaaa. ***En*** (els arbres genealògics (o genealogies))$_a$ ***es representen*** (tots els membres d'una família)$_b$. [Localització]

Acabarem amb el marcador *simular* que tan sols apareix en una ocasió en el corpus i que expressa la relació d'associació de manera clara. Aquest 100% de precisió s'ha de considerar amb prudència per tal com no disposem de contextos suficients per avaluar aquest marcador tal com hem fet en els altres casos. Caldria disposar d'un corpus més gran de documents i de caràcter tant general com especialitzat per poder-lo validar com a vehiculador de la relació associativa sense cap dubte. Tanmateix, creiem que val la pena exemplificar-lo aquí:

bbbbbb. Per veure amb claredat com canvia la informació en diversos tipus de mutacions, posarem un exemple d'(una frase formada per paraules de tres lletres)$_a$ que ***simulen*** (el codi genètic de triplets)$_b$.

La següent gràfica representa tota la informació que acabem de proporcionar sobre cadascun dels marcadors de manera succinta:

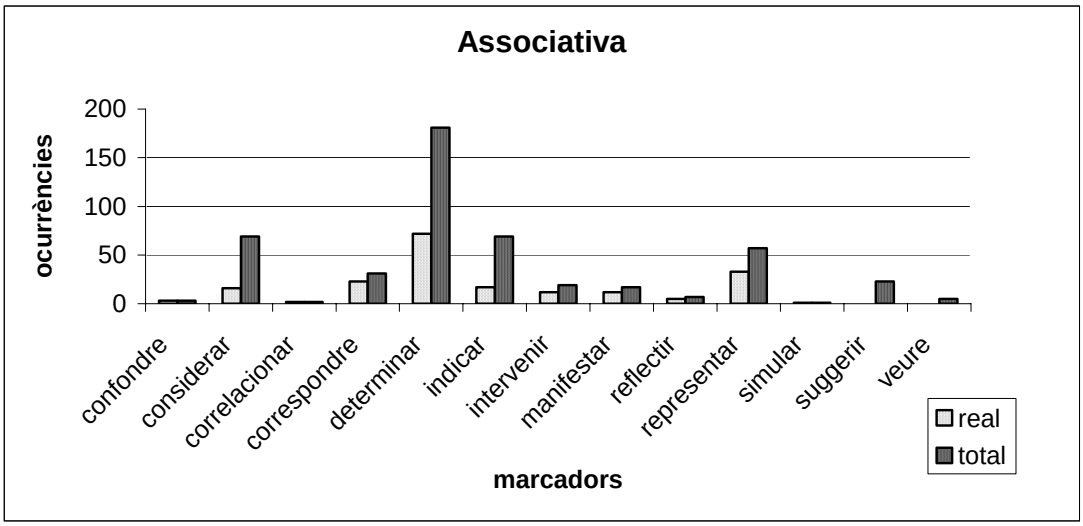


Figura 4-12. Gràfic dels marcadors d'associació.

4.4 A mode de síntesi

Seguidament, presentem una graella amb tota la informació que hem detallat anteriorment sobre els marcadors lingüístics verbals que expliciten cada tipus de relació conceptual. Ens proposem, amb aquesta taula, de recollir sintèticament tota la informació que hem descrit per tal d'organitzar la informació que haurà de ser reutilitzada en el capítol següent. La graella consta de la informació de relació conceptual en la columna de l'esquerra i dels percentatges de precisió de en fragments del 25% en la fila superior. A dins, cada marcador apareix col·locat en la cel·la corresponent segons el seu grau de precisió i la relació que expressa, tantes vegades com possibles relacions conceptuals vehiculi[4]. També fem una primera aproximació al patró sintàctic que estableix cadascun dels marcadors, sobretot pel que fa a la preposició que selecciona per tal de vehicular una o una altra relació conceptual. Avancem ja que, per al capítol d'estratègies de refinament i detecció de relacions conceptuals, ens centrarem en tots aquells marcadors que tinguin un índex de precisió del 25 al 100% tenint en compte, en els casos en què sigui necessari, els

[4] Indiquem en lletra cursiva els marcadors que apareixen com a unitats que vehiculen més d'una relació conceptual i amb un zero entre parèntesi els casos en què no hem trobat cap ocurrència del marcador.

altres tipus de relacions que expressen encara que en aquest cas la precisió sigui inferior al 25%.

Precisió dels marcadors per tipus de relació	0%-25%	25%-50%	50%-75%	75%-100%
SEMBLANÇA POSITIVA	-Això és (ser)	-Ser com *-Ser un*	-(ser) És a dir	-Assemblar-se a
SEMBLANÇA NEGATIVA	-Diferenciar (0) -Oposar (0) -Ser el contrari a (0)	*-Diferenciar(-se) de* *-Distingir*		*-Distingir(-se) de*
INCLUSIÓ	*-Distingir* *-Incloure*	*-Ser un (det.)*		
SEQÜENCIALITAT ESPACIAL LOCALITZACIÓ	*-Evidenciar* (0) *-Originar* -Produir a/en *-Produir durant* -Quedar -Indicar -Realitzar en -Veure's en (0) -Representar a/en	*-Aparèixer a/en* *-Iniciar en* -Tenir lloc a/en/dins -Trobar-se a/en	*-Produir(-se) en/a per/mitjançant-* *-Situar(-se) a/en* *-Localitzar*	
SEQÜENCIALITAT ESPACIAL DIRECCIÓ	-Apropar (0) -Mesurar (0) -Propagar (0) -Continuar	-Allunyar -Arribar		
SEQÜENCIALITAT TEMPORAL SIMULTANEÏTAT	*-Produir durant* *-Iniciar en* *-Aparèixer a/en* *-Originar* -Manifestar *-Situar a/en* *-Produir*			
SEQÜENCIALITAT TEMPORAL ANTERIORITAT-POSTERIORITAT	-Transcórrer	-Realitzar durant		-(Ser) seguit de
CAUSALITAT	-Reforçar -Trobar *-Aparèixer* -Contribuir *-Implicar*	*-Originar* -Tenir lloc ∅ / entre *-Produir*	*-Produir per/mitjançant* -Dependre de -Deure('s) a	-Causar / ser la causa de -Donar lloc a -Provocar
INSTRUMENTAL	-Produir -Fer a partir de / amb /gràcies a/mitjançant -Realitzar a partir de /amb/mitjançant -Indicar amb/mitjançant	-(Fer) Servir -Utilitzar	-Usar	
MERONÍMIA	-Mostrar (=tenir) -Aplegar -Reunir -Caracteritzar -Definir -Presentar (=tenir)	-Agrupar -Compondre -Constituir -Tenir	-Formar -Incloure -Integrar	-Constar de -Englobar

Precisió dels marcadors per tipus de relació	0%-25%	25%-50%	50%-75%	75%-100%
ASSOCIATIVA GENERAL	-Suggerir -Veure -Considerar	-*Determinar* -Indicar	-*Evidenciar* -Correspondre -Presentar -Intervenir -Representar -*Manifestar* -Reflectir	-Mostrar -Confondre(‘s) -Correlacionar(-se) -Simular
ASSOCIATIVA ESPECIALITZADA		-*Localitzar*	-*Diferenciar(-se) de* -*Manifestar* -*Determinar (genèticament)* -*Implicar* (en el patró *a* està implicat en *b*)	

Taula 4-1. Dades numèriques sobre la presència dels marcadors verbals.

Capítol 5

Estratègies de detecció de relacions conceptuals

Capítol V

5 Estratègies de detecció de relacions conceptuals

5.1 Introducció

En aquest capítol, ens proposem d'establir algunes estratègies de detecció de relacions conceptuals que ens portin a una posterior proposta de sistema de detecció semiautomàtica de relacions conceptuals a partir de l'anàlisi dels diferents fragments en què apareix una relació conceptual.

En el capítol anterior, hem presentat una anàlisi manual orientada a la detecció de les relacions conceptuals expressades mitjançant elements verbals a partir de la cerca en el corpus sobre el genoma humà que ja hem descrit anteriorment. El resultat de la cerca dels verbs en el corpus ens va donar un total de 3.992 contextos, els quals han estat objecte de l'anàlisi inicial que hem presentat en el capítol anterior.

El següent pas en el nostre procés de treball ha consistit en, a partir de la llista resultant de verbs que hem presentat en la taula que tanca el capítol anterior, modificar un programa de reconeixement d'unitats lèxiques especialitzades. La modificació d'aquest programa anomenat *Mercedes*[1] s'ha dut a terme amb l'objectiu

[1] Per a més informació sobre el funcionament d'aquest programa podeu consultar el manual d'ús intern VIVALDI, Jorge (2003a) *Sistema de reconocimiento de términos* Mercedes. *Manual de utilización*. Barcelona: Institut Universitari de Lingüística Aplicada.

que aquesta eina, construïda en el marc de l'Institut Universitari de Lingüística Aplicada, sigui capaç de detectar fragments textuals que contenen no només les unitats lèxiques especialitzades coincidents amb les que contenen els seus diccionaris de partida, sinó també les unitats verbals que inicialment indiquen o expressen una relació conceptual.

En aquest capítol, presentarem en primer lloc una breu descripció del funcionament del programa *Mercedes*. Seguidament, descriurem el disseny de la base de dades que conté els resultats de sortida d'aquest programa. A aquesta informació, que constitueix el nostre catàleg d'anàlisi per a una proposta d'estratègies de detecció de relacions conceptuals, hi hem afegit tots els aspectes que ens han portat a descartar o retenir una determinada unitat verbal com a marcador lingüístic explícit d'una o de diverses relacions conceptuals. A més, sobre la base d'aquest catàleg, hem indicat també si el marcador vehicula en un mateix context més d'un tipus de relació conceptual i, per tant, es tracta d'un marcador polisèmic.

De manera detallada, doncs, revisem quins són els aspectes que ens porten a descartar un determinat marcador perquè no expressa una relació conceptual i, d'altra banda, indiquem quins són els criteris que ens permeten retenir i aïllar una determinada relació conceptual. D'entre tots els casos en què es donen relacions conceptuals, aprofundim en el tractament dels marcadors polisèmics indicant alguns recursos que ens permeten desambiguar el tipus de relació conceptual vehiculada. Finalment, i atesa la diversitat de continguts que presentem en aquest capítol, creiem interessant d'introduir una síntesi dels elements més destacats que esdevindran estratègies per a la proposta de sistema de detecció semiautomàtica del capítol següent.

5.2 Un primer pas cap al reconeixement de relacions conceptuals

5.2.1 Eines i procés de treball

Per tal d'avançar en el reconeixement de les relacions conceptuals hem utilitzat l'eina *Mercedes*, que és un sistema de reconeixement d'unitats lèxiques especialitzades. Aquest sistema està compost per dos mòduls, un programa de reconeixement i un mòdul de diccionaris que inclou un diccionari de referència per a cada llengua i diversos glossaris electrònics de l'àmbit del genoma humà.

Pel que fa al seu funcionament, l'eina rep informació dels textos especialitzats que constitueixen el corpus, busca dins d'aquests textos les unitats lèxiques que estan incloses en el diccionari i, finalment, genera una llista d'unitats lèxiques especialitzades nominals que existeixen en el diccionari amb els seus respectius contextos d'aparició. En el moment actual de la recerca, el diccionari sobre el genoma humà consta de 6.000 entrades per a l'espanyol, 6.600 per a l'anglès i 850 per al català.

El següent pas ha consistit en introduir la llista de verbs resultants de l'anàlisi del capítol anterior en el programa per tal que el sistema detecti les unitats verbals que apareixen en els contextos on trobem alguna unitat lèxica especialitzada. Es tracta d'una opció del programa que permet obtenir una pàgina web addicional amb la indicació dels verbs sols que són potencials vehiculadors de relació conceptual[2]. La imatge següent mostra un exemple, amb les unitats lèxiques especialitzades en color vermell i les unitats verbals en color blau, dels resultats obtinguts amb l'ús d'aquest programa:

[2] Les dades resultants es troben disponibles i consultables en la seva totalitat en el CD-ROM que s'adjunta com a annex al final d'aquest volum.

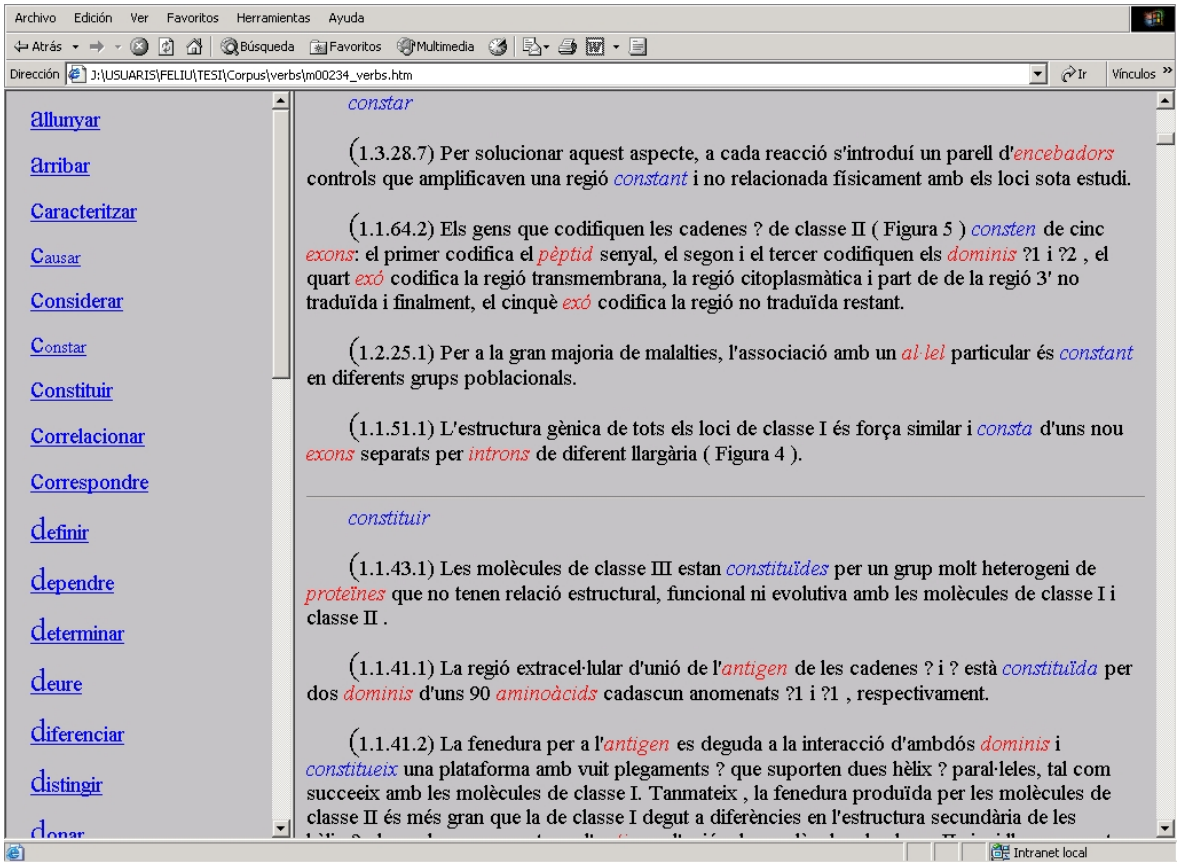


Figura 5-1. Mostra dels resultats obtinguts amb el programa *Mercedes*.

La llista de verbs que hem introduït al programa *Mercedes* prové dels resultats obtinguts en l'anàlisi del capítol 4, és a dir, que hem dut a terme un buidatge de la graella que presentem en l'apartat final del capítol anterior i el resultat ha estat una llista de 55 unitats verbals possibles vehiculadores de relació conceptual. El programa no admet, de moment, la introducció de la unitat verbal seguida del seu patró (principalment les preposicions) perquè compara cadenes de caràcters i si la preposició no apareix just darrere del verb trobem gran quantitat de silenci. A més, atès l'etiquetatge morfològic del Corpus Tècnic de l'IULA, tampoc no podem distingir entre la forma estàndard del verb i la seva forma pronominal[3]. Per aquests dos motius, els verbs introduïts individualment en aquesta funcionalitat del programa *Mercedes* són, per ordre alfabètic:

[3] Els resultats del programa *Mercedes* inclouen també les formes pronominals dels verbs.

allunyar, *aparèixer*, *apropar*, *arribar*, *caracteritzar*, *causar*, *compondre*, *considerar*, *constar*, *constituir*, *continuar*, *contribuir*, *correlacionar*, *correspondre*, *definir*, *dependre*, *determinar*, *deure*, *diferenciar*, *distingir*, *donar*, *englobar*, *evidenciar*, *fer*, *formar*, *implicar*, *incloure*, *indicar*, *inciar*, *integrar*, *intervenir*, *localitzar*, *manifestar*, *mesurar*, *mostrar*, *originar*, *presentar*, *produir*, *propagar*, *provocar*, *quedar*, *realitzar*, *reflectir*, *representar*, *reunir*, *ser*, *simular*, *situar*, *suggerir*, *tenir*, *transcórrer*, *trobar*, *usar*, *utilitzar*, *veure*.

Com es pot observar, existeix una diferència evident entre les unitats verbals llistades a la graella final del capítol anterior i la llista que acabem de detallar. En aquest primer pas cap a l'establiment d'estratègies que ens permetin detectar semiautomàticament les relacions conceptuals aprofitant al màxim les eines de què disposem en el marc de l'IULA, que és on funcionarà el sistema de detecció de relacions conceptuals, hem hagut de cenyir-nos a introduir la llista d'elements verbals com a unitats individuals. És a dir, el programa ens ha permès dur a terme la funcionalitat que combina la detecció d'unitats terminològiques coincidents amb les dels diccionaris font i, a més, obtenir els contextos amb els verbs marcats, però no hem introduït les preposicions i altres locucions que acompanyen en alguns casos als verbs ja que això produeix, inevitablement, silenci.

Posem un exemple: hem afegit a la llista de verbs de partida la unitat verbal *fer* però no, *fer mitjançant*, *fer a partir de*, *fer amb*, *fer gràcies* —estructures verbals indicadores de la relació conceptual instrument-funció— perquè el sistema compara cadenes de caràcters i, fins al moment, les dades del corpus no estan etiquetades sintàcticament. Si imaginem un context com: *L'anàlisi contrastiva del genoma es fa instantàniament mitjançant electroforesi*, ens adonem ràpidament que aquest context només apareixerà en la llista de contextos proporcionats pel programa *Mercedes* amb la indicació del verb *fer* però no *mitjançant*, atès que hi ha un element textual intermedi que faria que el sistema no detectés la unitat de partida *fer mitjançant*. Aquest mode de funcionament afecta totes les unitats que contenen algun element adjacent al verb i, per aquest motiu, hem decidit, en una primera etapa, treballar a partir de la llista de verbs individuals. D'aquesta manera hem assegurat que el sistema ens respongui amb el màxim d'informació, atès que els contextos que tenen

preposicions o locucions preposicionals evidentment, també queden recollits gràcies a la unitat verbal principal.

5.2.2 Base de dades de marcadors de relacions conceptuals[4]

L'aplicació del programa *Mercedes* amb la utilitat de detecció d'unitats verbals sobre el corpus que ja hem analitzat en el capítol anterior d'aquest treball ens ha donat com a resultat una llista de 3.114 contextos en què apareix una unitat verbal possible indicadora de relació conceptual i com a mínim una unitat lèxica especialitzada coincident amb la llista de termes dels diccionaris de partida en català sobre els quals funciona aquesta eina.

La reducció del nombre de contextos que havíem analitzat en el capítol anterior, 3.992, a la xifra actual de què disposem, 3.114, es deu al fet que el programa detecta els contextos en què apareix la unitat verbal però també, i obligatòriament, ha de retenir com a mínim una unitat terminològica en el context que va de punt a punt[5]. Volem fer explícit que, en el nostre cas, aquest mètode de funcionament del programa no representa cap problema atès que la nostra definició de relació conceptual comporta, necessàriament, l'aparició de com a mínim un concepte, expressat evidentment per una unitat terminològica, a cada banda de la unitat verbal vehiculadora de relació conceptual. Per tant, considerem que si el context ha estat descartat per manca d'unitats lèxiques especialitzades no es tractava d'un fragment de coneixement especialitzat que fos un bon candidat a ésser retingut per a futures aplicacions posteriors.

[4] Aquesta base de dades es troba disponible i consultable en la seva totalitat en el CD-ROM que s'adjunta com a annex al final d'aquest volum. Per facilitar la consulta de la informació, també hem constituït i afegit una aplicació en format html que permet navegar per la informació continguda a la base de dades.

[5] Els lectors d'aquest treball podran apreciar que, en alguns casos, els contextos tenen una extensió superior a la indicada, és a dir, de punt a punt. Cal recordar que els contextos provenen directament del Corpus Tècnic de l'IULA i, de vegades, el procés de segmentació d'inici i final de frase ens proporciona contextos una mica més llargs per qüestions tècniques.

Les dades proporcionades pel programa *Mercedes* apareixen en pàgines web, tal com hem vist al principi d'aquest capítol. Per tal de poder avançar una mica més en les estratègies que ens permetin detectar les relacions conceptuals de manera assistida hem creat una base de dades[6] que conté tots els contextos resultants amb informacions essencials per a la detecció futura. Aquesta base de dades serveix per determinar la validesa dels contextos verbals i, per tant, tots aquests contextos han estat revisats i etiquetats de nou sobre la base dels paràmetres que detallem en els apartats que segueixen. Abans però, volem descriure l'estructura general d'aquesta base de dades que, a grans trets, conté la informació següent:

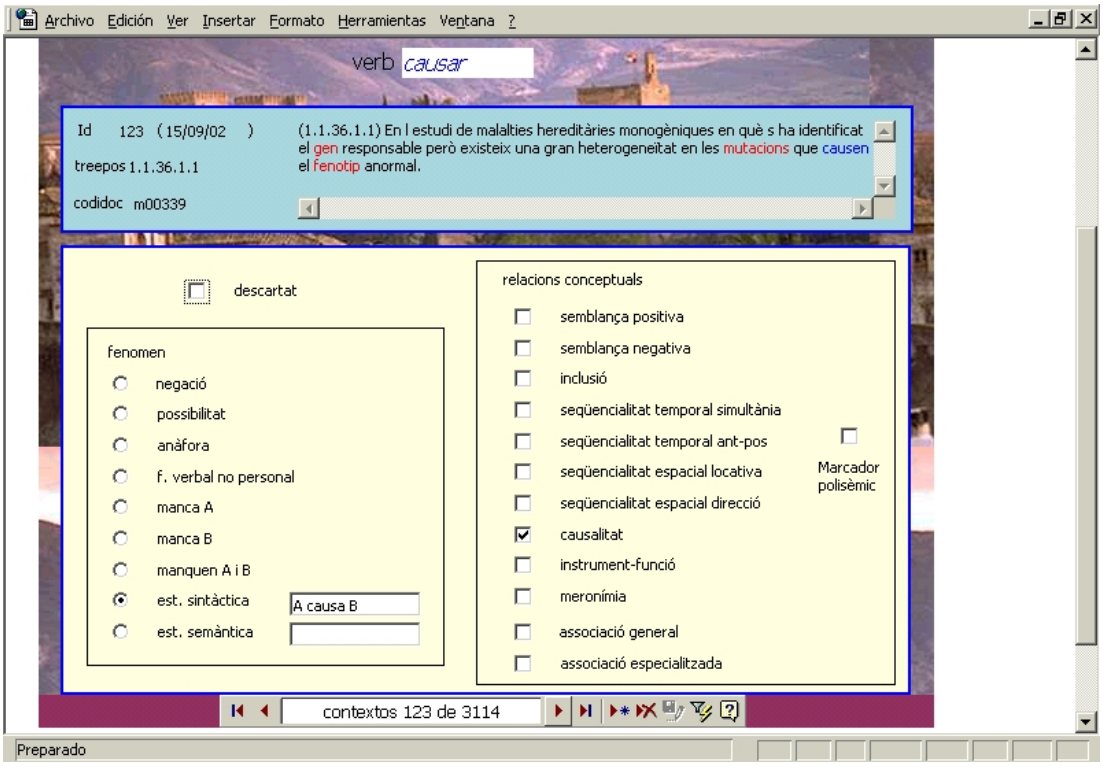


Figura 5-2. Estructura general de la base de dades de contextos verbals.

Aquesta base de dades conté tres parts diferenciades. La part superior amb la indicació del verb i el requadre sobre fons blau conté la informació que prové directament del traspàs automàtic de les dades extretes amb el programa *Mercedes*.

[6] La base de dades ha estat creada amb el programa Access 2000, de Microsoft, i només es pot consultar des d'aquesta versió del programa o superiors.

Després, el requadre groc central recull totes les indicacions que nosaltres hem col·locat manualment en aquest procés de revisió orientat a detectar estratègies per a la detecció i, finalment, a la barra d'eines inferior, se'ns indica en quin dels 3.114 contextos possibles vehiculadors de relació conceptual estem treballant, o revisant.

Així, trobem, en primera instància, una casella on apareix la unitat verbal, en el cas de la figura de mostra *causar*. Seguidament, el requadre blau recull al marge esquerre, i de dalt a baix, el número identificador del context i la data de creació del registre (*Id*), informació relativa al marcatge estructural del corpus que hem utilitzat que, com ja hem dit, constitueix una part del Corpus Tècnic de l'IULA (*treepos*) i, finament, el codi que té aquest document en el si del Corpus Tècnic (*codidoc*) que, en aquest cas concret, és el document de medicina[7] m00339. Encara en la franja de color blau, la base de dades recull el context en què apareix la unitat verbal indicada també en color blau (*causar*) i les unitats terminològiques coincidents amb les dels diccionaris de base del programa *Mercedes* indicades en color vermell (*gen*, *mutacions*, *fenotip*).

Si passem a la part central del registre, trobem dues zones diferenciades. A l'esquerra trobem, en primera instància, una casella per indicar els contextos que haguem descartat (*descartat*) i, a sota, la llista de fenòmens que ens permetran indicar per què hem descartat un determinat context. Aquest fenòmens els analitzarem amb més detall seguidament. A la dreta, i sota la indicació de *relacions conceptuals*, trobem la llista de les relacions conceptuals de què partim i amb què hem treballat al llarg de tot aquesta recerca. Com es pot observar, en el cas del context que analitzem la unitat verbal *causar* expressa una relació de *causalitat*. Per últim, la casella de la dreta ens permet indicar més clarament els casos en què trobem una unitat verbal que actua de marcador polisèmic, és a dir, que expressa més d'una relació conceptual, sigui en un mateix context o, com passa més sovint, a través de contextos diferents. Aquesta informació s'indica marcant tantes relacions conceptuals (normalment dues) com

[7] En el Corpus Tècnic de l'IULA els documents classificats temàticament com a textos sobre el genoma es troben dins de l'àrea temàtica major de medicina i, per aquest motiu, la lletra inicial que indica el document és una *m*.

sigui necessari i, a més, marcant la casella de *marcador polisèmic* d'un determinat registre.

La part inferior conté la informació numèrica sobre el total de registres i el número del registre que estem visualitzant (en aquest cas el 123 del total). Hem realitzat algunes consultes per tal d'ordenar les dades i, per aquest motiu, algunes de les possibilitats de consulta de la base de dades contenen, a més d'aquesta informació, l'ordenació del verb en relació a les 55 unitats verbals objecte d'estudi i el context que visualitzem del total de contextos en què apareix aquest possible marcador de relació conceptual.

5.2.2.1 Aspectes que ens porten a descartar el marcador verbal

A partir de l'observació de les dades del capítol anterior, i amb l'objectiu primordial de sistematitzar la informació de què disposem per poder establir el prototip de funcionament d'un sistema de detecció semiautomàtica, hem establert 7 paràmetres que permeten descartar una unitat verbal *a priori* vehiculadora de relació conceptual en un processament automàtic que només detecti les unitats verbals sense cap altre tipus d'informació addicional.

Alguns dels paràmetres que ens porten a descartar la unitat verbal es troben estretament relacionats entre ells, com indicarem oportunament, i es presenten per l'ordre en què creiem que resulten més efectius. Analitzem-los un a un.

NEGACIÓ

Es tracta de l'aparició d'una partícula negativa del tipus *no, ni, sense* que anul·la l'expressivitat semàntica de la unitat verbal pel que fa a la possible transmissió de relació conceptual. Vegem-ne alguns exemples a partir dels registres de la base de dades:

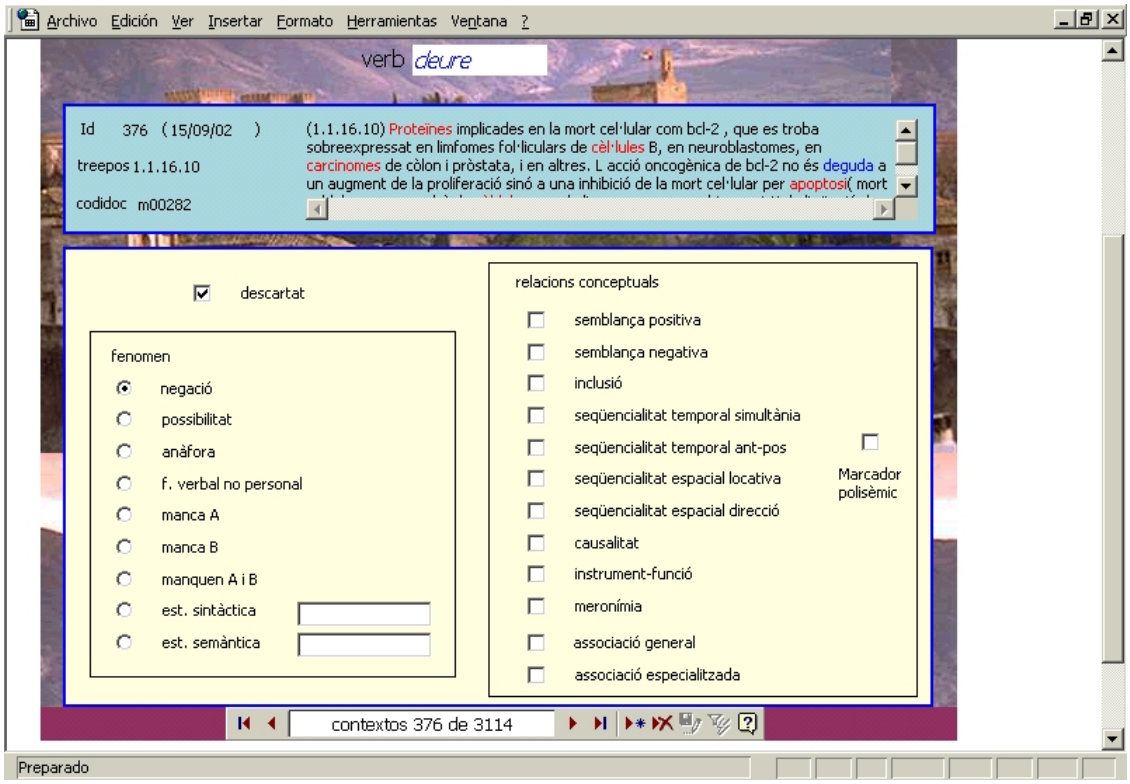


Figura 5-3. Exemple de negació.

En aquest cas, el coneixement especialitzat que es vehicula en el petit fragment on apareix la unitat verbal no recull una relació conceptual atès que la negació *no* de *l'acció oncogènica de bd-2 no és deguda a un augment de la proliferació* anul·la la possible relació conceptual de tipus causal, tot i que la resta del fragment sí que pot proporcionar informació conceptual rellevant.

El següent exemple segueix un patró semblant a l'anterior però, en aquest cas, la possible relació conceptual que no es vehicula és la relació meronímica, prototípicament expressada pel marcador *integrar*:

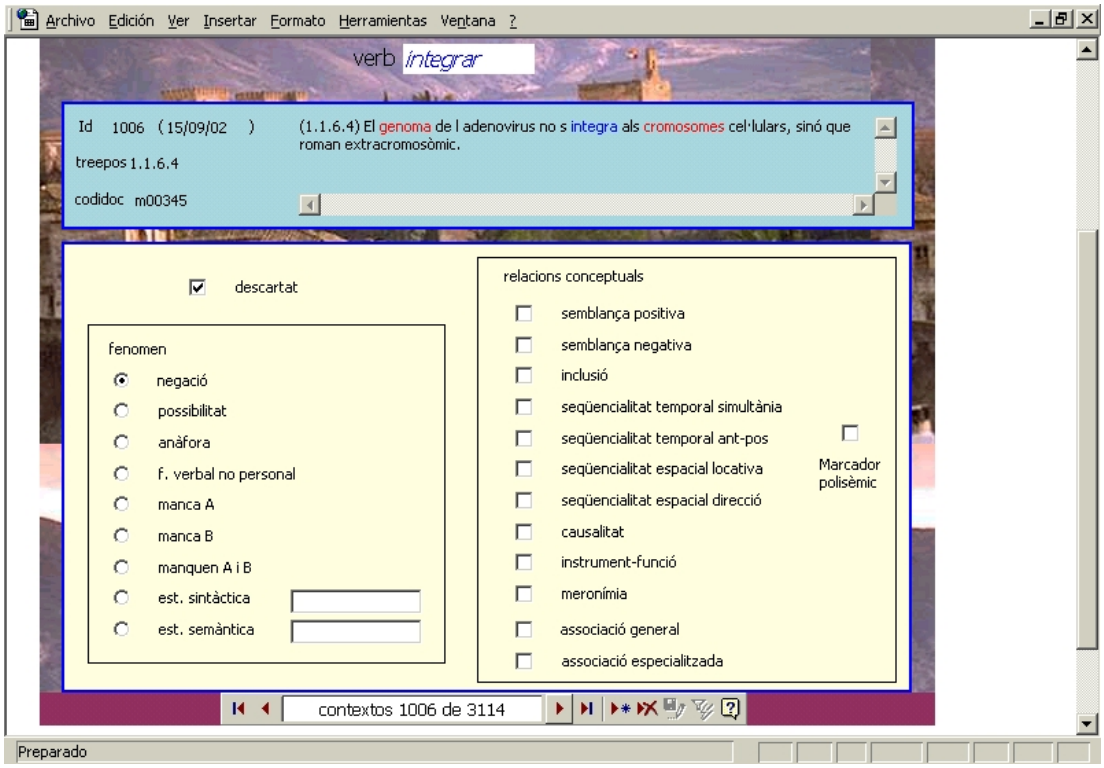


Figura 5-4. Exemple de negació.

POSSIBILITAT

De manera similar al paràmetre de la negació, considerem que no es dóna efectivament una relació conceptual quan apareix alguna marca que indica possibilitat, generalment expressada pel verb modal *poder* o per una forma verbal en el temps verbal del condicional. Així, creiem que una relació conceptual que es pugui aïllar, retenir i reutilitzar en futures aplicacions ha de donar-se en el text de manera afirmativa i, per aquest motiu, i en aquest primer intent de sistematització, deixem de banda contextos com els següent:

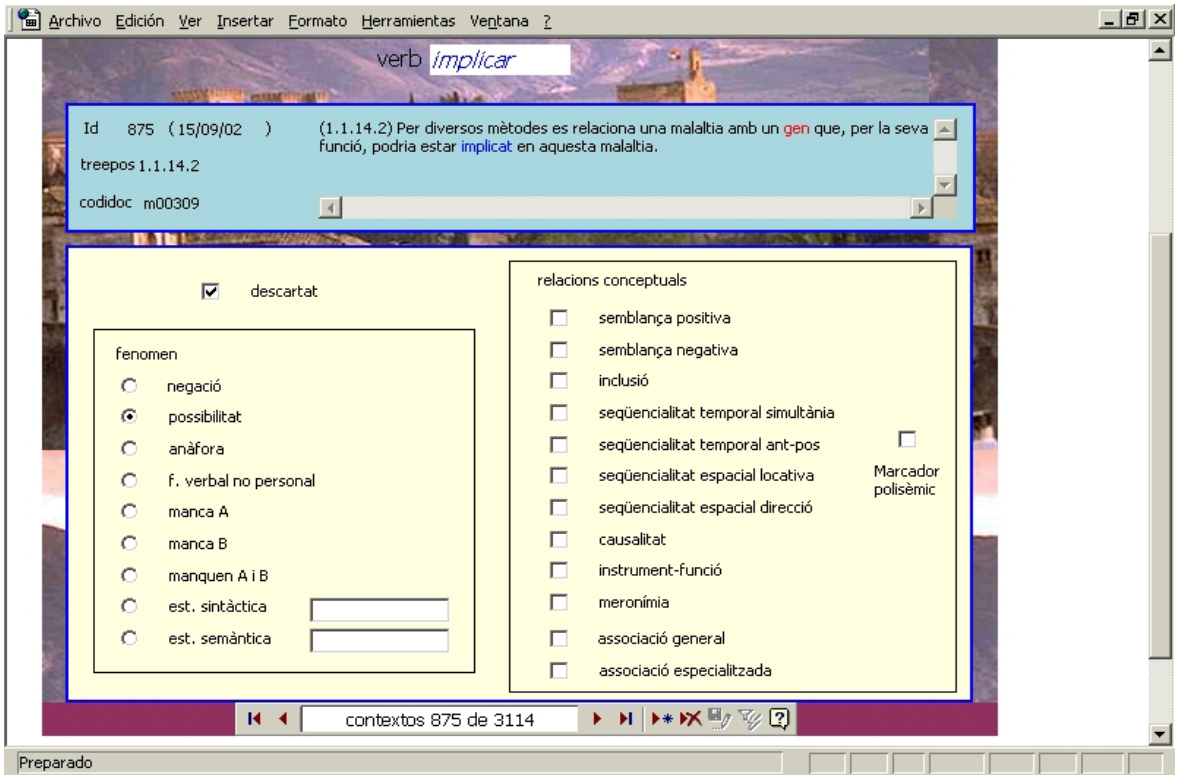


Figura 5-5. Exemple de possibilitat.

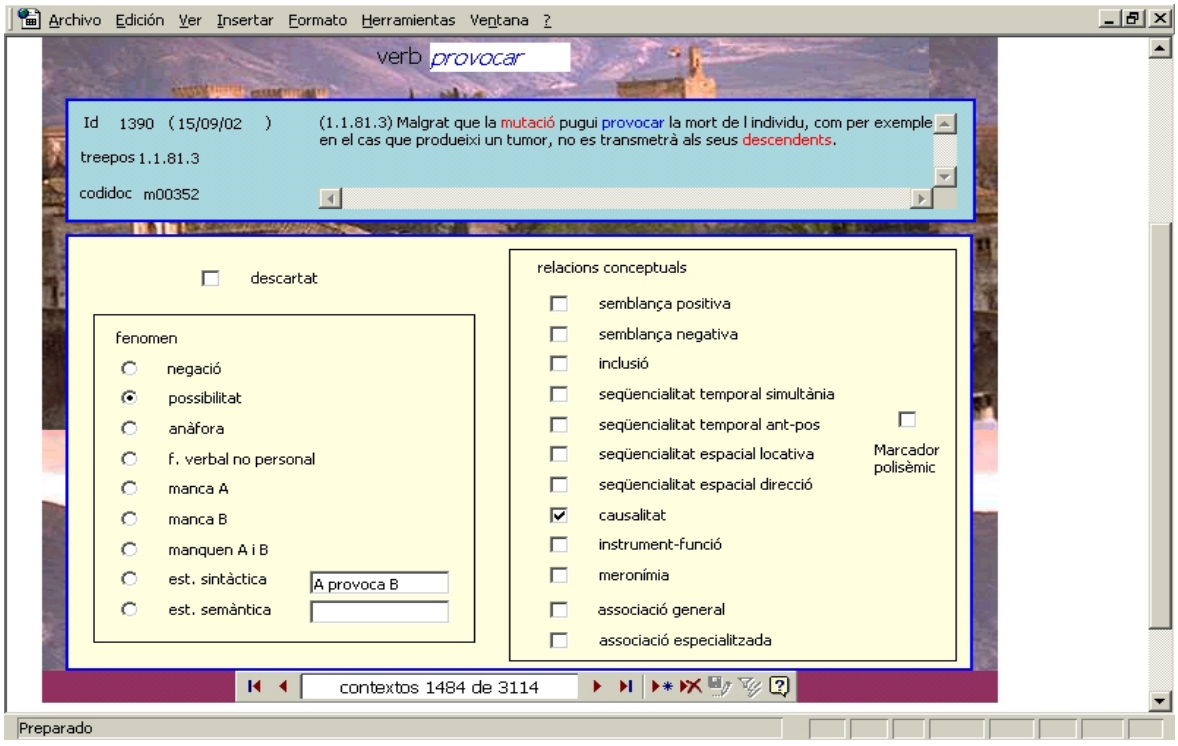


Figura 5-6. Exemple de possibilitat.

En cap dels dos casos es dóna la relació conceptual que correspondria prototípicament a les unitats verbals *implicar* (associació i, en menor mesura, causalitat) i *provocar* (causalitat) perquè el mitigador que indica possibilitat redueix el contingut semàntic de la relació.

ANÀFORA

L'anàfora és un procés discursiu molt efectiu i econòmic per als parlants però esdevé un element de difícil tractament en un sistema automàtic que no funcioni sobre la base d'un etiquetatge sintàctic i semàntic. En el nostre cas, el paràmetre de l'anàfora es troba estretament lligat amb els paràmetres de manca de conceptes, o d'unitats terminològiques, que tractarem més endavant. Evidentment, l'anàfora està reemplaçant algun element en el fragment textual però aquest no es troba explícit i, per tant, invalida la possibilitat de retenir un determinat fragment textual que mantingui l'esquema de *a*R*b*,n del qual partim.

Vegem-ne uns exemples per mostrar quins són els elements anafòrics més habituals amb què ens podem trobar:

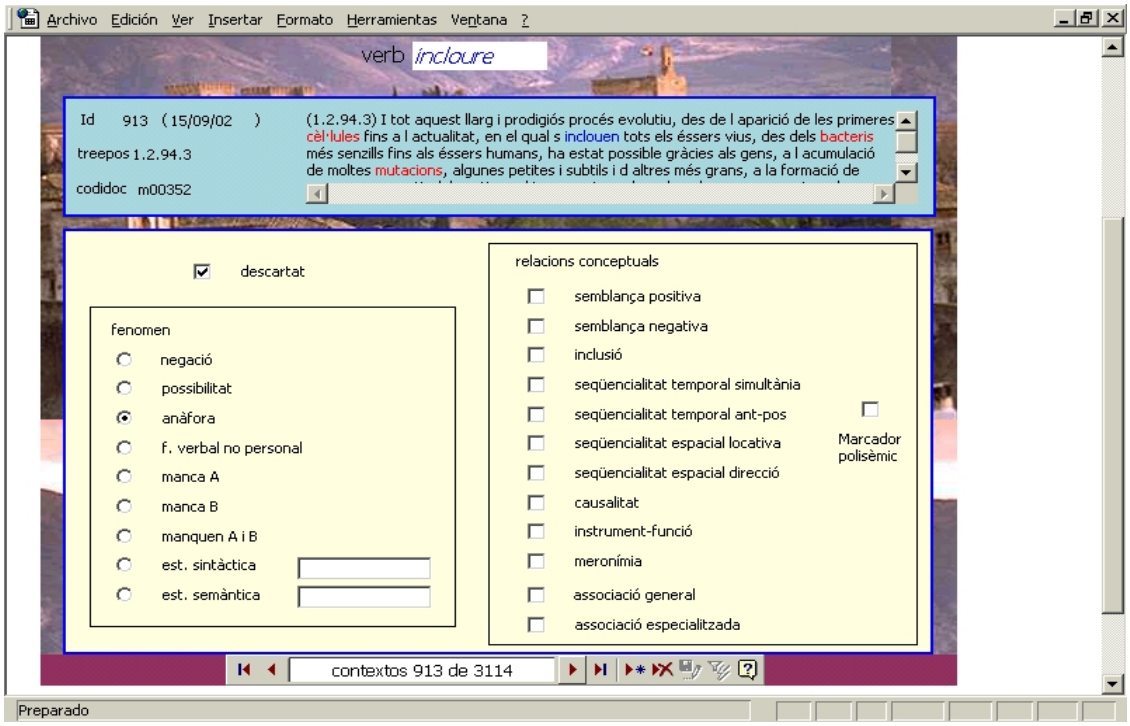
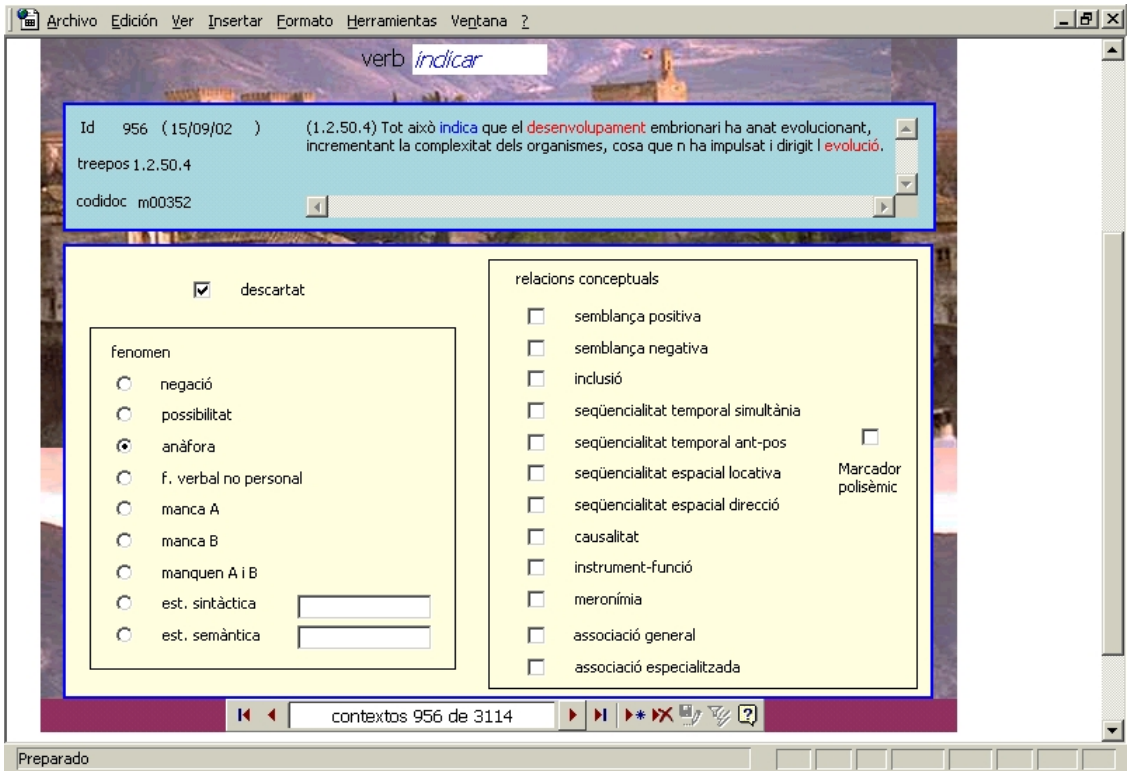


Figura 5-7. Exemple d'anàfora.

Figura 5-8. Exemple d'anàfora.



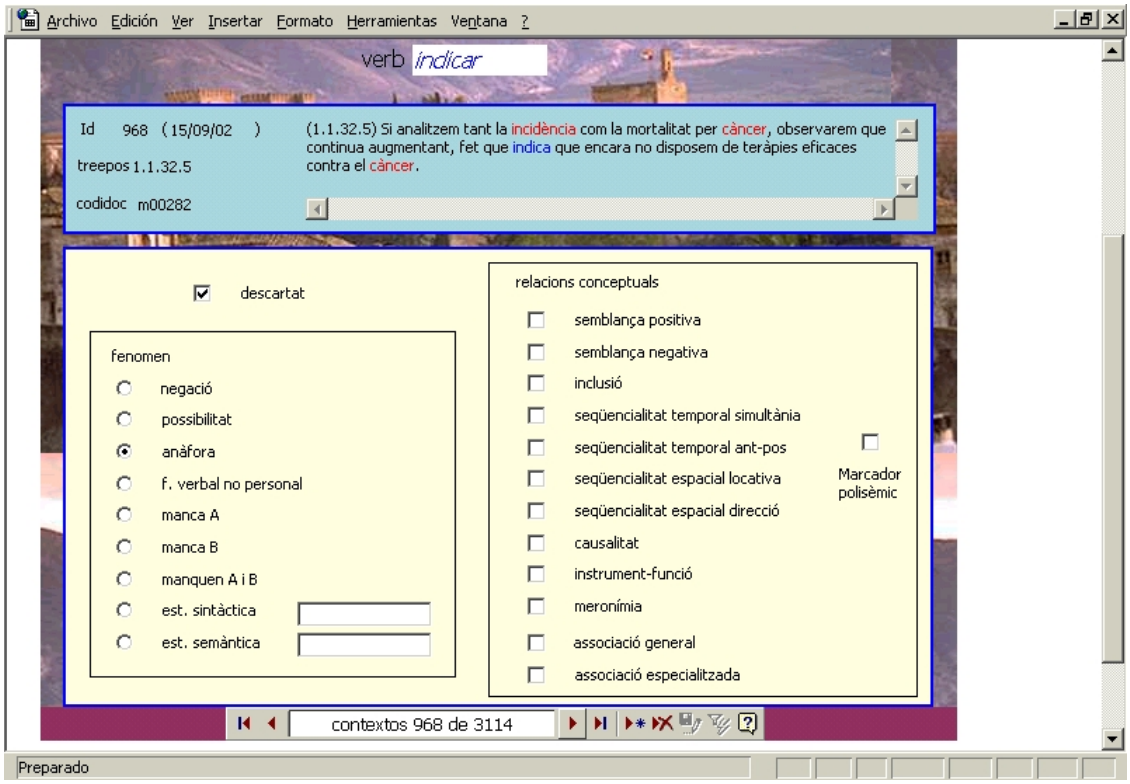


Figura 5-9. Exemple d'anàfora.

Unitats com *en el qual, tot això, fet que* i d'altres com *cosa que* emmascaren el veritable concepte, o unitat terminològica explícita, que hauria d'aparèixer en el fragment textual a banda i banda de la unitat verbal vehiculadora de relació conceptual. En aquests casos, indiquem que es tracta d'una anàfora però, igualment, podríem indicar que no apareix el concepte A o el concepte B perquè la noció d'aparició l'entenem com a materialització lingüística explícita d'una unitat terminològica.

FORMA VERBAL NO PERSONAL

També les unitats verbals, quan apareixen en forma verbal no personal del tipus infinitiu o gerundi, perden el seu valor de possibles vehiculadores de relació conceptual[8]. Aquesta pèrdua del valor es deu, eminentment, al fet que no coapareixen

[8] No és així en el cas dels participis atès que aquestes formes verbals apareixen acompanyades, majoritàriament, d'un verb amb menys pes semàntic i és precisament el participi el que vehicula la relació conceptual.

amb les unitats terminològiques necessàries per constituir un nus de coneixement complet amb la relació vehiculant dues o més unitats especialitzades. Una bona mostra la tenim en els exemples següents:

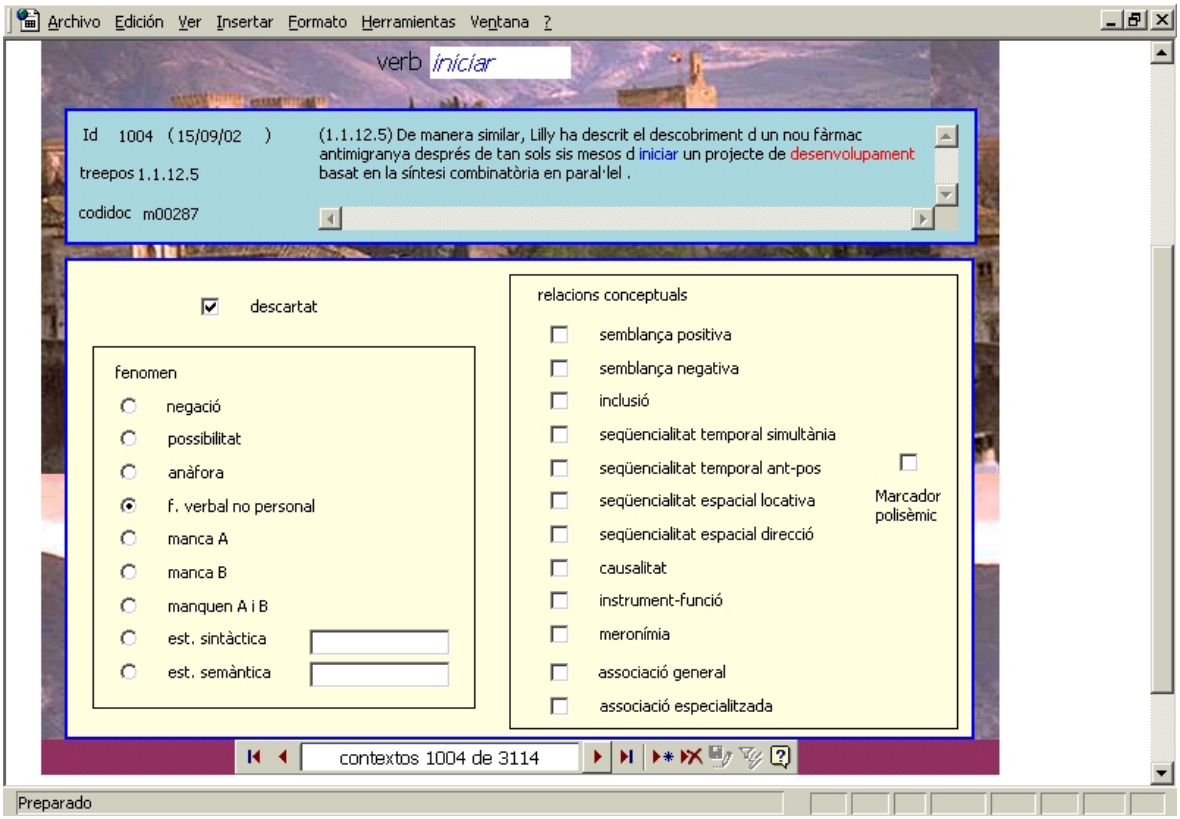


Figura 5-10. Exemple de forma verbal no personal.

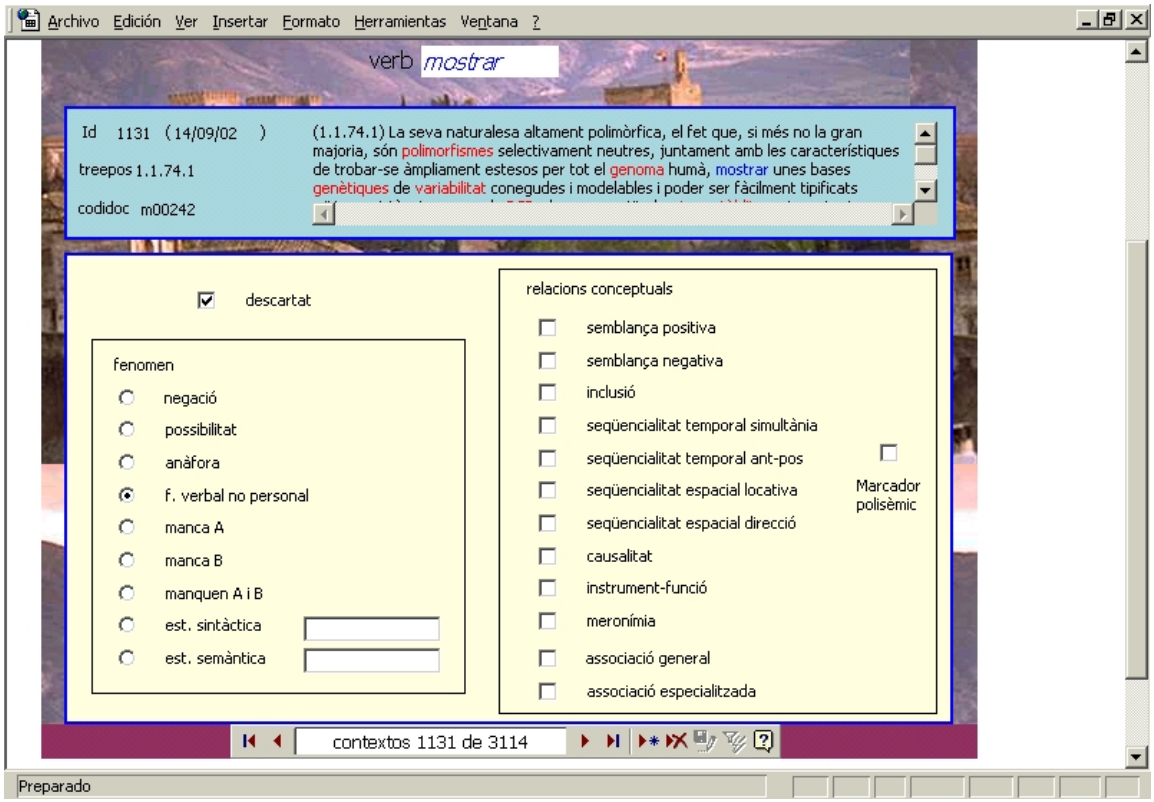


Figura 5-11. Exemple de forma verbal no personal.

En aquests casos, les unitats verbals *iniciar* i *mostrar* podrien expressar les relacions de temporalitat i associació però, com mostren els fragments textuals, l'aparició de les formes verbals no personals, tot i que corresponguin formalment als possibles marcadors de relació, impossibilita l'aparició de les unitats terminològiques requerides per a l'establiment d'una relació conceptual.

MANCA A / MANCA B / MANQUEN A I B

Agrupem, finalment, els tres últims paràmetres que apareixen per separat en el disseny de la base de dades per alleugerir la lectura de les dades i perquè es tracta de tres paràmetres molt propers i relacionats. Si recordem la definició de relació conceptual que hem anat manejant al llarg de tot el treball i que mantenim en tot moment, veiem que per considerar un fragment textual com a nus de coneixement en què apareix una relació conceptual requerim de la presència de *a*R*b*, *n*. Doncs bé, hi ha alguns casos en què els contextos ens proporcionen la unitat verbal potencialment indicadora de relació però aquesta no uneix com a mínim dos conceptes, representats

lingüísticament per unitats terminològiques. Pot ser que ens falti el primer element de la relació (*manca A*), el segon (*manca B*) o tots dos (*manquen A i B*). És per aquest motiu que els exemples següents, un per a cada tipus de paràmetre, han estat descartats:

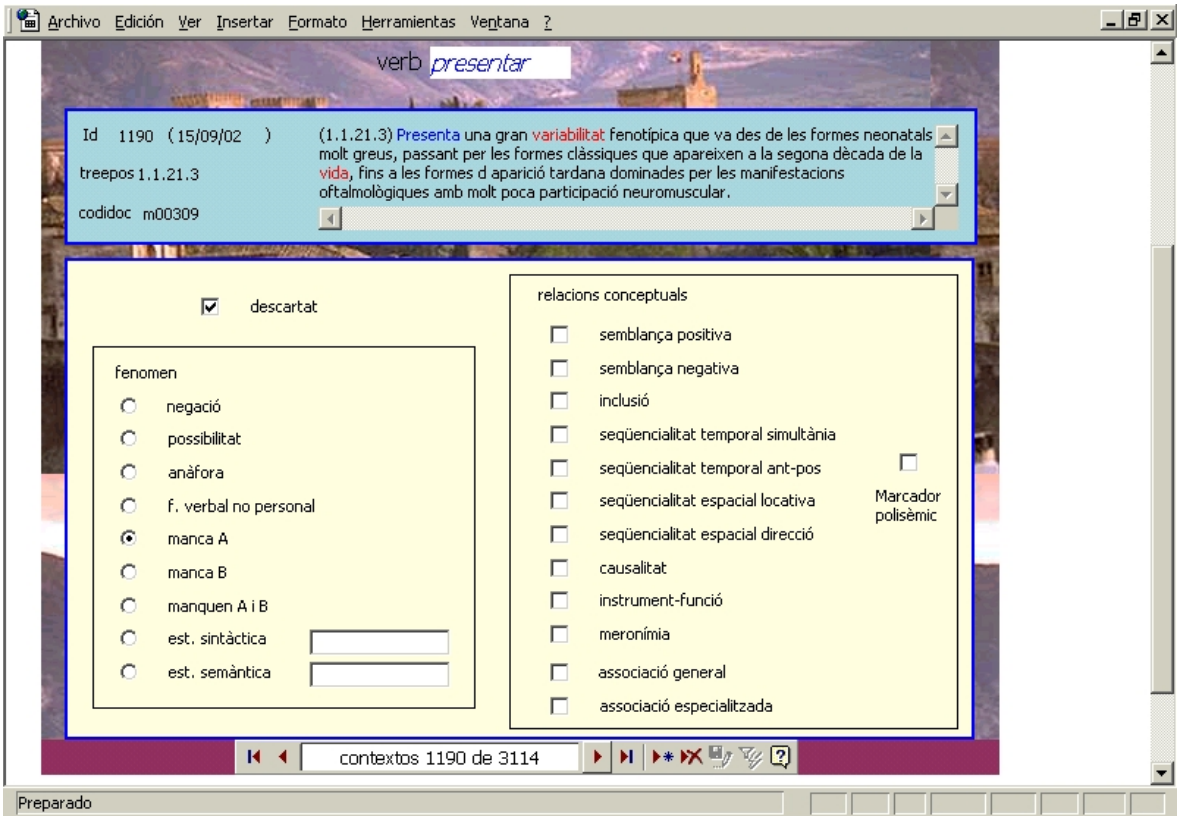


Figura 5-12. Exemple de manca A.

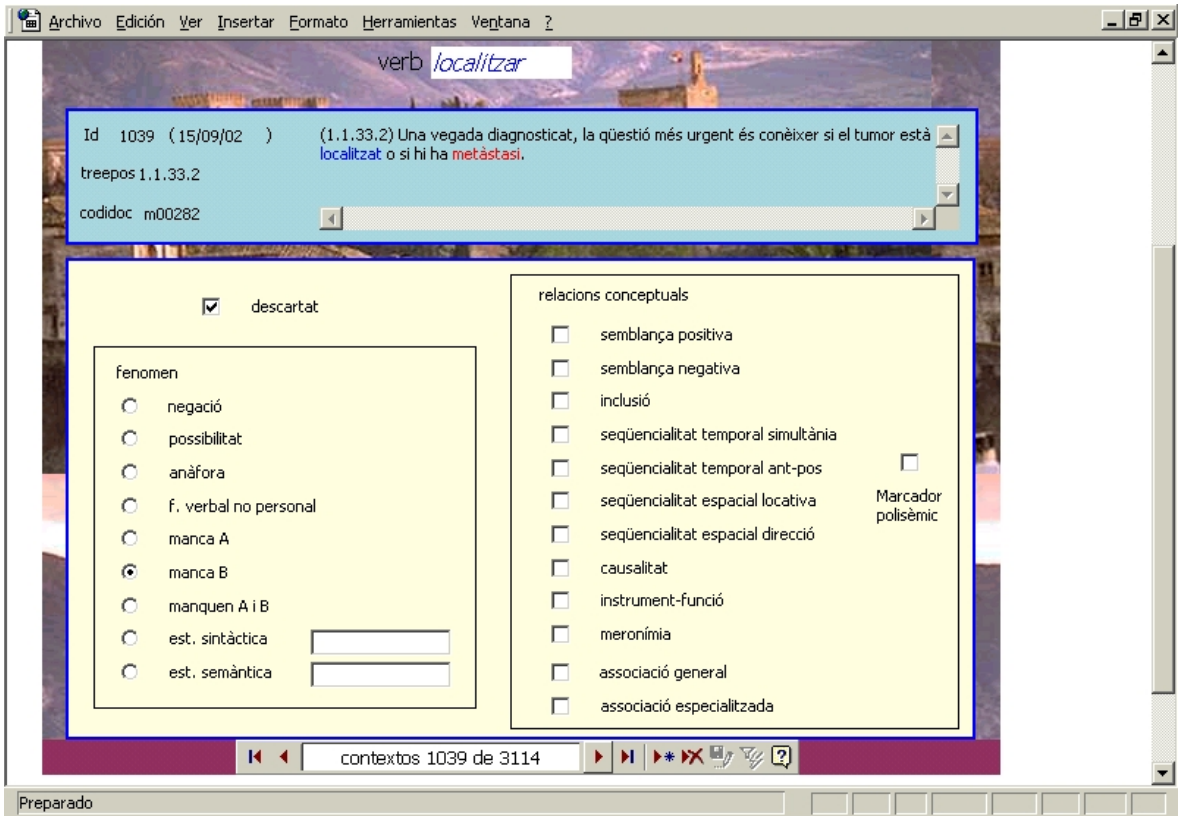


Figura 5-13. Exemple de manca B.

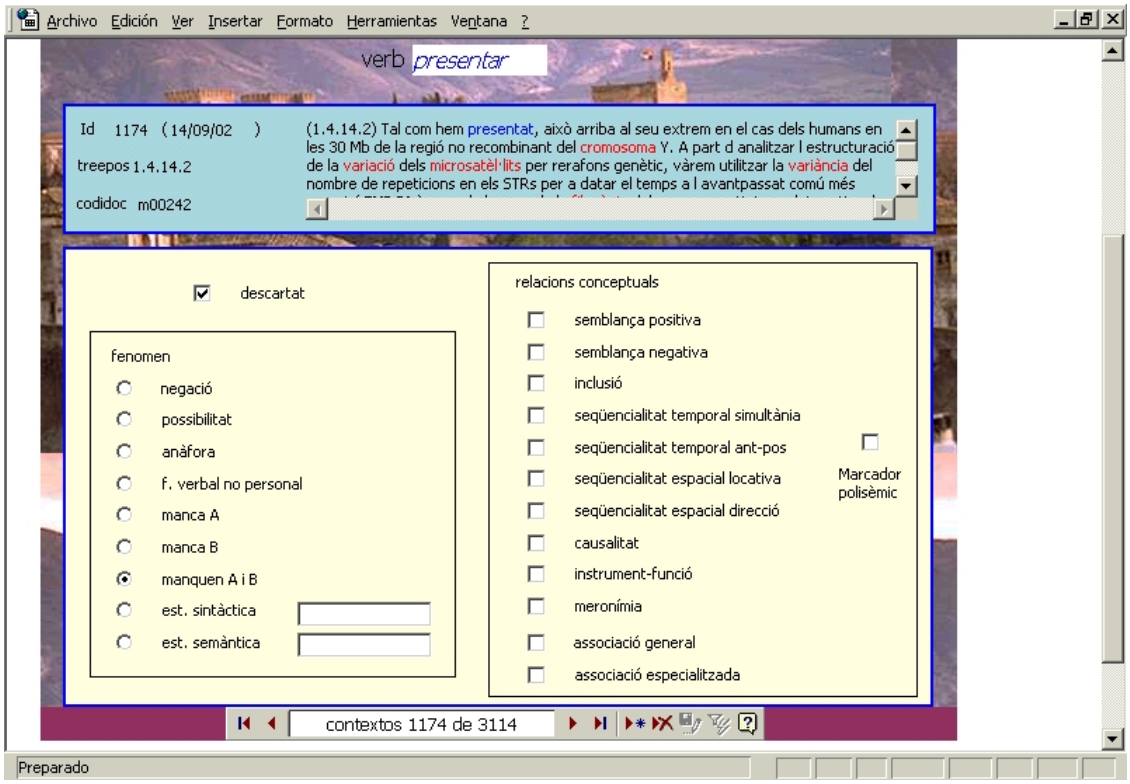


Figura 5-14. Exemple de manquen A i B.

Un cop vistos els criteris o paràmetres que ens han dut a descartar alguns dels contextos que, inicialment, el programa *Mercedes* ens presentava com a possibles integradors de relacions conceptuals, passem seguidament a endinsar-nos en el tractament dels casos en què efectivament s'explicita una relació conceptual.

5.2.2.2 Aspectes que ens porten a retenir el marcador verbal

En aquest apartat volem mostrar el tractament que han rebut els contextos en que el marcador verbal expressa efectivament una relació conceptual. El contextos que hem aïllat com a vàlids segueixen el patró *a*R*b*, *n*. En alguns casos, els conceptes *a*, *b*, i *n* apareixen indicats en color vermell perquè la unitat terminològica coincideix amb els diccionaris de base però, atès que el nombre d'unitats de què partim en el diccionari per al català, també hem retingut contextos en què apareixen unitats terminològiques que, de ben segur, amb un diccionari més ampli quedarien recollides.

En els casos en què existeix una relació conceptual explícita hem indicat una informació sintàctica, sobretot pel que fa als usos preposicionals, en la casella de la base de dades etiquetada *est. sintàctica*. Creiem que aquesta informació, tot i que no

sigui directament extrapolable al sistema de detecció semiautomàtica de relacions conceptuals, sí permetrà aprofundir en estudis futurs sobre els patrons sintàctics que expressen una determinada relació.

A més d'aquesta breu informació sintàctica, que recollim en un primer estadi de la recerca i que aprofundirem en futurs treballs, atès que no representa l'objecte central d'aquesta tesi, hem afegit, en els casos en què la unitat verbal expressa més d'una relació, informació de tipus semàntic, sobre la qual hem treballat detalladament. Aquesta informació semàntica per als marcadors polisèmics recull la categoria semàntica a què corresponen les unitats terminològiques lligades pel marcador. Aquest aspecte es tracta amb més detall en l'apartat següent d'aquest capítol però volíem, en aquesta primera descripció de la base de dades, posar un parell d'exemples que mostrin el tipus d'informació sintàctica i semàntica que de ben segur ajudarà al disseny del prototip de detecció semiautomàtica de relacions conceptuals.

Les dues imatges següents ens mostren el tipus d'informació sintàctica recollida per als verbs *utilitzar* (indicador de relació intrument-funció) i *trobar* (que expressa seqüencialitat espacial locativa):

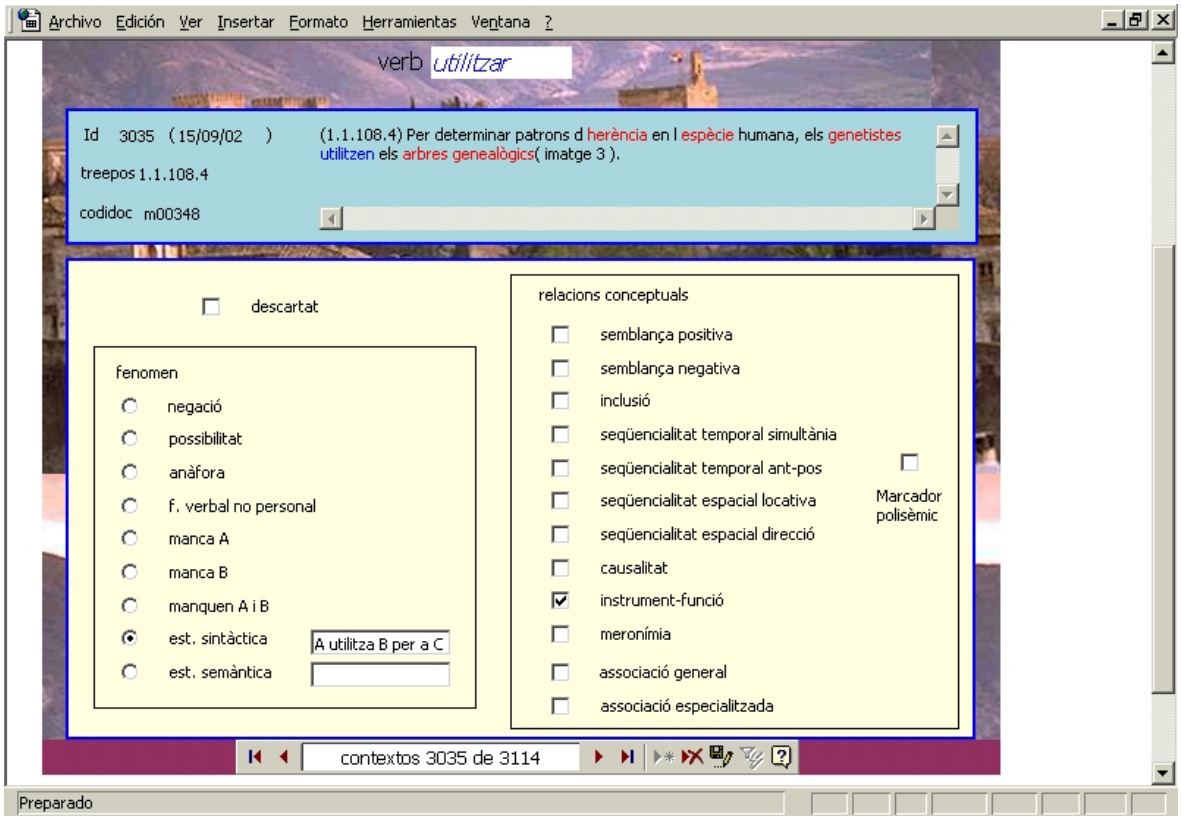


Figura 5-15. Mostra del patró sintàctic del marcador *utilitzar*.

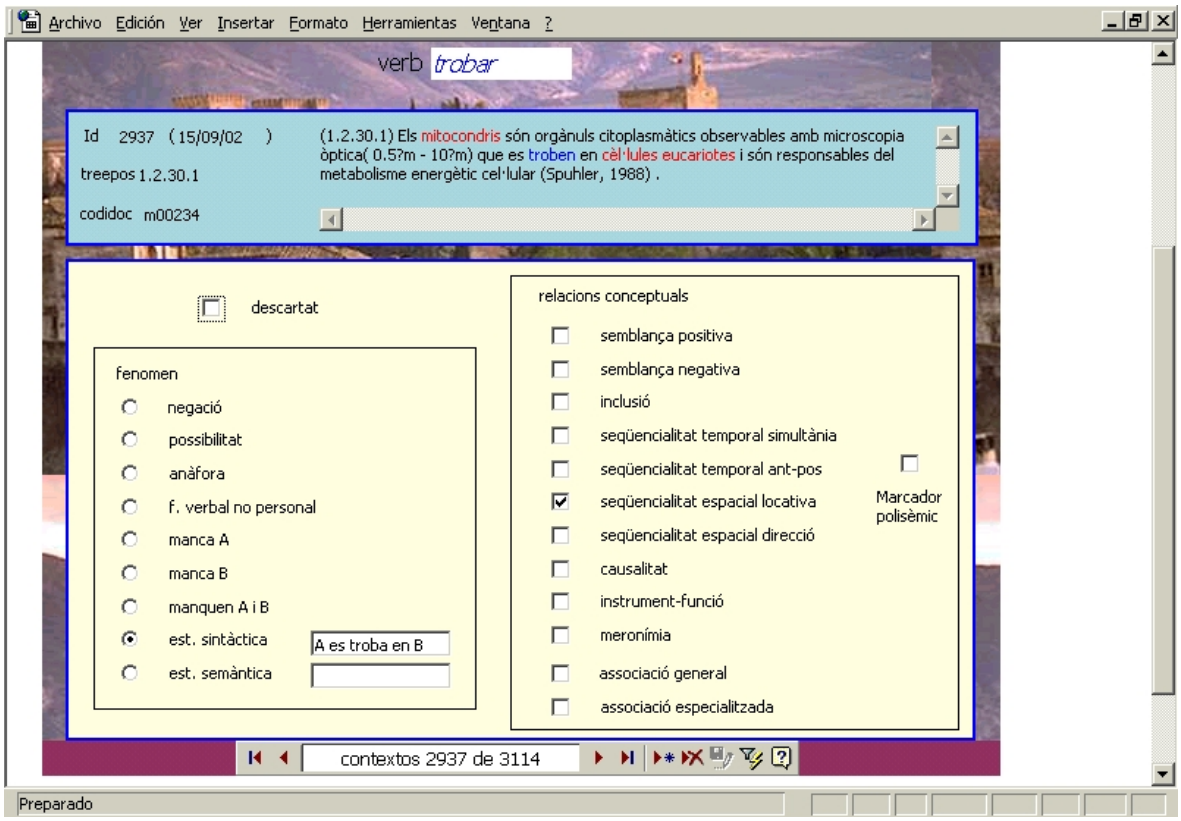


Figura 5-16. Mostra del patró sintàctic per al marcador *trobar*.

En el primer cas, s'indica que la relació d'instrument-funció s'expressa a partir del marcador verbal *utilitzar* seguint el patró [A utilitza B per a C], on entren en joc tres unitats terminològiques. En la segona imatge, que representa un registre del verb *trobar*, la relació de seqüencialitat espacial locativa s'expressa eminentment per la preposició *en* en l'estructura [A es troba en B]. Aquesta informació servirà d'ajuda per a la comprovació manual necessària dels resultats que proporcioni el sistema de detecció semiautomàtica de relacions conceptuals.

En relació a la informació semàntica, vegem un exemple del marcador polisèmic *situar* que, en dos contextos diferents, vehicula dues relacions conceptuals també diferents depenent de la categoria semàntica a què pertany el concepte B. En el

primer cas el concepte B pertany a la categoria *lloc* i, en el segon cas, B correspondria a la categoria conceptual *temps*. Vegem-ho:

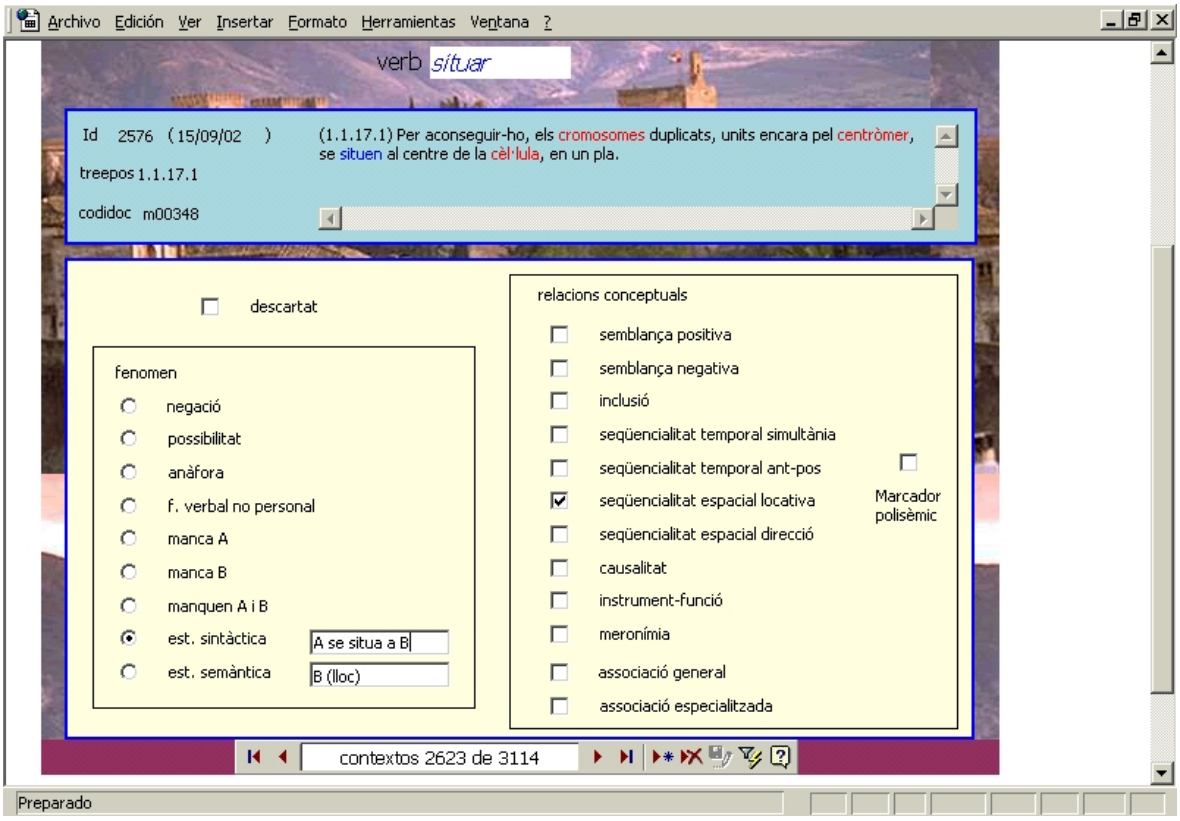


Figura 5-17. Informació semàntica del concepte *b* (lloc) per al marcador *situar*.

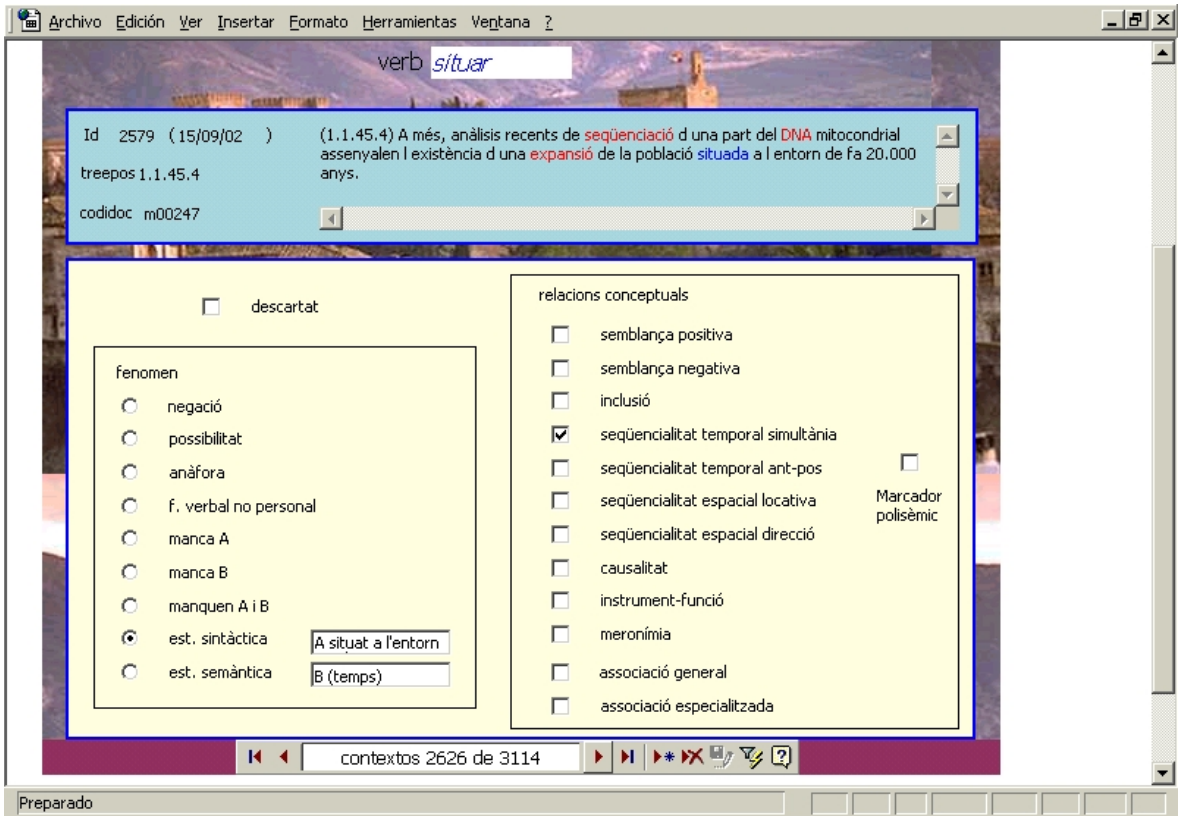


Figura 5-18. Informació semàntica del concepte *b* (temps) per al marcador *situar*.

5.2.2.3 Dades numèriques

Abans de tancar aquest apartat sobre la informació continguda en la base de dades i els paràmetres que ens han portat a descartar o retenir una determinada unitat verbal com a vehiculadora de relació conceptual volem esbossar de manera esquemàtica algunes dades numèriques sobre la proporció de contextos que expressen coneixement especialitzat sobre la base d'una relació conceptual explícita.

Marcador	Contextos retinguts	Contextos descartats	Total contextos	Percentatge de retenció	Soroll
allunyar	3	1	4	75%	25%
aparèixer	6	14	20	30%	70%
apropar	0	1	1	0%	100%
arribar	4	22	26	15,38%	84,62%
caracteritzar	22	11	33	66,7%	33,3%
causar	37	7	44	84,1%	15,9%
compondre	1	4	5	20%	80%
considerar	5	18	23	21,74%	78,26%

constar	3	6	9	33,3%	66,7%
Marcador	Contextos retinguts	Contextos descartats	Total contextos	Percentatge de retenció	Soroll
constituir	19	8	27	70,37%	29,63%
continuar	2	8	10	20%	80%
contribuir	3	1	4	75%	25%
correlacionar	0	1	1	0%	100%
correspondre	17	8	25	68%	32%
definir	3	12	15	20%	80%
dependre	12	6	18	66,7%	33,3%
determinar	42	16	58	72,41%	27,59%
deure	61	11	72	84,72%	15,28%
diferenciar	10	6	16	62,5%	37,5%
distingir	4	4	8	50%	50%
donar	30	58	88	34,1%	65,9%
englobar	5	0	5	100%	0%
evidenciar	6	3	9	66,7%	33,3%
fer	19	206	225	8,4%	91,6%
formar	73	34	107	68,22%	31,78%
implicar	35	14	49	71,43%	28,57%
incloure	30	3	33	90,90%	9,10%
indicar	39	13	52	75%	25%
iniciar	7	10	17	41,18%	58,82%
integrar	6	9	15	40%	60%
intervenir	10	4	14	71,43%	28,57%
localitzar	16	10	26	61,54%	17,86%
manifestar	11	5	16	68,75%	31,25%
mesurar	0	7	7	0%	100%
mostrar	39	27	66	59,1%	40,9%
originar	10	10	20	50%	50%
presentar	71	27	98	72,45%	27,55%
produir	47	64	111	42,34%	57,66%
propagar	0	1	1	0%	100%
provocar	24	11	35	68,57%	31,43%
quedar	4	23	27	14,81%	85,19%
realitzar	11	32	43	25,58%	74,42%
reflectir	1	1	2	50%	50%
representar	27	12	39	69,23%	30,76%
reunir	0	3	3	0%	100%
ser	346	699	1.045	33,11%	66,89%
simular	1	0	1	100%	0%
situar	22	7	29	75,86%	24,14%
suggerir	1	14	15	6,7%	93,3%
tenir	141	99	240	58,75%	41,25%
transcórrer	0	2	2	0%	100%
trobar	34	90	124	27,42%	72,58%

usar	0	1	1	0%	100%
Marcador	Contextos retinguts	Contextos descartats	Total contextos	Percentatge de retenció	Soroll
utilitzar	37	45	82	45,12%	54,88%
veure	2	46	48	4,17%	95,83%

Taula 5-1. Marcadors verbals amb freqüència d'aparició i percentatge de retenció.

Els casos que presenten un percentatge de retenció més elevat, és a dir, una precisió més alta sense haver aplicat, *a priori*, cap criteri de restricció a les dades i cercant només la unitat verbal aïllada són:

allunyar, *caracteritzar*, *causar*, *constituir*, *contribuir*, *correspondre*, *dependre*, *determinar*, *deure*, *diferenciar*, *distingir*, *englobar*, *evidenciar*, *formar*, *implicar*, *incloure*, *indicar*, *intervenir*, *localitzar*, *manifestar*, *mostrar*, *originar*, *presentar*, *provocar*, *reflectir*, *representar*, *simular*, *situar*, *tenir*, *utilitzar*.

Tanmateix, cal destacar que no tots els marcadors tenen la mateixa freqüència d'aparició i, per tant, no totes les dades són igualment significatives. Dels marcadors anteriors, només tenen una freqüència d'aparició superior a 25 ocurrències els verbs següents:

caracteritzar, *causar*, *constituir*, *corresponde*, *determinar*, *deure*, *formar*, *implicar*, *incloure*, *indicar*, *localitzar*, *mostrar*, *presentar*, *provocar*, *representar*, *situar* i *tenir*.

Les unitats verbals *allunyar*, *contribuir*, *dependre*, *diferenciar*, *distingir*, *englobar*, *evidenciar*, *intervenir*, *manifestar*, *originar*, *refectir*, *simular* i *utilitzar*, en canvi, manifesten una freqüència d'aparició força més baixa i, per tant, tot i que el seu percentatge de retenció sigui elevat, cal tenir en compte, si més no, que són unitats menys significatives quant a la freqüència en el corpus.

Així, i com es pot observar de manera general, hi ha diferències quant a la precisió en aquests 30 marcadors que superen, en tots els casos, un percentatge del 50%. Dels 25 marcadors restants, n'hi ha 7 que provoquen un percentatge de soroll total tot i que, en quatre dels set casos només disposem d'una ocurrència, indicada entre parèntesi. Aquesta marcadors són: *apropar* (1), *correlacionar* (1), *mesurar* (7),

propagar (1), *reunir* (3), *transcórrer* (2), *usar* (1). Considerem, per tant, que en aquests casos les dades no són prou representatives com per poder afirmar que aquests marcadors mai no expressen una relació conceptual. Tanmateix, el tractament d'aquestes unitats, amb una freqüència d'aparició força baixa, ha de ser per força diferenciat de la resta d'unitats amb un soroll elevat, com és el cas de les 20 unitats següents:

aparèixer, *arribar*, *compondre*, *considerar*, *constar*, *continuar*, *definir*, *donar*, *fer*, *iniciar*, *integrar*, *produir*, *quedar*, *realitzar*, *ser*, *suggerir*, *trobar*, *veure*.

Considerem que, a partir de l'aplicació dels paràmetres per descartar un determinat marcador verbal, les dades relatives al soroll canviarien completament i aquest es reduiria en gran mesura.

Pel que fa al silenci, i com hem comentat anteriorment, el fet de buscar unitats verbals aïllades incrementa el percentatge de soroll però elimina el risc de silenci que es provocaria si el programa funcionés amb unitats verbals seguides d'altres elements sintàctics com són les preposicions, els adverbis i les locucions preposicionals i adverbials per la raó que ja hem comentat anteriorment. Som conscients, però, que el silenci també afecta totes les unitats verbals que poden expressar relacions conceptuals i que no tenim recollides en el llistat de partida. La nostra voluntat és anar incrementant la llista a partir de la qual treballarà el sistema de detecció semiautomàtica sobre la base de treballs futurs, de manera que la nova unitat verbal o grups d'unitats vehiculadores d'una determinada relació hagin estat objecte d'un estudi previ.

5.3 Refinament de la detecció de relacios conceptuals: cap al recurs dels patrons sintàctics

Al llarg de tot aquest capítol hem anat presentant els marcadors verbals aïlladament atès que l'eina que hem utilitzat per detectar automàticament aquestes unitats no ens permet, fins a l'actualitat, cercar la informació de la unitat verbal tenint en compte el

seu patró sintàctic relatiu, sobretot, a la preposició que s'adjunta al verb per indicar una determinada relació conceptual. Som conscients, tanmateix, que aquesta informació és essencial, com demostra l'anàlisi del corpus que hem dut a terme al llarg del capítol 4, per tal de refinar el tipus de relació conceptual vehiculada per un determinat marcador i, sobretot, el pes que confereix una determinada preposició a aquesta unitat verbal. Per aquest motiu, i com a una primera aproximació a l'estructura sintàctica de cada marcador, volem reflectir aquí la informació que hem detallat en la base de dades en la casella d'estructura sintàctica que, de ben segur, ens ha de permetre realitzar cerques més acurades sobre el corpus d'anàlisi de manera semiautomàtica quan disposem d'un corpus etiquetat sintàcticament. En aquest sentit, preveiem que en un futur no gaire llunyà el corpus d'anàlisi contindrà informació sintàctica sobre les unitats objecte d'anàlisi i, a més, el sistema permetrà trobar aquestes estructures encara que els elments que les consitueixen no es trobin adjacents. Aquesta aplicació de patrons sintàctics, que combinarem amb els mètodes d'extracció proposats fins ara i també amb les estratègies que presentem en el capítol 6 ens portaran a reduir les possibles ambigüitats d'un determinat marcador verbal pel que fa a la relació conceptual que expressa.

Presentem, doncs, la llista de patrons sintàctics (amb els elements conceptuals *a*, *b*, *n* implicats en la relació conceptual) i la relació conceptual que vehiculen quan presenten una estructura diferent de [a *marcador verbal* b, n] —que és l'estructura bàsica de partida— per a cadascun dels marcadors verbals que acabem de presentar en la Taula 5.1.

allunyar

a *allunyat com* b, n [semblança negativa i seqüencialitat espacial locativa]

aparèixer

a *aparèixer a/en* b [seqüencialitat espacial locativa]

a *aparèixer durant* b [seqüencialitat espacial ant-pos]

a *aparèixer fa* b [seqüencialitat espacial simultània]

arribar

a *arribar a* b [seqüencialitat espacial direcció]

caracteritzar(-se)

b *és/està caracteritza per* a [associació general]

b *es caracteritza per* a [associació general]

causar

b *(ser/estar) causat per* a [causalitat]

a *és la causa de* b [causalitat][9]

considerar(-se)

a *es considera com* b [associació general]

constar

a *consta de* b [meronímia]

constituir

b *està constituït per* a [meronímia]

continuar

a *continua en* b [seqüencialitat temporal ant-pos]

[9] En aquest cas la unitat *causa* està mal etiquetada al corpus i apareix com a verb però ens és útil de retenir aquest patró per a aplicacions posteriors atès que sempre indica relació de causalitat.

contribuir

a *contribueix a* b [causalitat]

correspondre('s)

a *(es) correspon a* b [associació general]

a *(es) correspon en* b [associació general]

a *es correspon amb* b [associació general]

definir(-se)

a *es defineix per* b [instrument-funció]

a *es defineix amb* b [instrument-funció]

dependre

a *depèn de* b [causalitat]

deure('s)

b *és degut a* a [causalitat]

b *es deu a* a [causalitat]

diferenciar(-se)

a *diferenciar(-se) de* b [associació especialitzada i semblança negativa]

a *es diferencia en* b *de* n [semblança negativa]

a *es diferencia per* b [sembança negativa]

a *es diferencia en* b [associació especialitzada]

a *es diferencia cap a* b [associació especialitzada]

distingir(-se)

a *es distingeix de* b [semblança negativa]

donar

a *dóna lloc a* b [causalitat]

a *dóna origen a* b [causalitat]

b *ve donat per* a [causalitat]

englobar

b *englobat dins* a [meronímia]

fer(-se)

a *es fa gràcies a* b [instrument-funció]

a *es fa amb* b [instrument-funció]

a *es fa segons* b [instrument-funció]

a *es fa gràcies a* b [instrument-funció]

formar(-se)

A partir de b, *es forma* a [meronímia]

a *(està/és) format per* b [meronímia]

b *forma part de* a [meronímia]

implicar

a *(està) implicat en* b [associació especialitzada]

b *implica* a [causalitat]

incloure('s)

a, *incloent-hi* b [inclusió]

b *s'inclou en a* [inclusió i meromínia]

b *s'inclou dins* a [inclusió i meromínia]

b *s'inclou entre* a [inclusió i meromínia]

indicar(-se)

a *s'indica sota* b [seqüencialitat espacial locativa]

a *s'indica amb* b [instrument-funció]

a *s'indica en* b [seqüencialitat espacial locativa]

a *s'indica a* b [seqüencialitat espacial locativa]

iniciar(-se)

a *s'inicia en* b [seqüencialitat espacial locativa]

a *s'inicia cap a* b [seqüencialitat espacial simultània]

a *s'inicia poc temps després de* b [seqüencialitat espacial ant-pos]

a *s'inicia amb* b [seqüencialitat espacial simultània i instrument-funció]

integrar(-se)

b *s'integra en* a [meromímia]

b *és integrat dins de* a [meronímia]

intervenir

a *intervé en* b [associació general]

localitzar(-se)

a *(és/està) localitzat a* b [seqüencialitat espacial locativa i associació especialitzada]

a *(és/està) localitzat en* b [seqüencialitat espacial locativa i associació especialitzada]

a *es localitza en* b [seqüencialitat espacial locativa i associació especialitzada]

a *es localitza dins de* b [seqüencialitat espacial locativa]

manifestar(-se)

a *és manifestat per* b [associació especialitzada]

mostrar(-se)

b *es mostra en* a [associació general]

b *es mostra a* a [associació general]

a *mostra* b *en* c [associació general]

originar(-se)

a *origina primer* b *i després*, c [causalitat i seqüencialitat temporal ant-pos]

a *va originar-se a* b [seqüencialitat espacial locativa]

a *(és/està) originat per* b [causalitat]

presentar(-se)

a *es presenta a* b [associació general]

a *es presenta dins de* b [associació general]

a *es presenta com a* b [associació general]

a *és presentat per* b [associació general]

produir(-se)

a *es produeix en* b [seqüencialitat espacial locativa]

a *es produeix a* b [seqüencialitat espacial locativa]

a *es produeix gràcies a* b [causalitat i instrument funció]

a *es produeix per* b [causalitat]

provocar

b *(és/està) provocat per* a [causalitat]

quedar

a *queda + participi en* b [seqüencialitat espacial locativa]

a *queda a* b [seqüencialitat espacial locativa]

realitzar(-se)

a *(és/està) realitzat amb* b [instrument-funció]

a *es realitza en* b [seqëncialitat espacial locativa]

a *es realitza a* b [seqëncialitat espacial locativa]

a *es realitza durant* b [seqüencialitat temporal ant-pos]

reflectir(-se)

a *es reflecteix en* b [associació general]

representar(-se)

a *es representa en* b [seqüencialitat espacial locativa]

a *(és/està) representat per* b [associació general]

a *(és/està) representat amb* b [instrument-funció]

ser

a *és un* b [inclusió, semblança positiva i associació general]

a *és el* b [inclusió, semblança positiva i associació general]

a *és dins de* b [seqüencialitat espacial locativa]

a, *és a dir*, b [semblança positiva][10]

a *és com* b [semblança positiva]

a *és en* b [seqüencialitat espacial locativa]

a *és semblant a* b [semblança positiva]

a *és present en* b [seqüencialitat espacial locativa]

situar(-se)

a *(és/està) situat a* b [seqüencialitat espacial locativa]

[10] *És a dir*, es tracta generalment com a connector reformulatiu però, atès que conté el verb *ser*, sembla interessant de retenir-lo per a estudis futurs sobre la relació de sembança positiva.

a *(és/està) situat en* b [seqüencialitat espacial locativa]

a *se situa a* b [seqüencialitat espacial locativa]

a *(és/està) situat a l'entorn de* b [seqüencialitat temporal simultània i espacial locativa]

a *situa* b *amb* c [seqüencialitat temporal simultània]

a *situa* b *en c* [seqüencialitat espacial locativa]

a *se situa entre* b *i* c [seqüencialitat espacial locativa]

a *(és/està) situat dins de* b [seqüencialitat espacial locativa]

a *(és/està) situat per* b [seqüencialitat temporal simultània]

tenir

a *té lloc dins* b [seqüencialitat espacial locativa]

a *té lloc en* b [seqüencialitat espacial locativa]

a *té lloc a* b [seqüencialitat espacial locativa]

a *té lloc entre* b *i* c [seqüencialitat espacial locativa]

a *té una funció* b [instrument-funció]

a *té la funció de* b [instrument-funció]

a *té com a funció* b [instrument-funció]

trobar(-se)

a *es troba en* b [seqüencialitat espacial locativa]

a *es troba dins* b [seqüencialitat espacial locativa]

a *es troba a l'interior de* b [seqüencialitat espacial locativa]

a *es troba localitzat a* b [seqüencialitat espacial locativa]

a *es troba situat en* b [seqüencialitat espacial locativa]

utilitzar(-se)

a *(és/està) utilitzat com a* b [instrument-funció]

a *s'utilitza com a* b [instrument-funció]

a *s'utilitza en* b [instrument-funció]

a *s'utilitza per a* b [instrument-funció]

a *utilitza* b *per a* c [instrument-funció]

veure('s)

a *es veu en* b [seqüencialitat espacial locativa i associació general]

Aquests patrons sintàctics han d'ajudar, sens dubte, a la detecció d'alguns dels tipus de relacions conceptuals i, com es pot observar, permeten gràcies a l'ús de les preposicions establir un gran nombre de marcadors que indiquen relació de seqüencialitat espacial locativa i d'instrument funció. Tot i això, encara ens queden uns quants marcadors que resulten polisèmics pel que fa a la relació conceptual que expressen i, per aquest motiu, aquest treball integra la llista de patrons sintàctics amb una proposta semàntica aprofundida que presentem seguidament.

5.4 Tractament dels marcadors polisèmics: cap al recurs semàntic de l'ontologia

Al llarg d'aquest treball s'ha demostrat que un gran nombre dels marcadors verbals vehiculadors de relació conceptual poden expressar més d'un tipus de relació conceptual. En alguns casos, com hem vist en analitzar les dades contingudes a la base de dades, en un mateix context un marcador pot esdevenir polisèmic i indicar dues relacions conceptuals diferents. Majoritàriament, però, la polisèmia dels marcadors es dóna en contextos on apareixen unitats terminològiques de naturalesa diferent.

En aquest apartat pretenem indicar una possible via de resolució dels casos de polisèmia que hem anomenat, ja a la introducció d'aquest treball, *factor del context*, noció que apareixerà a partir d'aquest moment com a element clau per a la desambiguació de l'etiquetatge de contextos amb una o més indicacions de relació conceptual. Creiem que el recurs semàntic de l'ontologia permetria desambiguar, o com a mínim ajudar a decidir, quin tipus de relació conceptual té lloc en cada context. Recordem que algunes unitats que solen comportar-se com a marcadors polisèmics:

- *allunyar* pot expressar semblança negativa i seqüencialitat espacial locativa;
- *aparèixer* pot expressar seqüencialitat espacial locativa i temporal d'anterioritat-posterioritat;
- *constituir* pot expressar meronímia i, en alguns casos, quan actua com a sinònim del verb *ser*, inclusió;
- *diferenciar* pot expressar associació especialitzada i semblança negativa;
- *fer* pot expressar causalitat i instrument-funció;
- *implicar* pot expressar associació especialitzada i causalitat;

- *incloure* pot expressar inclusió i meronímia;
- *indicar* pot expressar seqüencialitat espacial locativa, instrument-funció i associació general;
- *iniciar* pot expressar els dos tipus de seqüencialitat temporal, la seqüencialitat locativa i instrument-funció;
- *localitzar* pot expressar seqüencialitat espacial locativa i instrument-funció;
- *originar* pot expressar els dos tipus de seqüencialitat temporal, seqüencialitat espacial locativa i causalitat;
- *presentar* pot expressar associació general i meronímia (quan és sinònim de *tenir*);
- *produir* pot expressar seqüencialitat espacial locativa, causalitat, seqüencialitat temporal simultània i anterioritat-posterioritat, i instrument-funció;
- *realitzar* pot expressar instrument-funció, seqüencialitat espacial locativa i seqüencialitat temporal anterioritat-posterioritat;
- *representar* pot expressar associació general i seqüencialitat espacial locativa;
- *ser* pot expressar inclusió, associació general, seqüencialitat espacial locativa, semblança positiva i negativa, i causalitat (quan, per exemple, va acompanyat del nom *causa* com, per exemple, en el registre 2480 del document m00282 "Avui, la sida és la causa principal de mort en adults").
- *situar* pot expressar seqüencialitat temporal simultània i seqüencialitat espacial locativa;
- *tenir* pot expressar meronímia (i així ho fa en la majoria de casos) però també associació general, instrument-funció i seqüencialitat espacial locativa;

- *trobar* pot expressar seqüencialitat espacial locativa, causalitat i en algun cas, inclusió;

- *veure* pot expressar associació general i seqüencialitat espacial locativa.

Observem clarament que un gran nombre de marcadors poden expressar més d'un tipus de relació conceptual i creiem que amb l'ajuda de l'ontologia, la diversitat i el contingut semàntic de la qual han estat àmpliament descrits en el capítol 3 d'aquest treball, podríem refinar la proposta de detecció de relacions conceptuals.

A continuació exemplificarem amb contextos reals els beneficis que creiem que pot tenir l'ús de l'ontologia. Els contextos com el següent són, com hem pogut veure al llarg del treball, poc nombrosos. Es tracta d'un context en què la unitat *ser* acompanyada pel determinant *un* expressa una relació d'inclusió clara:

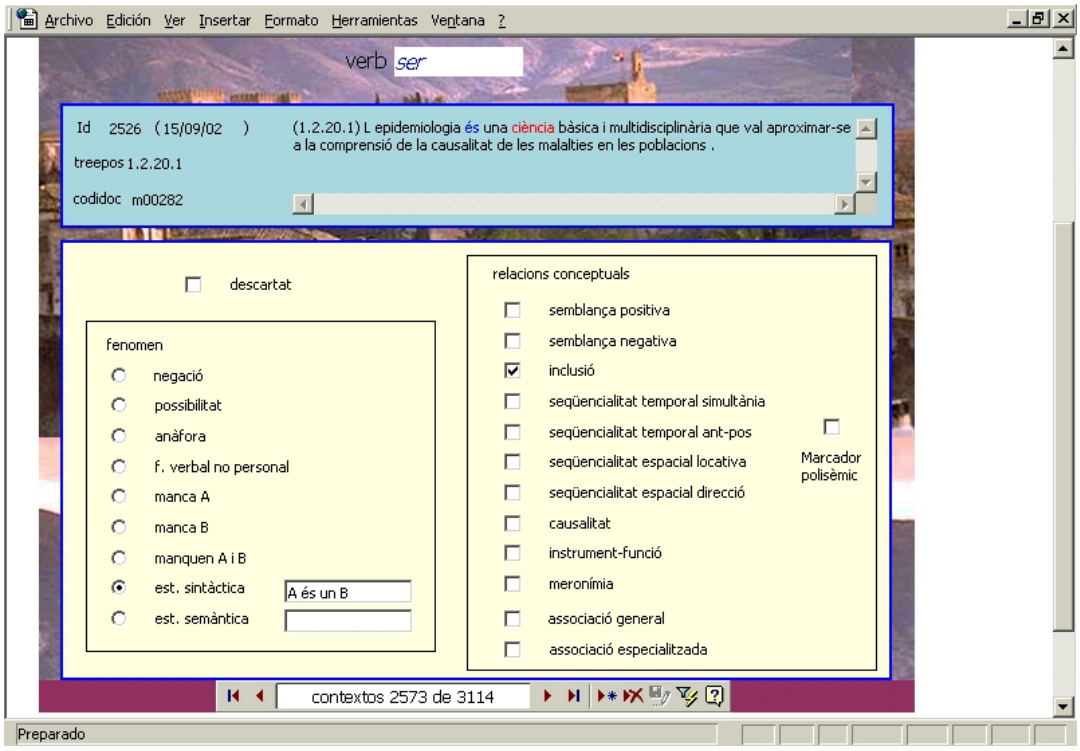


Figura 5-19. Relació d'inclusió expressada per *ser un*.

En aquest cas, desitjablement el sistema detectaria les unitats terminològiques *epidemiologia* i la unitat *ciència* (com ja ha fet), unides pel vincle *ser_un* i

l'etiquetaria, encara que per a nosaltres és clar que es tracta d'una relació d'inclusió, amb les etiquetes de semblança positiva i negativa, inclusió, associació general, seqüencialitat espacial locativa, i, fins i tot, causalitat. Per al cas de la causalitat, el sistema buscaria la unitat nominal *causa* en el context i, en no trobar-la, eliminaria aquesta possibilitat. En el cas de l'etiqueta de seqüencialitat espacial locativa, el sistema també l'eliminaria en no trobar cap element preposicional locatiu (preposicions o locucions preposicionals del tipus *en*, *a prop de*, etc.). Ara bé, el sistema ja no sabria com decidir, de les tres etiquetes que queden quin és el tipus de relació conceptual que efectivament es vehicula en aquest fragment. Fixem-nos en el casos que mostrem seguidament, on el context pot expressar, *a priori*, semblança positiva i també inclusió amb l'estructura *ser_un* idèntica al context anterior.

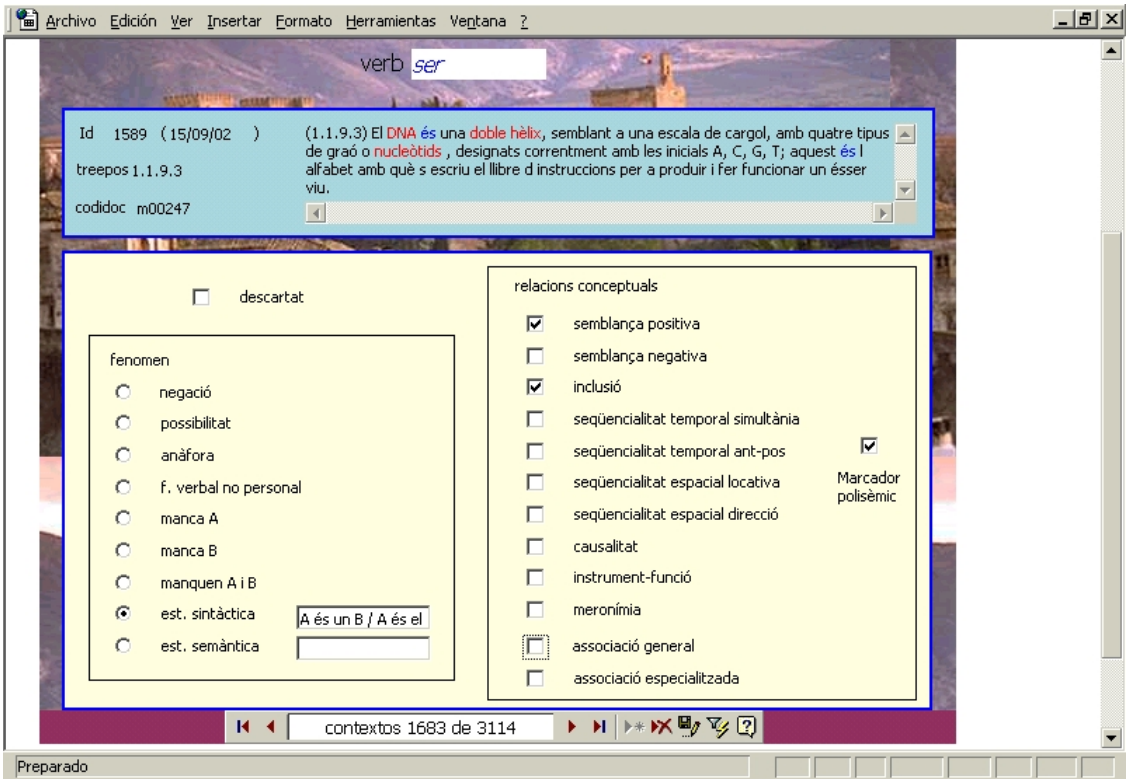


Figura 5-20. Exemple de polisèmia expressada mitjançant el marcador *és un*.

Si aïllem els dos fragments inicials dels dos registres exemplificats trobem que *L'epidemiologia és una ciència* i que *El DNA és una doble hèlix* i, per tant, si l'única relació conceptual que preveiéssim per a *ser_un* fos la d'inclusió, en una

estructuració jeràrquica el concepte superordinat d'*epidemiologia* seria *ciència* però el de *DNA* seria *doble hèlix*.

Per tal de solucionar aquests casos, podem recórrer a l'ontologia, en el nostre cas una ampliació de EuroWordNet amb conceptes sobre genètica i l'ontologia sobre el genoma humà implementada amb OntoTerm, eines que estem complementant en el primer cas i construint en el marc de l'Institut Universitari de Lingüística Aplicada, en el segon. Si busquem a l'ontologia d'OntoTerm trobem, fins al moment, el concepte *DNA* que penja del concepte superordinat *acid* i té com a cohipònim *RNA* i com a subtipus *cDNA*, per exemple. En el cas d'*epidemiologia* i *doble hèlix* es trobem que aquests conceptes encara no s'han introduït. Continuem la cerca a la versió ampliada d'EuroWordNet[11] i trobem que *DNA* equival a *àcid desoxiribonucleic*, que és un *àcid nucleic* i que aquest és un *àcid* i, en la cadena d'hiperonímia, l'*àcid* és un *compost químic*, que és una *substància* que és un *objecte inanimat* que, finalment, penja del concepte *top*, *entity*. En cap cas, doncs, apareix una cadena d'hiperonímia que tingui cap relació amb *doble hèlix*. En canvi, si busquem *doble hèlix* no obtenim cap resultat. Optem, doncs, per buscar *hèlix* i ens apareixen tres camins d'hiperonímia que indiquem amb el signe >:

a) *hèlix>construcció/estructura>artefacte>objecte_inanimat>entitat* (objecte)

b) *hèlix>propulsor>dispositiu_mecànic>mecanisme>utillatje>artefacte>objecte_inanimat>entitat* (objecte)

c) *hèlix>corba>línia>forma* (propietat)

Tenint en compte que el concepte *DNA* només té una cadena d'hiperonímia, el sistema podria deduir que el fragment *El DNA és una doble hèlix* està expressant una relació de semblança, basant-se en la cadena d'hiperonímia *c*) en què s'estableix la indicació d'una propietat. Aquest etiquetatge encara resultaria més fàcil si en

[11] La informació sobre genètica s'està introduint bàsicament en castellà però seria una tasca força abastable introduir els equivalents en català per a què el sistema funcionés en aquestes dues llengües.

l'ontologia sobre el genoma humà implementada en OntoTerm aparegués la relació *similar_to* ja descrita i introduïda en el programa entre *DNA* i *doble hèlix*.

En canvi, per al cas de l'*epidemiologia*, el sistema troba la següent cadena d'hiperonímia:

> *epidemiologia>ciència_mèdica>ciències_biològiques>ciències_naturals>disciplina_*
> *científica>branca_de_coneixement>àrea_de_coneixement>contingut_mental>coneixement/saber>tret_psicològic*

Per tant, la cadena d'hiponímia correspon amb el terme que apareix després del marcador *ser_un* i el sistema etiquetaria el fragment com a relació d'inclusió.

Ocupem-nos ara d'un segon cas on hem etiquetat un mateix context vehiculat pel marcador *incloure* com a possible vehiculador de la relació conceptual d'inclusió i la de meronímia. Partim de la base que no disposem de suficient coneixement especialitzat sobre la matèria, que és com actuaria el sistema de detecció semiautomàtica a partir de la llista de verbs i, per tant, ens veiem obligats a indicar que es tracta d'un marcador polisèmic, com mostra el registre següent:

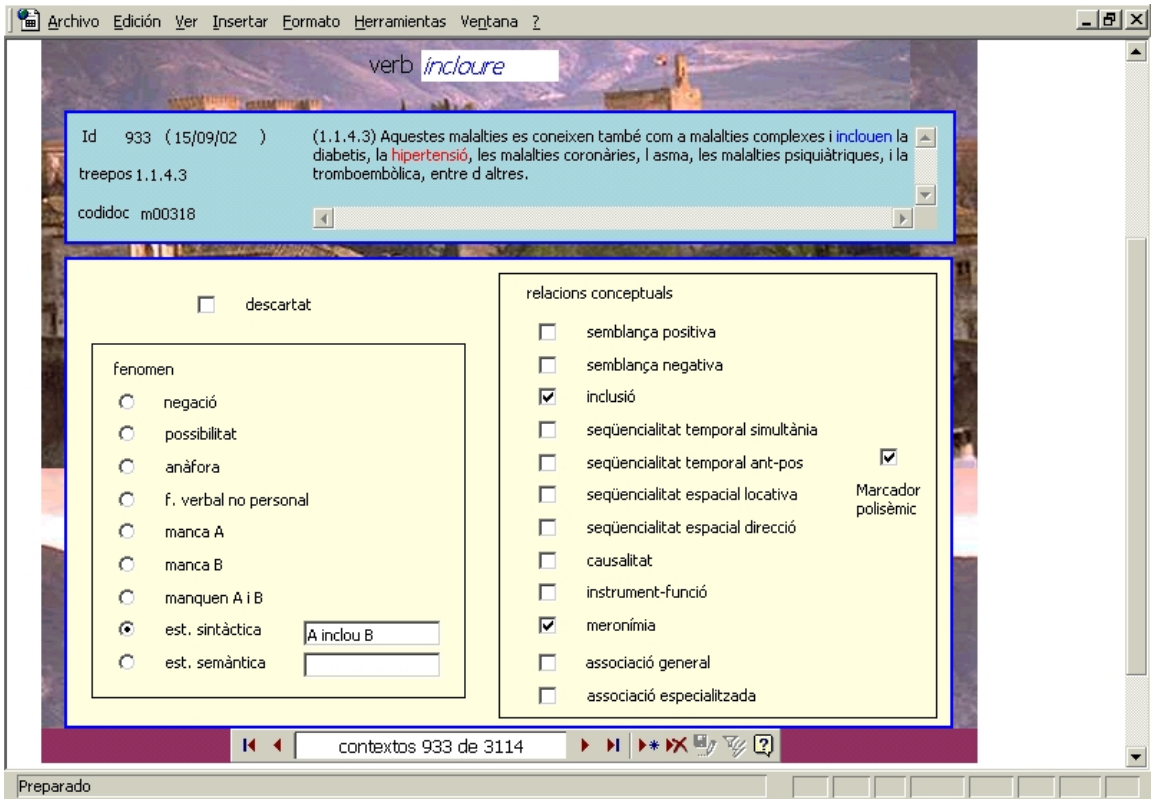


Figura 5-21. Exemple de polisèmia expressada mitjançant el marcador *incloure*.

Procedim de manera anàloga als exemples anteriors i busquem a l'ontologia la cadena d'hiponímia i hiperonímia i la relació de meronímia del terme *malaltia* per tal d'intentar determinar si els termes simples o complexos que apareixen a la banda dreta del marcador[12], *diabetis*, *hipertensió*, *malalties coronàries*, *asma*, *malalties psiquiàtriques* i *tromboembòlica* són parts del concepte *malaltia*, i per tant ens trobaríem amb una relació meronímica o si són subtipus o hipònims i, per tant, tindríem una relació d'inclusió.

Els primers resultats en la cerca de la cadena d'hiponímia troben *asma* com a subtipus de malaltia, fet que podria indicar que la resta d'unitats llistades com a grup

[12] Aquestes unitats es poden detectar automàticament amb el programa d'extracció de candidats a terme *Yate*, implementat per J. Vivaldi. Per a més informació sobre el càlcul del factor de context que ajuda a detectar termes que es troben en contextos terminològicament rics, vegeu VIVALDI, Jorge (2001: 112-113).

són cohipònims d'*asma*. Busquem ara la relació de meronímia a l'ontologia i trobem *mal* i *afecció/dolència*. Per tant, cap de les unitats terminològiques que apareixen a la dreta del marcador *incloure*. Tanmateix, com que la cadena d'hiponímia només ha trobat un element coincident. Proposem al sistema que busqui individualment cadascuna d'aquestes unitats terminològiques per intentar decidir si, efectivament, tenen com a hiperònim el terme *malaltia*. En el moment concret de la nostra recerca, la cerca per a cada cas dóna els següents resultats:

- *diabetis>malaltia_glandular>desordre/malestar/trastorn>condició/situació >estat*
- *hipertensió>malaltia_cardiovascular> desordre/malestar/trastorn>condició/situació>estat*
- *malalties coronàries* (no hi ha resultat tot i que sí apareix *malaltia_cardiovascular*)
- *asma>malaltia_respiratòria>mal/malaltia>problema_de_salut>estat_fisiològic> condició/situació>estat*
- *malalties psiquiàtriques* (no hi ha resultat)
- *tromboembòlica* (el sistema no troba aquest adjectiu aïllat en la cerca)

Els resultats mostren que, majoritàriament, les unitats terminològiques que pareixen a la dreta del marcador verbal constitueixen semànticament un grup de cohipònims de l'hipònim *malaltia* que ocorre a l'esquerra del vincle. Per aquest motiu, el sistema disposaria d'informació suficient per a decidir que la relació conceptual expressada en aquest fragment textual és la relació d'inclusió.

5.5 A mode de síntesi

Abans de tancar definitivament aquest capítol, voldríem presentar una síntesi dels aspectes o idees més rellevants que hem tractat en els diversos apartats i que

esdevindran elements clau en la descripció del prototip de sistema de detecció semiautomàtica de relacions conceptuals que presentem en el capítol següent.

En primer lloc, voldríem destacar que hem intentat integrar en les estratègies de detecció de relacions conceptuals totes les eines al nostre abast disponibles en el marc de l'IULA. Així, al marge del Corpus Textual, volem destacar la utilització del programa *Mercedes*, l'ontologia sobre el genoma humà que s'està construint amb l'eina OntoTerm i, finalment, l'adaptació d'EuroWordNet per a l'àmbit del genoma i una primera exploració amb el programa *Yate* d'extracció de termes. Creiem que un treball que vulgui ser efectiu i realista ha de basar-se en els recursos que es troben realment disponibles i intentar, a més, proporcionar noves informacions que els puguin enriquir.

Dit això, voldríem resumir les idees del capítol en quatre grans blocs. Primerament, hem establert els criteris que ens duran a descartar una unitat verbal inicialment possible vehiculadora de relació conceptual sobre la base de l'anàlisi manual del corpus del nostre corpus. Aquests paràmetres (negació, possibilitat, anàfora, forma verbal no personal, manca d'A, manca de B i manca d'A i B) han estat descrits i exemplificats al llarg d'aquestes pàgines.

En segon lloc, volem fer èmfasi en el fet que la informació sintàctica i semàntica s'ha recollit de manera inicial, com a un primer estadi de la recerca, per tal de facilitar el refinament posterior del sistema semiautomàtic però que, en un primer moment, no han estat criteris decisius per a descartar o retenir una determinada unitat verbal.

En tercer lloc, els resultats numèrics ens permeten constatar que alguns elements verbals mostren percentatges de precisió força elevats mentre que d'altres semblen comportar un soroll més elevat.

Finalment, volem destacar la novetat que representa integrar la informació semàntica, a més dels patrons sintàctics, vehiculada a través de l'ús d'una ontologia, a un sistema de detecció inicialment basat en la comparació de cadenes de caràcters a partir d'una llista prèvia. Creiem que aquest recurs semàntic esdevindrà la clau per a

la majoria de decisions que el sistema haurà de prendre per refinar al màxim els seus resultats.

Capítol 6

Proposta de sistema de detecció semiautomàtica de relacions conceptuals

Capítol VI

6 Proposta de sistema de detecció semiautomàtica de relacions conceptuals

6.1 Introducció

En aquest últim capítol, que precedeix les conclusions i les futures vies de recerca que se'n deriven, presentem una proposta de sistema de detecció semiautomàtica de relacions conceptuals. Com ja hem dit a la introducció d'aquest treball el perfil de l'autora cobreix l'àmbit lingüístic però, més difícilment, els coneixements informàtics necessaris per poder implementar una proposta d'aquesta índole. Per aquest motiu, aquest capítol recull de la manera més sistematitzada possible l'esquema d'implementació que hauria de seguir el sistema de detecció semiautomàtica de relacions conceptuals.

Per tal d'assolir aquest grau màxim de sistematització proposem la utilització d'un arbre de decisions, complementat a partir del recurs a d'altres eines. Les condicions de l'arbre de decisions i els recursos a altres eines provenen directament de l'estudi sobre les estratègies de detecció descrit àmpliament en el capítol anterior.

Així, i més concretament, aquest capítol defineix, en primer lloc, què entenem per arbre de decisions i justifica per què hem escollit aquesta opció de sistematització de la informació de cara a proposar un prototip de sistema de detecció que s'acosti al

màxim a la possible aplicació final de la tasca teòrica i metodològica que hem dut a terme.

Finalment, tancarem el capítol amb una descripció de la viabilitat de la proposta tenint en compte, principalment, de quines eines i recursos disposem en el marc de l'IULA per tal de poder implementar una proposta de detecció de relacions conceptuals que, de moment, haurà de ser validada per l'usuari quan n'obtingui els resultats.

6.2 Sistema de detecció semiautomàtica de relacions conceptuals

6.2.1 Què és un arbre de decisions?

A partir de la bibliografia explorada per tal de decidir si un arbre de decisions ens permetria d'esquematitzar la informació de què disposem, hem escollit les següents definicions d'arbre de decisions que creiem que, per la seva simplicitat, facilitaran la nostra tasca. Entenem que un arbre de decisions és, seguint la definició de Murthy, K.V.S (1995, cap. 1):

> «***decision trees*** *are a way to represent rules underlying data. Decision trees are hierarchical, sequential classification structures that recursively partition the set of observations (data). (...) Each node of the decision tree consists of either a test that partitions the data, or a decision about the object. Once a tree is constructed from data, it can be used to classify objects of unknown category».*

Com veiem en aquesta definició, els arbres de decisió permeten representar regles a partir de les dades, propietat que s'adequa perfectament a les nostres necessitats. I, a més, un arbre de decisions es pot utilitzar per classificar objectes de categoria desconeguda que és, en la majoria dels casos, la situació en què es troben els marcadors verbals que expressen una o més d'una relació conceptual i dels quals no

sabem amb exactitud la categoria, entesa com el tipus de relació conceptual que vehiculen.

Un arbre de decisions consta de nodes, on es representen els atributs d'entrada, i els arcs amb els diferents valors que aquests nodes poden adquirir. Un arbre de decisions també es pot convertir en regles expressades mitjançant *if-then*.

La bibliografia sobre els arbres de decisions és força extensa i variada i, arriba a nivells de detall i combinació d'estratègies que queden lluny del propòsit d'aquest treball. Els arbres de decisions s'estan utilitzant eficaçment per implementar diversos formalismes en el camp de l'aprenentatge automàtic (*machine learning*), principalment pel que fa a tasques bàsiques del processament del llenguatge natural com són el reconeixement de la parla, l'etiquetatge morfològic i sintàctic, i la desambiguació semàntica, entre d'altres. En el nostre cas, utilitzem la noció d'arbre de decisió per sistematitzar la informació rellevant que hem descrit al llarg del capítol cinquè i per facilitar la futura implementació informatitzada de la nostra proposta.

6.2.2 Per què un arbre de decisions?

Reprenent el que acabem de dir, considerem que un arbre de decisions que, a més, es pot reconvertir en regles del tipus *if-then* representa un mitjà per estructurar la informació de la manera més propera possible a com entenen la informació les màquines. Per aquest motiu, aquest últim capítol recull l'esforç de sistematització de tota la informació que hem anat esbossant al llarg de totes aquestes pàgines.

A més, un arbre de decisions permet estructurar seqüencialment la informació de manera privilegiada. Amb això volem dir que s'estableixen diversos nivells i que segons quin sigui el tipus d'informació resultant en els primers nodes, el sistema ja no avançarà més en la possible detecció d'una relació conceptual.

Finalment, un arbre de decisions com a punt d'inici del sistema no impedeix, afortunadament, poder combinar altres estratègies o recursos al llarg del procés de detecció de relacions conceptuals. Aquesta possibilitat d'ordenació de la informació i d'integració i combinació d'altres recursos com són un extractor de terminologia i el

recurs semàntic a l'ontologia que ja hem esmentat anteriorment permeten completar el procés de detecció semiautomàtica, o si es prefereix assistida, de relacions conceptuals.

6.2.3 Arbre de decisions i combinació de recursos per a la construcció del prototip de sistema de detecció semiautomàtica de relacions conceptuals

En primer lloc, presentem un esquema en forma d'arbre molt simplificat de l'estructura de funcionament del prototip i, seguidament, anirem desglossant en forma de regles *if-then*, que es convertirà en el nostre cas en *si* x, *aleshores* y.

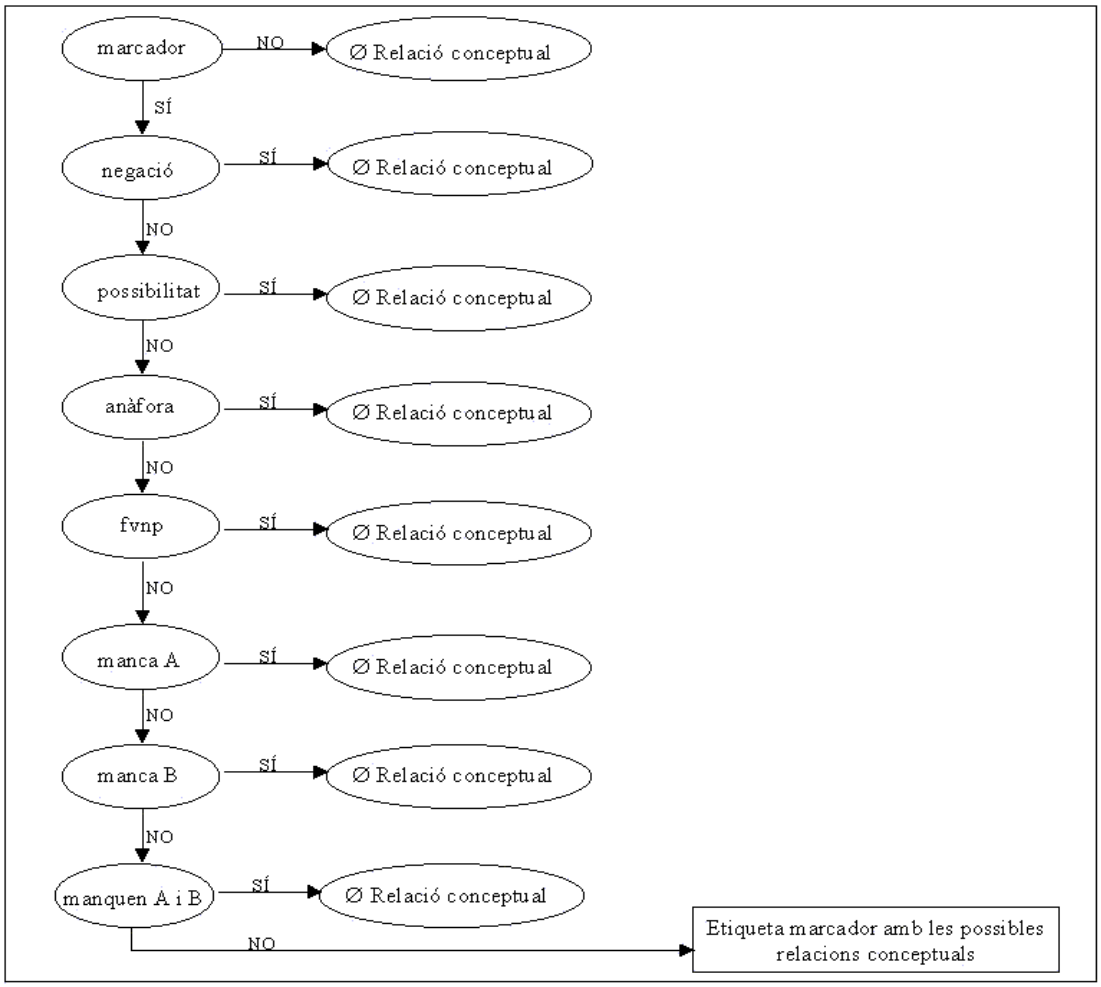


Figura 6-1. Estructura de l'arbre de decisions.

Aquest arbre de decisions es llegeix de la manera següent:

- *Si* troba un marcador de la llista predefinida en el programa, *aleshores* passa al node següent. *Si* no troba cap dels marcadors de la llista predefinida en el programa *aleshores* no hi ha relació conceptual.

- *Si* troba algun element de negació de la llista predefinida en el programa (adverbis de negació) en un context de com a màxim 3 elements a l'esquerra del marcador, *aleshores* no hi ha relació conceptual. *Si* no troba cap dels marcadors de la llista predefinida en el programa, *aleshores* passa al node següent.

- *Si* troba algun element de possibilitat de la llista predefinida en el programa (principalment el verb poder) en un context de com a màxim 3 elements a l'esquerra del marcador o bé troba el marcador en forma verbal condicional, *aleshores* no hi ha relació conceptual. *Si* no troba cap dels marcadors de la llista predefinida en el programa, *aleshores* passa al node següent.

- *Si* troba algun element anafòric de la llista predefinida en el programa (pronoms relatius) en un context de com a màxim 5 elements a la dreta i a l'esquerra del marcador, *aleshores* no hi ha relació conceptual. *Si* no troba cap dels marcadors de la llista predefinida en el programa, *aleshores* passa al node següent.

- *Si* detecta que el marcador es troba en forma verbal no personal, infinitiu o gerundi, *aleshores* no hi ha relació conceptual. *Si* no troba el marcador en cap d'aquestes dues formes verbals no personals, *aleshores* passa al node següent.

- *Si* no troba cap unitat terminològica a l'esquerra del marcador d'entre totes les unitats fins a la primera paraula del context d'anàlisi, *aleshores* no hi ha relació conceptual. *Si* troba alguna unitat terminològica a l'esquerra del marcador, *aleshores* passa al node següent.

- *Si* no troba cap unitat terminològica a la dreta del marcador d'entre totes les unitats fins a l'última paraula del context d'anàlisi, *aleshores* no hi ha relació

conceptual. *Si* troba alguna unitat terminològica a la dreta del marcador, *aleshores* passa al node següent.

- *Si* no troba cap unitat terminològica ni a l'esquerra ni a la dreta del marcador d'entre totes les unitats que formen el context d'anàlisi, *aleshores* no hi ha relació conceptual. *Si* troba com a mínim una unitat terminològica a l'esquerra i a la dreta del marcador, *aleshores* etiqueta el marcador amb totes les possibles indicacions de relació conceptual d'aquest marcador.

Fins a aquesta etapa, volem destacar dues qüestions pel que fa al mode de funcionament del programa. En cadascun dels nodes resultants obtenim una llista de contextos descartats de manera automàtica pel sistema. Proposem que el programa guardi un arxiu amb els contextos descartats per a cadascuna de les etapes que ha anat passant en el procés d'anàlisi. Per tant, en la fase final d'aquest arbre de decisions disposarem d'un arxiu "nomarcador" on trobarem tots els contextos que no tenen un marcador verbal coincident amb els de la llista inicial però que podrem revisar per tal d'obtenir nous marcadors que expressin alguna de les relacions conceptuals que proposem en nostra tipologia. També disposarem d'un arxiu "negació", on podrem observar si apareixen nous indicadors d'aquest fenomen i si efectivament la detecció ha estat correcta. Tindrem dos arxius, "possibilitat" i "anàfora" que ens permetran de verificar el funcionament del sistema de detecció i, finalment, tindrem tres arxius per separat ("mancaa", "mancab" i mancaab") que podrem analitzar per tal de detectar si apareixen unitats terminològiques que el sistema no ha pogut reconèixer a partir de les eines d'extracció de terminologia amb què treballa.

Així, el sistema basat en l'arbre de decisions ens permet obtenir un conjunt de marcadors verbals que queden retinguts per tal de passar a l'establiment del tipus de relació conceptual que vehicula cadascun dels marcadors. En aquest sentit, i després d'haver establert les condicions negatives per tal de descartar un determinat marcador en un context especialitzat determinat, el sistema estableix com a condicions positives en primer lloc el patró sintàctic que es manifesta en un determinat context i, seguidament, recorre al recurs semàntic de l'ontologia per tal de

determinar amb el màxim de precisió possible la relació conceptual expressada pel marcador en un determinat fragment de coneixement especialitzat.

En aquest sentit, i conscients que els contextos d'anàlisi inicials de què partim apareixen amb unitats terminològiques marcades a partir de diccionaris en català però el nombre d'unitats que conté el diccionari de base és força reduït, volem indicar que la revisió manual d'alguns contextos ens pot dur a engruixir el nombre d'entrades del diccionari. Tanmateix, creiem essencial de combinar, en aquesta etapa de detecció de les unitats terminològiques el sistema d'extracció de candidats a terme *Yate*[1], implementació de la tesi doctoral duta a terme per J. Vivaldi (2001), que permet combinar estratègies heterogènies per obtenir candidats a terme.

En el marc de la tesi, i per demostrar que l'extractor de terminologia és una ajuda efectiva en la detecció de candidats a terme en contextos especialitzats, hem aplicat inicialment *Yate* a tres dels documents que formen part del nostre corpus i als quals pertanyen els contextos que hem comentat extensament en l'apartat del recurs semàntic de l'ontologia. Els documents sobre els quals hem aplicat aquest programa d'extracció de candidats a termes són m00282, m00247 i m00318 (que corresponen als contextos de la Figura 5.18, Figura 5.19 i Figura 5.20). En aquests casos, per tal de desambiguar el marcador de relació conceptual hem partit d'unitats terminològiques detectades per *Mercedes* i per una detecció manual d'altres unitats. En aplicar *Yate* a aquests contextos hem ampliat el nombre d'unitats terminològiques detectades automàticament, proposades pel sistema com a candidats a termes amb un alt grau de terminologibilitat o *termhood*.

Així, i a tall d'exemple, per al context de la Figura 5.18:

L'epidemiolocia és una ciència bàsica i multidisciplinària que val aproximar-se a la comprensió de la causalitat de les malalties en les poblacions.

Yate detecta com a candidats a termes *epidemiologia* i *ciència bàsica*.

[1] Per a més informació sobre el funcionament d'aquest programa podeu consultar el manual d'ús intern VIVALDI, Jorge (2003b) *Sistema de extracción de Candidatos a Término* YATE. *Manual de utilización*. Barcelona: Institut Universitari de Lingüística Aplicada.

Per al context de la Figura 5.19:

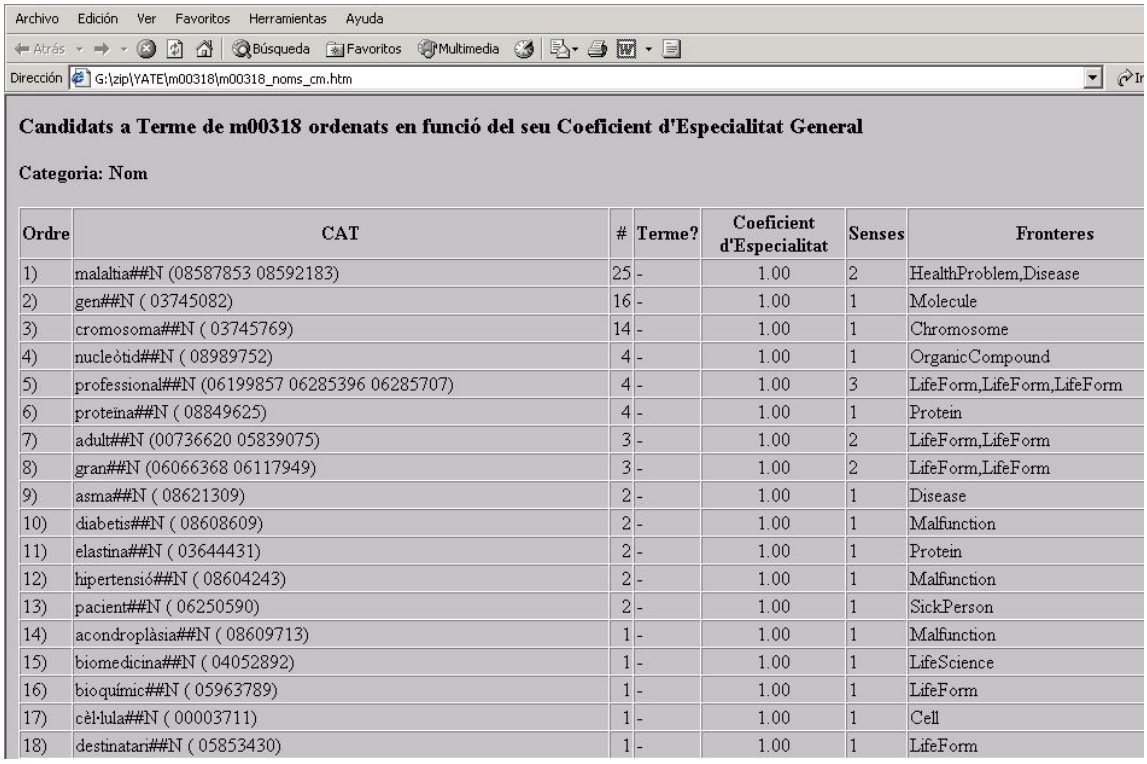
El DNA és una doble hèlix, semblant a una escala de cargol, amb quatre tipus de graó o nucleòtids, designats correntment amb les inicials A, C, G, T; aquest és l'alfabet amb què s'escriu el llibre d'instruccions per a produir i fer funcionar un ésser viu.

El sistema de detecció de candidats a terme detecta les mateixes unitats que *Mercedes* i, a més, la unitat complexa *ésser viu*.

I finalment, per al context de la Figura 5.20, l'extractor ens indica:

Aquestes malalties es coneixen també com a malalties complexes i inclouen la diabetis, la hipertensió, les malalties coronàries, l'asma, les malalties psiquiàtriques i la tromboembòlica entre d'altres.

no només *hipertensió*, coincident amb el resultat de *Mercedes*, sinó també *malalties*, *diabetis*, *hiertensió*, *malaltia coronària*, *asma* i *malaltia psiquiàtrica*, per tal com els adjectius especialitzats es troben introduïts com a base per poder detectar patrons nom + adjectiu com a candidats a unitats terminològiques. Vegem-ne els resultats en les figures següents:



Archivo Edición Ver Favoritos Herramientas Ayuda

Atrás Búsqueda Favoritos Multimedia

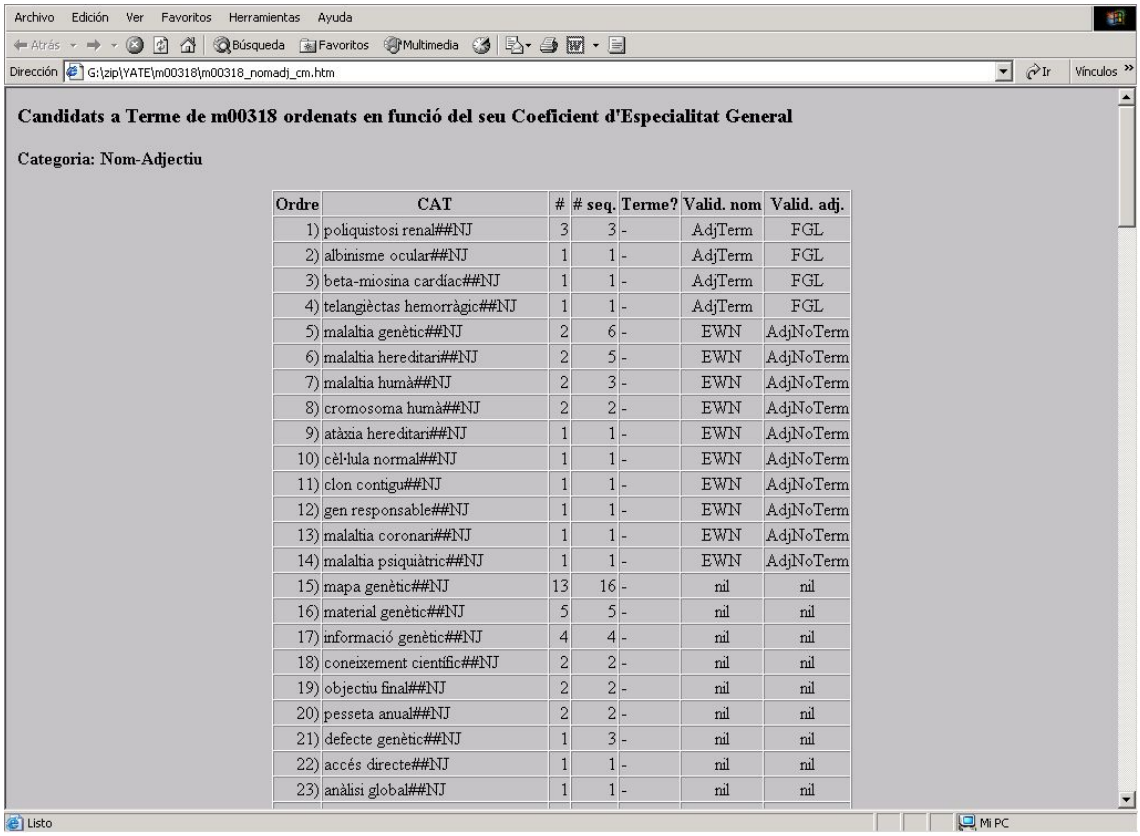
Dirección G:\zip\YATE\m00318\m00318_noms_cm.htm

Candidats a Terme de m00318 ordenats en funció del seu Coeficient d'Especialitat General

Categoria: Nom

Ordre	CAT	#	Terme?	Coeficient d'Especialitat	Senses	Fronteres
1)	malaltia##N (08587853 08592183)	25	-	1.00	2	HealthProblem,Disease
2)	gen##N (03745082)	16	-	1.00	1	Molecule
3)	cromosoma##N (03745769)	14	-	1.00	1	Chromosome
4)	nucleòtid##N (08989752)	4	-	1.00	1	OrganicCompound
5)	professional##N (06199857 06285396 06285707)	4	-	1.00	3	LifeForm,LifeForm,LifeForm
6)	proteïna##N (08849625)	4	-	1.00	1	Protein
7)	adult##N (00736620 05839075)	3	-	1.00	2	LifeForm,LifeForm
8)	gran##N (06066368 06117949)	3	-	1.00	2	LifeForm,LifeForm
9)	asma##N (08621309)	2	-	1.00	1	Disease
10)	diabetis##N (08608609)	2	-	1.00	1	Malfunction
11)	elastina##N (03644431)	2	-	1.00	1	Protein
12)	hipertensió##N (08604243)	2	-	1.00	1	Malfunction
13)	pacient##N (06250590)	2	-	1.00	1	SickPerson
14)	acondroplàsia##N (08609713)	1	-	1.00	1	Malfunction
15)	biomedicina##N (04052892)	1	-	1.00	1	LifeScience
16)	bioquímic##N (05963789)	1	-	1.00	1	LifeForm
17)	cèl·lula##N (00003711)	1	-	1.00	1	Cell
18)	destinatari##N (05853430)	1	-	1.00	1	LifeForm

Figura 6-2. Aplicació de *Yate* al document m00318. Visualització de la categoria nom.



Archivo Edición Ver Favoritos Herramientas Ayuda

Atrás Búsqueda Favoritos Multimedia

Dirección G:\zip\YATE\m00318\m00318_nomadj_cm.htm Ir Vínculos

Candidats a Terme de m00318 ordenats en funció del seu Coeficient d'Especialitat General

Categoria: Nom-Adjectiu

Ordre	CAT	#	# seq.	Terme?	Valid. nom	Valid. adj.
1)	poliquistosi renal##NJ	3	3	-	AdjTerm	FGL
2)	albinisme ocular##NJ	1	1	-	AdjTerm	FGL
3)	beta-miosina cardíac##NJ	1	1	-	AdjTerm	FGL
4)	telangièctas hemorràgic##NJ	1	1	-	AdjTerm	FGL
5)	malaltia genètic##NJ	2	6	-	EWN	AdjNoTerm
6)	malaltia hereditari##NJ	2	5	-	EWN	AdjNoTerm
7)	malaltia humà##NJ	2	3	-	EWN	AdjNoTerm
8)	cromosoma humà##NJ	2	2	-	EWN	AdjNoTerm
9)	atàxia hereditari##NJ	1	1	-	EWN	AdjNoTerm
10)	cèl·lula normal##NJ	1	1	-	EWN	AdjNoTerm
11)	clon contigu##NJ	1	1	-	EWN	AdjNoTerm
12)	gen responsable##NJ	1	1	-	EWN	AdjNoTerm
13)	malaltia coronari##NJ	1	1	-	EWN	AdjNoTerm
14)	malaltia psiquiàtric##NJ	1	1	-	EWN	AdjNoTerm
15)	mapa genètic##NJ	13	16	-	nil	nil
16)	material genètic##NJ	5	5	-	nil	nil
17)	informació genètic##NJ	4	4	-	nil	nil
18)	coneixement científic##NJ	2	2	-	nil	nil
19)	objectiu final##NJ	2	2	-	nil	nil
20)	pesseta anual##NJ	2	2	-	nil	nil
21)	defecte genètic##NJ	1	3	-	nil	nil
22)	accés directe##NJ	1	1	-	nil	nil
23)	anàlisi global##NJ	1	1	-	nil	nil

Listo Mi PC

Figura 6-3. Aplicació de *Yate* al document m00318. Visualització de la categoria NA.

Queda demostrat, doncs, que la combinació de *Mercedes* i *Yate* en la detecció d'unitats terminològiques que constitueixen els elements *a*, *b* i *n* de la nostra estructura de partida per a l'establiment d'una relació conceptual dóna uns resultats força millorats i facilita quant a la rapidesa la tasca de detecció d'unitats terminològiques en un fragment de coneixement especialitzat.

En un futur esperem, també, poder utilitzar la base de dades sobre el genoma humà, lligada a una ontologia sobre aquesta mateixa temàtica, com a font per detectar unitats terminològiques en el corpus textual. Totes dues eines, *Yate* i la base de dades de genoma humà, treballen sobre la base d'una ontologia i, per tant, la detecció de les unitats terminològiques comporta també la inclusió de la informació semàntica basada en una ontologia que servirà al sistema, en una etapa final, per intentar desambiguar els marcadors polisèmics.

A partir dels resultats que contenen els contextos etiquetats amb les possibles indicacions d'una o més d'una relació conceptual, el programa pot intentar integrar una mínima informació sobre els patrons sintàctics, basada eminentment en l'aparició d'una determinada preposició després del marcador, per refinar els resultats (per exemple, l'aparició de la preposició *mitjançant* o *gràcies a* comporta, prototípicament, la vehiculació de la relació conceptual instrument-funció).

En darrer terme, el programa recorrerà a la semàntica de les unitats terminològiques, sobre la base de l'ontologia, per tal de decidir quina és amb major possibilitat la relació conceptual expressada mitjançant un marcador verbal en un context determinat. En aquest sentit, el sistema treballarà tal com s'indica en el capítol 5 fent servir el recurs semàntic de l'ontologia, i intentant determinar a quina categoria semàntica pertany la unitat, per tal de decidir quin tipus de relació conceptual s'està expressant. Recordem que podem utilitzar aquest recurs principalment per als casos de polisèmia entre les relacions conceptuals de semblança i inclusió; d'inclusió i meronímia, i de seqüencialitat espacial i temporal.

En resum, doncs, el sistema treballarà sobre un conjunt de regles que segueixen l'esquema de l'arbre de decisions que hem presentat però integrarà, al mateix temps, un sistema d'extracció de candidats a unitats terminològiques, una anàlisi sintàctica mínima basada en la coaparició del marcador i una determinada preposició o locució preposicional que pugui comportar un refinament de les dades i una anàlisi semàntica basada en una ontologia, entesa com a eina d'ajuda a la desambiguació de les unitats verbals que, en expressar relacions conceptuals, esdevenen marcadors polisèmics.

6.3 Viabilitat de la proposta

Atès que no podem donar resultats del funcionament del sistema de detecció semiautomàtica de relacions conceptuals per tal com no hem pogut implementar-ne un prototip i provar-lo, volem indicar els motius pels quals creiem viable aquesta proposta en una feina futura no gaire llunyana.

En primer lloc, el sistema haurà de treballar sobre documents marcats estructuralment i morfològicament. En aquest sentit, disposem d'un corpus textual multilingüe sobre cinc àrees diferents del coneixement especialitzat que esdevindran una font per al processament de les dades essencial i, a més, serviran com a banc de proves no només per als textos de l'àrea temàtica del genoma o la medicina, amb els quals hem treballat fins al moment, sinó que també ens permetrà ampliar la recerca a d'altres àrees temàtiques o facilitar la feina a recerques futures.

En segon lloc, disposem de dues eines en funcionament i amb resultats força satisfactoris que són els programes *Mercedes* i *Yate* que ja hem descrit. Aquests dos programes combinats, juntament amb la possibilitat d'integrar la base de dades sobre el genoma humà que estem constituint actualment com a projecte de recerca, ens permetrà incrementar el nombre d'unitats candidates a ser unitat terminològica. Creiem que el sistema de detecció semiautomàtica de relacions conceptuals pot, d'una banda, beneficiar-se dels resultats del programa *Yate* però, de l'altra, pot ser útil la seva implementació en l'estratègia del càlcul del factor del context (Vivaldi, 2001: 112-113) que ja hem esmentat.

Més concretament, aquesta estratègia es basa en la cerca per part de l'extractor de candidats a terme que es troben en contextos terminològicament rics, com era el cas de l'exemple de polisèmia del marcador *incloure* en el capítol anterior (*Aquestes malalties es coneixen també com a malalties complexes i inclouen la diabetis, la hipertensió, les malalties coronàries, l'asma, les malalties psiquiàtriques, i la tromboembòlica, entre d'altres*). Doncs bé, hem observat que, d'una banda, l'extractor ens pot proposar, *a priori*, unitats terminològiques com *diabetis*, *hipertensió*, *malaltia coronària*, *asma* i *malaltia psiquiàtrica*, però l'existència d'un marcador de relació conceptual vehiculant un nombre *n* d'unitats, algunes detectades com a terminològiques però d'altres no ([malaltia] *tromboembòlica*), pot ser un paràmetre més a tenir en compte en el funcionament de l'extractor[2].

[2] Aquesta possible aplicació del sistema de detecció de relacions conceptuals a l'extractor de terminologia pel que fa al càlcul del factor del context ha estat debatuda i acceptada per l'autor del programa i es troba pendent d'implementació a partir de la finalització d'aquest treball.

En l'anàlisi del context, si el sistema s'enfronta a un context terminològicament ric pel que fa a les unitats, però aquest context és, a més, terminològicament ric pel que fa a la relació conceptual que vehicula aquestes unitats, estem convençuts que la integració de la informació de la relació conceptual pot atorgar a les unitats no detectades en un primer moment un pes terminològic més elevat que serveixi per proposar-les com a candidates a unitats terminològiques en una segona fase.

Finalment, també disposem de l'ontologia corresponent al genoma humà, estretament lligada amb la base de dades terminològiques que ja hem esmentat i, de manera més útil i remarcable, tenim a l'abast l'ontologia d'EuroWordNet, principalment la versió modificada per tal de contenir informació sobre el genoma humà.

Aquests tres blocs de recursos i eines ens permeten aventurar que el sistema de detecció semiautomàtica de relacions conceptuals no és una proposta basada només en la il·lusió sinó que partirà d'una feina feta fins al moment i materialitzada en uns recursos en funcionament per tal de cobrir un aspecte més del tractament del coneixement especialitzat que fins ara no s'ha pogut detectar de manera automatitzada. Pensem que la suma de dues eines, una orientada a la detecció de candidats a termes i l'altra a la detecció de possibles relacions entre candidats a termes, ens ha de dur a obtenir un primer esquelet del coneixement especialitzat contingut en un text i, per tant, aquest coneixement podrà ser reaprofitat en la compleció del mòdul lèxic d'una base de coneixement especialitzada per la temàtica lligada a una ontologia.

Capítol 7

Conclusions

Capítol VII

7 Conclusions

En aquest capítol final recollim a mode de síntesi les conclusions que es desprenen del treball que hem realitzat. A més, presentem quines són les aportacions d'aquesta tesi doctoral en la recerca sobre terminologia i coneixement especialitzat que estem duent a terme en el marc de l'IULA i, més concretament, en el si del grup IULATERM. Finalment, indiquem algunes de les futures vies de recerca per les quals en agradaria continuar a partir de la conclusió d'aquesta tesi doctoral.

7.1 Conclusions generals

En aquest apartat presentem les conclusions generals a què hem arribat a partir del desenvolupament de cadascun dels capítols d'aquest treball. En primer lloc, creiem interessant de recordar que aquest treball té els seus orígens en un treball de recerca previ i que un dels objectius inicials, orientat a validar la tipologia de relacions conceptuals que havíem establert prèviament, ha estat acomplert de manera satisfactòria. Així, hem continuat treballant sobre la base de la nostra definició de relació conceptual, fet que ens ha permès dur a terme la recerca de manera sistemàtica i, alhora, ens ha portat a la validació de la tipologia inicial de relacions conceptuals de què partíem tot i que haguem canviat l'àrea temàtica del corpus d'on hem extret totes les dades que hem analitzat.

En aquest sentit, hem observat que la majoria de marcadors verbals que expressaven una determinada relació no són temàticament dependents sinó que s'han mantingut tot i el canvi d'àmbit temàtic i l'ampliació del corpus de treball. Destaquem, doncs,

que existeix una llista tancada de tipus de relacions conceptuals que funciona i es materialitza en el discurs especialitzat i que són les diverses manifestacions de cadascuna de les relacions conceptuals, és a dir, els diversos marcadors verbals els que són diversos per tal d'expressar cada relació. De la llista de marcadors verbals inicials de què partíem alguns es mantenen i expressen sempre una o més d'una relació conceptual i d'altres tenen una major presència depenent de l'àrea temàtica objecte d'estudi. Aquesta distinció entre la tipologia de relacions conceptuals i els marcadors verbals que la vehiculen fomenta encara més la idea que presentàvem en la introducció d'aquest treball sobre la separació de la noció de relació conceptual i la unitat lingüística explícita que la materialitza. En aquest sentit, creiem que ha quedat constatat al llarg del treball que les unitats verbals que expressen cadascun dels diferents tipus de relació conceptual quedarien recollides en un mòdul lèxic mentre que la noció o tipus de relació conceptual, més en abstracte, quedaria recollida com un concepte del mòdul ontològic en un sistema de representació de la informació especialitzada que contingui dos mòduls, un de semàntic o ontològic, i un de lèxic amb les unitats terminològiques separades d'una banda i els marcadors de relació conceptual de l'altra, però sempre amb els dos mòduls estretament relacionats.

És per aquest motiu que el capítol tercer del treball descriu abastament el tractament que s'ha donat a les relacions conceptuals en cinc ontologies diferents que s'han utilitzat, i s'utilitzen encara *a posteriori*, per a recerques més àmplies. Hem demostrat que, en la majoria de casos, el tractament de les relacions conceptuals en les ontologies és insuficient. Considerem que és insuficient en dos aspectes, o bé perquè no hi ha un estudi previ teòric sobre les relacions conceptuals i s'integren en l'ontologia les relacions que tradicionalment s'han tractat sense cap mena d'estructuració interna; o bé perquè es confon la noció de relació conceptual amb les unitats de la llengua que les expressen i això comporta repeticions d'alguns dels tipus de relació conceptual. En el nostre cas, i volem destacar aquesta feina que ja hem explicat en el capítol tercer i que valida la hipòtesi de partida, hem implementat per separat la noció de tipus de relació conceptual en el mòdul ontològic i la unitat vehiculadora de relació conceptual en una part del mòdul lèxic. Fins al moment, la indicació de relacions conceptuals entre conceptes funciona plenament i dóna compte

de les relacions conceptuals que efectivament es donen entre els conceptes que s'han introduït en l'ontologia. A més, la separació entre la tipologia de relacions conceptuals i els marcadors verbals evita repeticions en el mòdul ontològic i una millor classificació conceptual de les unitats verbals que expressen cada tipus de relació conceptual.

La constitució i anàlisi de les dades del capítol quart ens ha servit per incrementar i, al mateix temps refinar, la llista de marcadors verbals de què partíem i que esdevenen el punt de partida per a l'establiment d'estratègies que permetin detectar relacions conceptuals que apareixen en context. L'objectiu d'aquestes conclusions no és repetir cadascuna de les estratègies, però sí deixar constància de la innovació que representa integrar els paràmetres que hem recollit a la base de dades de marcadors verbals de relació conceptual per tal de proposar un sistema de marcatge i posterior detecció semiautomàtica de relacions conceptuals. Aquesta innovació o novetat se centra, d'una banda, en la integració de paràmetres contextuals que es poden detectar a partir de l'anàlisi morfològica dels contextos aïllats (negació, possibilitat, etc.) i, de l'altra, en el recurs dels patrons sintàctics i la informació semàntica de l'ontologia que es pot usar per millorar els resultats quant a l'etiquetatge de marcador polisèmic d'una unitat. Creiem que aquesta combinació d'estratègies representa un dels aspectes més destacables d'aquest treball.

Finalment, tot i ser conscients que no podem donar resultats sobre el funcionament del sistema de detecció semiautomàtica de relacions conceptuals per tal com la implementació s'haurà de fer en una etapa posterior de la recerca, creiem que l'esforç de sistematització i modelització de les dades facilitaran enormement la tasca de materialització de les regles derivades de l'arbre de decisions en un programa informàtic que donarà resultats que es podran reutilitzar per retroalimentar l'ontologia i la base de dades i, també l'extractor de terminologia i el mapeig del coneixement especialitzat contingut en un text.

7.2 Aportacions de la tesi doctoral

Des del nostre punt de vista, una tesi doctoral ha d'aportar alguna novetat a la recerca que s'ha fet fins al moment. En el nostre cas, creiem que aquest objectiu s'acompleix en els aspectes següents:

- Hem establert una nova definició de la noció de relació conceptual (*a* R *b*, *n)* basada en la realitat que representa treballar amb un corpus especialitzat i detectar les unitats *in vivo* que apareixen vehiculant conceptes especialitzats per mitjà d'una relació. La nostra definició i esquematització de la noció de relació conceptual respon a l'estructuració del coneixement especialitzat en els textos i es mostra útil de cara a la sistematització en una base de dades lligada a una ontologia on tenen cabuda, no només les clàssiques relacions d'hiponímia i de meronímia, sinó un ventall més ampli de relacions conceptuals.

- Disposem a partir d'aquest treball d'una llista de marcadors verbals vehiculadors de relació conceptual. Aquests marcadors s'han presentat en forma de catàleg per a la descripció més teòrica del tipus de relació, però s'han usat ja com a font de partida del programa *Mercedes* per tal de detectar a un primer nivell els contextos que contenen unitats terminològiques però també possibles materialitzacions de relacions conceptuals.

- A més, hem integrat la tipologia de relacions conceptuals en una ontologia sobre el genoma humà i les unitats terminològiques que s'entren en la base de dades terminològiques que estem construint tenen el seu lligam directe a un concepte determinat de l'ontologia que, al seu torn, es troba relacionat amb els altres conceptes a partir d'una o diverses relacions conceptuals de la tipologia.

- La tasca de detecció de relacions conceptuals a partir de textos integra els programes *Mercedes* i *Yate* per intentar evitar al màxim la feina de detectar manualment les unitats candidates a ser termes. A partir d'aquí, disposem d'unitats terminològiques lligades per una relació. El següent pas és observar

si aquestes unitats ja són a la base de dades terminològica i si el seu corresponent concepte ja forma part de l'ontologia. En cas negatiu, aquesta informació es pot introduir per actualitzar la informació dels mòduls ontològic i lèxic de la base de coneixements sobre el genoma. A més, els fragments textuals en què apareix una determinada relació conceptual són inicialment vàlids per a formar part dels contextos que il·lustren l'ús d'una determinada unitat terminològica. Volem destacar, doncs, que a partir de la detecció de relacions conceptuals podem enriquir l'ontologia pel que fa al lligam entre els conceptes i, a més, si aquest lligam s'estableix entre conceptes nous, detectats pels sistemes d'extracció d'unitats terminològiques, la informació lingüísitica sobre la unitat també podrà ser directament extrapolada a partir del context en què la relació conceptual té un pes important.

- La descripció de noves estratègies lingüístiques per descartar o retenir un determinat marcador verbal servirà com a punt de partida per al disseny i la posterior implementació d'un primer prototip de detecció semiautomàtica de relacions conceptuals. Els resultats del sistema es podran utilitzar per a complementar la base de dades terminològiques sobre el genoma i també l'ontologia però, a més, creiem que el sistema pot beneficiar, en petit grau, el sistema d'extracció de candidats a terme pel que fa al factor del context terminològicament ric i, d'aquesta manera, la combinació de la cerca de termes i de relacions conceptuals podrà donar, de ben segur, resultats més acurats.

- A més llarg termini, i a partir del que acabem de dir, la combinació del sistema de detecció semiautomàtica de relacions conceptuals i el programa d'extracció de candidats a terme donarà com a resultat fragments de coneixement especialitzat, és a dir, nusos i relacions expressats per unitats terminològiques i verbs, que permetran visualitzar en forma de mapa conceptual el contingut especialitzat d'un text. Aquests resultats podran ser aprofitats en el marc de la recerca que actualment s'està fent en la representació tridimensional del coneixement especialitzat d'un text.

- Per últim, i lligat amb una de les futures vies de recerca que creiem que més aprofundirem en el futur, creiem que la nostra proposta també pot ser integrada en un sistema de recuperació d'informació especialitzada. A partir de les dades recollides en aquest treball podríem intentar aprofitar el reconeixement automàtic de les relacions conceptuals per tal de desenvolupar un sistema de recuperació d'informació que compatibilitzi les relacions expressades a través del llenguatge natural des usuaris, en les diferents consultes, amb les relacions existents en els documents sobre els quals es fa la cerca per tal de millorar els resultats pel que fa a la rellevància del document en relació a la consulta de l'usuari. A més, l'establiment dels diversos marcadors verbals per a cada tipus de relació conceptual permetrà expandir els termes de la consulta de l'usuari a d'altres termes relacionats. Si imaginem un usuari que està interessat en trobar malalties causades per l'alteració d'un gen, si l'usuari interroga el sistema de recuperació d'informació amb les paraules clau *malaltia* i *gen*, els resultats seran molt nombrosos i molt menys refinats que en el cas que l'usuari pogués indicar que vol les *malalties* CAUSADES PER *gen*. El sistema tindria la informació estructurada en funció de les relacions que s'estableixen en els diversos documents i proporcionaria resultats força més refinats. A més, a partir de la llista de les diferents unitats que poden expressar causalitat, no només trobaria els documents on efectivament apareix el marcador *causar* sinó que també podria expandir la consulta a unitats com *provocar*, per exemple. La integració de les relacions conceptuals en un sistema de recuperació d'informació comportaria, sens dubte, uns resultats més propers a la necessitat d'informació de l'usuari expressada en llenguatge natural i reconvertida i expandida per tal de proporcionar els millors resultats.

7.3 Futures vies de recerca

En arribar a aquest punt és quan realment, en lloc de finalitzar la recerca, s'obren diverses possibilitats de continuïtat en una tasca que tot just acaba de començar. En

aquest sentit, la voluntat de l'autora del treball és continuar investigant en l'estudi de les relacions conceptuals, principalment en els aspectes següents:

- Ampliar la llista de marcadors verbals que expressen una determinada relació conceptual i aprofundir en els patrons sintàctics en què es materialitzen aquest marcadors per tal d'arribar a una proposta d'implementació automàtica de detecció d'aquests patrons.

- Aplicar la llista de marcadors de relació conceptual a textos especialitzats d'altres àrees temàtiques per col·laborar en la construcció de l'estructuració del coneixement especialitzar a partir de corpus reals en d'altres àrees especialitzades.

- Expandir i complementar la llista de marcadors que vehiculen les relacions conceptuals amb altres unitats que no siguin verbs i que, de ben segur, també contribueixen a l'estructuració del coneixement especialitzat mitjançant les relacions conceptuals

- Implementar el sistema de detecció semiautomàtica de relacions conceptuals a partir d'un prototip inicial. Aquest prototip ens permetrà analitzar i avaluar els resultats i afegir nous criteris per tal d'aconseguir una resposta més refinada i acurada.

- Treballar en una proposta d'inclusió en l'ontologia de les relacions conceptuals detectades en els textos especialitzats per tal de sistematitzar i fer més lleugera la tasca d'introducció d'aquesta informació semàntica en el mòdul ontològic d'una base de coneixements.

- Establir les característiques que hauria de tenir un apartat del mòdul lèxic d'una base de coneixements que contingui totes les unitats amb els seus respectius patrons sintàctics i que pugui ser reutilitzada per a recuperació d'informació tenint en compte l'agrupació dels diversos marcadors verbals per a cada tipus de relació conceptual i la caracterització lingüística d'aquestes unitats.

- I finalment, i estretament lligat amb el punt anterior, avançar en la recerca sobre la creació d'un sistema de recuperació d'informació que integri les relacions conceptuals com a element clau de l'estructuració del coneixement especialitzat i que permeti acostar els resultats del sistema de recuperació d'informació a les necessitats reals expressades per l'usuari a partir d'una consulta en llenguatge natural.

Capítol 8

Referències bibliogràfiques

Capítol VIII

8 Referències bibliogràfiques

ALARCÓN, Rodrigo; SIERRA, Gerardo (2002) «Hacia la extracción automática de conceptos». A: *VIII Simposio Iberoamericano de Terminología. La Terminología entre la globalización y la localización*. Cartagena de Indias (Colòmbia). [CD-ROM]

ARISTÒTIL (1997) *Organon. I Catégories. II De l'interprétation*. París: Librairie Philosophique J. Vrin. (Bibliothèque des Textes Philosophiques).

ARNTZ, Reiner; PICHT, Heribert (1989) *Einführung in die Terminologiearbeit*. Traducció de l'alemany: DE IRAZAZÁBAL, Amelia (1995) *Introducción a la terminología*. Madrid: Fundación Germán Sánchez Ruipérez, Pirámide. (Biblioteca del libro).

BACH, Carme (2001) *Els connectors reformulatius catalans: anàlisi i proposta d'aplicació lexicogràfica*. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra. [Tesi doctoral]

BARRIÈRE, Caroline (2001) «Investigating the causal relation in informative texts». A: *Terminology*, 7, 2, p. 135-154.

BARRIÈRE, Caroline (2002) «Hierarchical refinement and representation of the causal relation». A: *Terminology*, 8, 1, p. 91-111.

BARRIÈRE, Caroline; HERMET, Matthieu (2002) «Causality taking root in Terminology». *TKE 2002. Terminology and Knowledge Engineering Proceedings. 6th International Conference 28th-30th August 2002*, p. 15-20.

BATEMAN, John A.; KASPER, Robert T.; MOORE, Johanna D.; WHITNEY, Richard (1990) *A General Organization of Knowledge for Natural Language Processing: the Penman Upper Model*. *Technical Report*. Marina del Rey (Califòrnia): USC/ISI.

BATEMAN John A.; MAGNINI, Bernardo; FABRIS, Giovanni (1995) «The Generalized Upper Model Knowledge Base: Organization and Use». A: MARS, Nicolaas J. I. (ed.) *Towards very large knowledge bases: knowledge building and knowledge sharing*. Amsterdam: IOS Press, p. 60-72.

BEAUGRANDE, Robert-Alain de; DRESSLER, Wolfang Ulrich (1997) *Introducción a la lingüística del texto*. Barcelona: Ariel, p. 135-168.

BEL, Núria; VILLEGAS, Marta (2000) «An introduction to SIMPLE». Workshop de l'Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra. Barcelona, desembre de 2000.

BENJAMINS, Richard; FENSEL, Dieter; DECKER, Stefan; GÓMEZ-PÉREZ, Asunción (1999) «$(KA)^2$ Building ontologies for the Internet: a mid term report». A: *International Journal of Human Computer Studies*, 51, 3, p. 687-712.

BIBER, Douglas (1993) «Representativeness in Corpus Design». A: *Literary and Linguistic Computing*, 8, 4, p. 243-257.

BIBER, Douglas *et al*. (1998) *Corpus linguistics. Investigating language structure and use*. Cambridge: Cambridge University Press.

BORST, Pim; AKKERMANS, Hans; TOP, Jan (1996) «Engineering ontologies». *International Journal of Human-Computer Studies*, 46, p. 365-406. [Report INF-96-09 de la University of Twente]

BOWDEN, Paul R.; EVETT, Lindsay; HALSTEAD, Peter (1998) «Automatic Acronym Acquisition in a Knowledge Extraction Program». A: *Computerm '98*, p. 43-49.

BUDIN, Gerhard (1996) «Terminology Science as Applied Philosophy of Science». A: MYKING, Joahn; SÆBØE, Randi; TOFT, Bertha (red.) *Terminologi-system og*

kontekst. Nordisk minisymposium, 1996. *KULTs skiriftserie*, 71. Bergen: Universiteter Bergen, p. 59-71.

CABRÉ, M. Teresa (dir.) (1996) *Terminologia. Selecció de textos d'E. Wüster*. Barcelona: Servei de Llengua Catalana, Universitat de Barcelona.

CABRÉ, M. Teresa; MOREL, Jordi; TEBÉ, Carles (1996) «Las relaciones conceptuales de tipo causal: un caso práctico». A: *Actas del V Simposio Iberoamericano de Terminología: Terminología, ciencia y tecnología*. Ciudad de México, 3-8 de novembre de 1996. Mèxic: Unión Latina, p. 82-94.

CABRÉ, M. Teresa (dir.) (1999a) *La terminología: Representación y comunicación. Una propuesta de base comunicativa y otros artículos*. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra. (Sèrie Monografies, 3).

CABRÉ, M. Teresa (dir.) (1999b) «La enseñanza de la terminología en España: problemas y propuestas». A: *Hermeneus. Revista de Investigación en Traducción en Interpretación*, 2/2000, p. 41-94.

CABRÉ, M. T.; FELIU, J. (ed.) (2001) *La terminología científico-técnica: reconocimiento, análisis y extracción de información formal y semántica (DGES PB96-0293)*. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra. (Sèrie Materials, 2).

CABRÉ, M. T.; FELIU, J.; TEBÉ, C. (2001) «Bases cognitivas de la terminología: hacia una visión comunicativa del concepto». A: *Sendebar*, 12, p. 301-310.

CABRÉ, M. T. (2002) «Textos especializados y unidades de conocimiento: metodología y tipologización». A: GARCÍA PALACIOS, Joaquín; FUENTES, M. Teresa (ed.) *Texto, terminología y traducción.* Salamanca: Ediciones Almar, p. 15-36.

CABRÉ, M. T. (2002) «Análisis textual y terminología, factores de activación de la competencia cognitiva en la traducción». A: ALCINA CAUDET, A.; GAMERO PÉREZ, S. (ed.) *La traducción científico-técnica y la terminología en la sociedad de la información*. Castellón: Publicacions de la Universitat Jaume I, p. 87-105.

CABRÉ, M. T. (2003) «Theories of terminology. Their description, prescription and explanation». A: *Terminology*, 9, 2, p. 163-200.

CEDERBERG, Scott; WIDDOWS, Dominic (2003) «Using LSA and Noun Coordination Information to Improve the Precision and Recall of Automatic Hyponymy Extraction». A: *Conference on Natural Language Learning* (CoNLL-2003). Edmonton (Canadà), p. 111-118.

CHAFFIN, Roger (1992) «The Concept of a Semantic Relation». A: LEHRER, Adrienne; FEDER, Eva (ed.) *Frames, Fields and Contrasts*. New Jersey: Lawrence Elbaum Associates Publishers, p. 253-288.

CHAFFIN, Roger; HERRMANN, Douglas J. (1988) «The nature of semantic relations: a comparison of two approaches». A: EVENS, Martha (ed.) *Relational models of the lexicon. Representing knowledge in semantic networks*. Cambridge: Cambridge University Press, p. 289-334.

COLLIER, Nigel; NOBATA, Chikashi; TSUJII, Junichi (2001) «Automatic acquisition and classification of terminology using a tagged corpus in the molecular biology domain». A: *Terminology*, 7, 2, p. 239-257.

Computers and the Humanities. Special Issue on EuroWordNet, 32. Dordrecht: Kluwer Academic Publishers, 1998.

CONDAMINES, Anne (1995) «Terminology: New needs, new perspectives». A: *Terminology*, 2, 2, p. 219-238.

CONDAMINES, Anne; REBEYROLLE, Josette (1997) «Construction d'une base de connaissances terminologiques à partir de textes: expérimentation et définition d'une méthode». A: *Actes des Journées Ingénierie des connaissances et Apprentissage Automatique (JICAA '97)*. Roscoff, 20-22 de maig de 1997, p. 191-206.

CONDAMINES, Anne; REBEYROLLE, Josette (1998) *CTKB: A corpus-based approach to a Terminological Knowledge Base*. A: BOURIGAULT, Didier; JACQUEMIN, Christian; L'HOMME, Marie-Claude (ed.) *Computerm '98. First Workshop on*

Computational Terminology. Proceedings of the Workshop. COLING-ACL '98, 15 d'agost de 1998. Montreal (Quebec): Université de Montréal, p. 29-35.

CONDAMINES, Anne (1999) «Approche sémasiologique pour la constitution de Bases de Connaissances Terminologiques» A: DELAVIGNE, Valérie; BOUVERET, Myriam (ed.) *Sémantique des termes spécialisés*. Rouen: Université de Rouen, p.101-118.

CONDAMINES, Anne; REBEYROLLE, Josette (2001) «Searching for and identifying conceptual relationships via a corpus-based approach to a Terminological knowledge Base (CTKB)». A: BOURIGAULT, Didier; JACQUEMIN, Christian; L'HOMME, Marie-Claude (2001) *Recent Advances in Computational Terminology*. Amsterdam / Philadelphia: John Benjamins Publishing Company, p. 125-148.

CONDAMINES, Anne (2002) «Corpus analysis and conceptual relation patterns». A: *Terminology*, 8, 1, p. 141-162.

CRAIG, Colette (ed.) (1986) *Noun Classes and Categorization. Proceedings of a Symposium on Categorization and Noun Classification*. Eugene (Oregon), octubre de 1983. Amsterdam / Filadèlfia: John Benjamins Publishing Company.

CRUSE, David A. (1986) *Lexical Semantics*. Cambridge: Cambridge University Press.

CRUSE, David A. (2000) *Meaning in Language. An Introduction to Semantics and Pragmatics*. Nova York: Oxford University Press.

DAVIDSON, Laura (1997) *Knowledge Extraction Technology for Terminology*. Ottawa: School of Translation and Interpretation. [Tesi doctoral]

DAVIDSON, Laura; KAVANAGH, Judy; MACKINTOSH, Kristen; MEYER, Ingrid.; SKUCE, Douglas (1998) «Semi-automatic Extraction of Knowledge-rich Contexts from Corpora». A: BOURIGAULT, Didier; JACQUEMIN, Claude; L'HOMME, Marie-Claude (ed.) *Computerm '98. First Workshop on Computational Terminology. Proceedings of the Workshop*. COLING-ACL '98, 15 d'agost de 1998. Montreal (Quebec): Université de Montréal, p. 50-56.

DELEUZE, Gilles; GUATTARI, Felix (1977) *Rizoma (Introducción)*. València: Pre-textos.

DESCLÉS, Jean-Pierre (1996) «Appartenance/inclusion, localisation, ingrédience et possession». A: *Faits de langues*, 7, p. 91-100.

DÍEZ ORZAS, Pedro Luis (1999) «La relación de meronimia en los sustantivos del léxico español: contribución a la semántica computacional» [en línia]. A: *Estudios de Lingüística Española*, vol.2. *http://elies.rediris.es/elies2* [Consulta: 31 de maig de 1999].

DIN 2330 (1979) *Begriffe und Bennenungen: Allgemeine Grundsätze* [Conceptes i denominacions: principis generals]. Berlín/Colònia: Beuth.

DIN 2331 (1980) *Begriffsysteme und ihre Darstellun* [Sistemes conceptuals i la seva representació]. Berlín/Colònia: Beuth.

DIN 2342 (1986) *Begriffe der Terminologielehre: Grundbegriffe* [Conceptes de la teoria de la terminologia: conceptes fonamentals]. Berlín/Colònia: Beuth.

DOMÈNECH, Meritxell (1998) *Unitats de coneixement i textos especialitzats: primera proposta d'anàlisi*. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra. [Treball de recerca de doctorat no publicat]

DUBOIS, Danièle (dir.) (1993) *Sémantique et cognition. Catégories, prototypes, typicalité*. París: CNRS Editions.

DUBOIS, Danièle (1999) «Le lexique: fixateur des représentations et producteur d'ontologie». Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra. [Ponència no publicada presentada en el Seminari de Terminologia Teòrica, 28-29 de gener de 1999]

DUBOIS, Danièle (2001) «Lexique(s) et catégories: de la perception individuelle aux connaissances partagées». A: *Terminología y cognición*. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra. (Sèrie Activitats, 7).

DUINEVELD, A. J.; STOTER, R.; WEIDEN, M. R.; KENEPA, B.; BENJAMINS, V. R. (2000) «WonderTools? A comparative study of ontological engineering tools». A: *International Journal of Human-Computer Studies*. Academic Press, 52, 6, p. 1.111-1.133.

EAGLES (1999) «Preliminary Recommendations on Lexical Semantic Encoding. Final Report». The EAGLES Lexicon Interest Group. EAGLES Document: LE3-4244 [en línia]. *http://www.ilc.pi.cnr.it/EAGLES96/browse.html#wg2* [Consulta: 10 de juliol de 2002].

ESTOPÀ, Rosa (1999) *Extracció de Terminologia: Elements per a la Construcció d'un SEACUSE (Sistema d'Extracció Automàtica de Candidats a Unitats de Significació Especialitzada)*. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra. (Sèrie Tesis, 2). [Tesi doctoral publicada en CD-ROM, 2003]

EVENS, Martha W.; LITOWITZ, Bonnie E.; MARKOWITZ, Judith A.; SMITH, Raoul N.; WERNER, Oswald (1980) *Lexical-Semantic Relations: A Comparative Survey*. Canadà: Linguistic Research Inc.

Faits de langues. Revue de linguistique, 7 («La relation d'appartenance»). París: Ophrys, 1996.

FABER, Pamela; LÓPEZ RODRIGUEZ, Clara Inés; TERCEDOr, María Isabel (2001) «Utilització de técnicas de corpus en la representación del conocimiento médico». A: *Terminology*, 7, 2, p. 167-198.

FABER, Pamela; JIMÉNEZ, Catalina (ed.) (2002) *Investigar en terminología*. Granada: Editorial Comares.

FACHBEREICH, Vom (1996) *Relational Concept Analysis: Semantic Structures in Dictionaries and Lexical Databases*. Darmstadt: Gesellschafts- und Geschichtswissenchaften der Technischen. [Tesi doctoral]

FELBER, Heribert (1984) *Terminology Manual*. París: Unesco, Infoterm, United Nations Educational, Scientific and Cultural Organization.

FELBER, Helmut; PICHT, Heribert (1984) *Métodos de terminografía y principios de investigación terminológica*. Madrid: Instituto Miguel de Cervantes (CSIC).

FELIU, Judit (2000) *Relacions conceptuals i variació funcional: elements per a un sistema de detecció automàtica*. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra. [Treball de recerca de doctorat]

FELIU, Judit (2001) «Propuesta de clases conceptuales y de relaciones conceptuales: recopilación y análisis». A: CABRÉ, M. T.; FELIU, J. (ed.) *La terminología científico-técnica: reconocimiento, análisis y extracción de información formal y semántica (DGES PB96-0293)*. Barcelona, Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra, p. 143-154.

FELIU, Judit; QUIXAL, Martí (2002) *Manual d'ús d'OntoTerm. Versió 0.98.* Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra. [Document de treball]

FELIU, Judit; CABRÉ, M. Teresa (2002) «Conceptual relations in specialized texts: new typology and an extraction system proposal». A: *TKE 2002. Terminology and Knowledge Engineering Proceedings. 6th International Conference 28th-30th August 2002*, p. 45-49.

FELIU, Judit; SOLÉ, Elisabet; TEBÉ, Carles; CABRÉ, M. Teresa (2002) «Las relaciones conceptuales: un elemento esencial en la estructuración del conocimiento especializado» A: *Actas del VIII Simposio Iberoamericano de Terminología*. Cartagena de Indias (Colòmbia), 28-31 d'octubre de 2002. [CD-ROM]

FELIU, Judit; VIVALDI, Jorge; CABRÉ, M. Teresa (2002a) *Ontologies: a review* [en línia]. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra (Papers de l'IULA, 34). *ftp://ftp.iula.upf.es/pub/publicacions/02inf034.pdf* [Consulta: 20 d'agost de 2002].

FELIU, Judit; VIVALDI, Jorge; CABRÉ, M. Teresa (2002b) «Towards an Ontology for a Human Genome Project». A: *LREC2002. Third International Conference on*

Language Resources and Evaluation. Proceedings. Las Palmas de Gran Canaria, maig de 2002, p. 1.885-1.890.

FELIU, Judit; SOLÉ, Elisabet; TEBÉ, Carles (2003) «Las relaciones meronímicas en terminología: análisis semántico-textual y aplicaciones». A: *Terminología e industrias da língua. Actas do VII Simpósio Ibero-Americano de Terminologia*. Lisboa: ILTEC, Instituto de Lingüística Teórica e Computacional, p. 389-402.

FELLBAUM, Christiane (ed.) (1998) *WordNet: An Electronic Lexical Database*. Cambridge: MIT Press.

FRIDMAN, Natalya (1997) *Knowledge Representation for Intelligent Information Retrieval in Experimental Sicences*. Boston: Northeastern University. [Tesi doctoral]

FULFORD, Heather (2001) «Exploring terms and their linguistic environment in text». A: *Terminology*, 7, 2, p. 259-279.

FODOR, Jerry A. (1998) *Concepts. Where Cognitive Science Went Wrong*. Oxford: Clarendon Press.

GALINSKI, Christian; GOEBEL, Jürgen W. (1996) *Guide to Terminology Agreements*. Viena: Infoterm.

GARCÍA, Daniela (1998) *Analyse automatique des textes par l'oganisation causale des actions. Réalisation du système informatique COATIS*. París: Université de Paris-Sorbonne (París IV), UFR: Institut des Sciences Humaines Appliquées (ISHA). [Tesi doctoral]

GAUDIN, François (1991) «Terminologie et travail scientifique: mouvements de signes, mouvements de connaissances». A: *Cahiers de Linguistique Sociale*, 18, p. 111-131.

GIVÓN, Talmy «Prototypes: between Plato and Wittgenstein». A: CRAIG, Colette (ed.) (1986) *Noun Classes and Categorization*. Amsterdam/Filadèlfia: John Benjamins Publishing Company, p. 77-102.

GRUBER, Thomas R. (1992) *Ontolingua: A mechanism to Support Portable Ontologies.* Report KSL 91-66. Stanford: Stanford University.

GRUBER, Thomas R. (1993a) *Toward Principles for the Design of Ontologies Used for Knowledge Sharing* [en línia]. Stanford Knowledge Systems Laboratory. Document presentat al Padua workshop on Formal Ontology, març 1993. *http://gicl.mcs.drexel.edu/people/regli/Classes/KBA/Readings/onto-design.pdf* [Consulta: 20 d'agost de 2003].

GRUBER, Thomas R. (1993b) *What is an Ontology?* [en línia]. Stanford Knowledge Systems Laboratory. *http://www-ksl.stanford.edu/kst/what-is-an-ontology.html* [Consulta: 15 d'agost de 2003].

GRUBER, Thomas R. (1993c) «A Translation Approach to Portable Ontology Specifications». A: *Knowledge Acquisition*, 5, 2, p. 199-220.

GRUBER, Thomas R.; OLSEN, Gregory R. (1994) «An ontology for engineering mathematics». A: DOYLE, P. Torasso; SANDEWALL, E. (ed.) *Fourth International Conference on Principles of Knowledge Representation*. San Mateo, CA.: Morgan Kaufmann, p. 258-269.

HAMON, Thierry (2000) *Variation sémantique en corpus spécialisé: Acquisition de relations de synonymie à partir de ressources lexicales*. París: Université Paris-Nord. [Tesi doctoral]

HEARST, Marti A. (1992) «Automatic Acquisition of Hyponyms from Large Text Corpora». A: *Proceedings of the Fourteenth International Conference on Computational Linguistics*. Nantes, p. 539-545.

HEMPEL, Carl Gustav (1952) *Fundamentals of concept formation in empirical science*. Chicago: Chicago University Press.

HOFFMAN, Lothar (1991) «Texts and Text Types in LSP». A: SCHRÖDER, Hartmut (ed.) *Subject-oriented Texts. Languages for Special Purposes and Text Theory*. Offprint, Berlín-Nova York: Zalter de Gruyter, p. 58-166.

HOFFMAN, Lothar (1998) *Llenguatges d'especialitat. Selecció de textos.* Edició a càrrec de J. Brume. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra. (Sèrie Monografies, 1).

IRIS, Madelyn A.; LITOWITZ, Bonne E.; EVENS, Martha (1988) «The problems of the part-whole relation». A: EVENS, Martha (ed.) (1998) *Relational models of the lexicon. Representing knowledge in semantic networks*. Cambridge: Cambridge University Press, p. 260-287.

ISO/704 (1987) *Principes et méthodes de la terminologie — Principles and methods of terminology*. Norme internationale — International Standard.

ISO/1087 (1990) *Terminology - Vocabulary — Terminologie - Vocabulaire.* Norme internationale — International Standard.

JACKENDOFF, Ray (1990) *Semantic Structures*. Cambridge, Massachussets: The MIT Press, 2a ed. 1991.

JACKIEWICZ, Agata (1996) «L'expression lexicale de la relation d'ingrédience (partie-tout)». A: *Faits de langues*, 7, p. 53-62.

JONES, Steven (1998) «Approaching Antonymy Afresh». A: *LWPAL (Liverpool Working Papers in Applied Linguistics*, 4, 1, p. 71-85.

KAMEL, Maged N.; ROUDSARI, Abdul V.; CARSON, Ewart R. (2002) «Towards a semantic medical Web: HealthCyberMap's tool for building an RDF metadata base of health information resources based on the Qualified Dublin Core Metadata Set» [en línia]. A: *Med Sci Monit*, 8, 7, p. 124-136. *http://www.MedSciMOnit.com/pub/vol_8/ no_7/2615.pdf* [Consulta: 19 de juliol de 2002].

KANG, Sin-Jae; CHUNG, You-Jin; LEE, Jong-Hyeok (2002) «Language Independent and Practical Ontology in Korean-Japanese Machine Translation Systems». A: *Literary and Linguistic computing*, 17, 1, p. 19-36.

KAROLAK, Stanislas (1996) «Considérations sur le concept d'appartenance». A: *Faits de langues*, 7, p. 101-110.

KEMPSON, Ruth M. (1992) *Semantic Theory*. Cambridge: Cambridge University Press.

KLEIBER, Georges (1990) *La sémantique du protoype*. París: Presses Universitaires de France.

KOCOUREK, Rotislav (1982) *La langue française de la technique et de la science. Vers une linguistique de la langue savante*. Wiesbaden: Oscar Brandstetter Verlag Gmbh & Co. KG.

LAKOFF, Georges (1986) «Classifiers as reflection of Mind». A: CRAIG, Colette (ed.) (1986) *Noun Classes and Categorization*. Amsterdam/Filadèlfia: John Benjamins Publishing Company, p. 13-51.

LAKOFF, Georges (1987) *Women, Fire, and Dangerous Things. What Categories Reveal about the Mind*. Chicago: The University of Chicago Press.

LANGACKER, Ronald W. (1997) «The contextual basis of cognitive semantics». A: NUYTS, Jan; PEDERSON, Eric (ed.) (1997) *Language and conceptualization*. Cambridge: Cambridge University Press, p. 229-252.

LARA, Luis Fernando (1998/1999) «"Concepts" and Term Hierarchy». A: *Terminology*, 5, 1, p. 59-76.

LE2-4003 (1997) Deliverable D005: «Definition of the links and subsets for nouns of the EuroWordNet project» [en línia]. *http://www.ley.una.ln/~ewn/docs/* [Consulta: 10 de juny de 2002].

LE2-4003a (1998) Deliverable D027: «EuroWordNet Subset2 for Dutch, Spanish and Italian» [en línia]. *http://www.ley.una.ln/~ewn/docs/* [Consulta: 10 de juny de 2002].

LE2-4003b (1998) Deliverable D017: «The EuroWordNet Base Concepts and Top Ontology» [en línia]. http://www.ley.una.ln/~ewn/docs/ [Consulta: 10 de juny de 2002].

LE3-4244 (1999) «Preliminary Recommendations on Lexical Semantic Encoding. Final Report». The EAGLES Lexicon Interest Group [en línia]. http://www.ilc.pi.cnr.it/EAGLES96/browse.html#wg2 [Consulta: 10 de juny de 2002].

LEECH, Geoffrey (1974) *Semantics. The study of meaning*. 2a ed. Gran Bretanya: Penguin Books, 1990.

LENAT, Douglas B.; GUHA, Ramanathan V. (1990) «Building Large Knowledge-based systems: Representation and Inference in the CYC project». Boston: Addison Wesley Publishing.

LENCI, Alessandro *et al*. (1999) *SIMPLE Work Package. Linguistic Specificatiions, Deliverable D2.1* [en línia]. http://www.ub.es/gilcub/SIMPLE/simple.html [Consulta: 22 de maig de 2002].

LERAT, Pierre (1983) *Sémantique descriptive*. Paris: Hachette, p. 5-38 (Col. Langue, Linguistique).

L'HOMME, Marie-Claude (2003) «Indices de relations conceptuelles dans les définitions terminologiques. Application au domaine de l'informatique». *I Jornada Internacional de Terminologia* [en línia]. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra, p. 44-51. (Sèrie Activitats, 12). ftp://ftp.iula.upf.es/pub/ publicacions/publi031.pdf [Consulta: 5 d'octubre de 2002]

LOCKE, John (1975) *An Essay concerning Human Understanding*. Oxford: Oxford University Press (8a ed. 1991).

LÓPEZ, Clara Inés (2000) *Tipología textual y cohesión en la traducción biomédica inglés español: un estudio de corpus*. Granada: Universidad de Granada, Departamento de Traducción e Interpretación. [Tesi doctoral]

LYONS, John (1978) *Élements de sémantique*. París: Larousse.

LYONS, John (1995) *Linguistic Semantics. An Introduction*. Cambridge: Cambridge University Press.

MAEDCHE, Alexander; STAAB, Steffen (2000) «Discovering Conceptual Relations from Text». A: HORN, W. (ed.) *ECAI 2000. Proceedings of the 14th European Conference on Artificial Intelligence. Berlin, August 21-25, 2000*. Amsterdam: IOS Press.

MÀRQUEZ, Lluís (1999) *Part-of-speech Taggng: A Machine Learning Approach based on Decision Trees*. Barcelona: Departament de Llenguatges i Sistemes Informàtics, Universitat Politècnica de Catalunya. [Tesi doctoral]

MARSHMAN, Elisabeth; MORGAN, Tricia; MEYER, Ingrid (2002) «French patterns for expressing concept relations». A: *Terminology*, 8, 1, p. 1-29.

MARSHMAN, Elisabeth (2002) «The Cause-Effect Relation in a Biopharmaceutical Corpus: English Knowedge Patterns». *TKE 2002. Terminology and Knowledge Engineering Proceedings. 6th International Conference 28th-30th August 2002*, p. 89-94.

MAYNARD, Diana (1999) *Term recognition using combined knowledge sources*. Manchester: Manchester Metropolitan University, Faculty of Science and Engineering. [Tesi doctoral]

MEL'CUK, Igor (1984) *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques I*. Montréal: Les Presses de l'Université de Montréal, p. XIII-XVI i 2-16.

MEL'CUK, Igor (1988) *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques II*. Montréal: Les Presses de l'Université de Montréal, p. 1-47.

MEL'CUK, Igor (1992) *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques III*. Montréal: Les Presses de l'Université de Montréal, p. 1-58.

MEYER, Ingrid (1984) «Computer-assisted Concept Analysis for Terminology Work». A: WRIGHT, Sue-Ellen; BUDIN, Gerhard (ed.) (1984) *Handbook of Terminology Management. Volume I. Basic Aspects of Terminology Management*. Amsterdam/Filadèlfia: John Benjamins Publishing Company.

MEYER, Ingrid (1990) «Computer-assisted Concept Analysis for Terminology Work». A: *Proceedings of the Nordic Post Graduate Course in Terminology*. Mariehamn (Finlàndia), setembre de 1990. Estocolm: Tekniska nomenklaturcentralen, p. 193-212.

MEYER, Ingrid (1998) *The COGNITERM Project*. [en línia] *http://aix1.uottawa.ca/~imeyer/ research.htm* [Consulta: 5 de novembre de 1998].

MEYER, Ingrid (2001) «Extracting knowledge-rich contexts for terminography». A: BOURIGAULT, Didier; JACQUEMIN, Christian; L'HOMME, Marie-Claude (2001) *Recent Advances in Computational Terminology*. Amsterdam/Filadèlfia: John Benjamins Publishing Company, p. 279-302.

MILLER, George A.; BECKWITH, Richard; FELBAUM, Christiane; GROSS, Derek; MILLER, Katherine (1993) *Introduction to WordNet: An On-line Lexical Database* [en línia]. *http://www.cogsci.princeton.edu/~wn/5papers.pdf* [Consulta: 15 de desembre de 2000].

MORENO, Antonio (1999) «An introduction to OntoTerm». Workshop a l'Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra. Barcelona, juny de 1999.

MORENO, Antonio (2000) *Diseño e implementación de un lexicón computacional para lexicografía y traducción automática* [en línia]. A: *Estudios de Lingüística Española*, vol. 9. *http://elies/rediris/es/elies9/* [Consulta: 6 de maig de 2001]. [Tesi doctoral]

MORENO, Antonio; PÉREZ, Chantal (2000) «Reusing the MikroKosmos Ontology for Concept-Based Multilingual Terminology Databases». A: *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC-2000). May 31th-June 2nd*. Atenes, p. 1.061-1.067.

MORIN, Emmanuel (1999a) *Extraction de liens sémantiques entre termes à partir de corpus de textes techniques*. Nantes: Laboratoire IRIN, Institut de Recherche en Informatique, Université de Nantes. [Tesi doctoral]

MORIN, Emmanuel (1999b) «Automatic Acquisition of Semantic Relations between Terms from Technical Corpora». A: SANDRINI, Peter (ed.) *TKE'99 Terminology and Knowledge Engineering. Proceedings Fifth International Congress on Terminology and Knowledge Engineering 23-27 agost 1999*. Viena: Termnet, p. 268-278.

MURTHY, Kolluru Venkata Sreerama (1995) *On Growing Better Decision Trees from Data* [en línia]. Baltimore (Maryland): John Hopkins University. http://www.tigr.org/~salzberg/murthy_thesis/thesis.html [Consulta: 7 d'abril de 2003]. [Tesi doctoral]

NISTRUP, Bodil; SANDFORD, Bolette, ERDMAN, Hanne (2001) «Defining Semantic Relations for OntoQuery». A: JENSEN, Per Anker; SKADHAUGE, Peter (ed.) *Proceedings of the First International OntoQuery Workshop, January 17-18*. University of Southern Denmark, Department of Business Communication and Information Science, p. 57-58.

NIREMBURG, Sergei; RASKIN, Victor (en preparació) *Ontological Semantics.* Massachussets: MIT Press.

NLM (1998) *UMLS Knowledge Sources*. 8a ed. National Library of Medicine. U.S. Dept. of Health and Human Services.

NOY, N. F.; FERGERSON, R. W.; MUSEN, M. A. (2000) «Knowledge-Acquisition Interfaces for Domain Experts: An Empirical Evaluation of Protege-2000». A: *Proceedings of the Twelfth International Conference on Software Engineering and Knowledge Engineering (SEKE2000)*. Chicago. [en línia]. http://www-

smi.stanford.edu/pubs/SMI_Reports/SMI-2000-0825.pdf [Consulta: 15 de juny de 2002].

NUOPPONEN, Anita (1994a) «Causal Relations In Terminological Knowledge Representation». A: *Terminology Science & Research* , 5, 1, 36-44.

NUOPPONEN, Anita (1994b) «On Causality and Concept Relationships». A: DRASKAU, Jennifer; PICHT, Heribert (ed.): *Terminology Science and Terminology Planning, IITF-Workshop on Theoretical Issues of Terminology Science*. Viena: TermNet, p. 217-230.

NUOPPONEN, Anita (1994c) «Wüster revisited: On Causal Concept Relationships and Causal Concept Systems». A: BREKKE, Magnar; ANDERSEN, Øivin; DAHL, Trine; MYKING, Johan (ed.) *Applications and Implications of Current LSP Research, Proceedings of the 9th European Symposium on LSP*. Vol. II. Bergen: Fagbokforlaget, p. 532-539.

NUOPPONEN, Anita (1997) *A model for systematic terminological analysis*. A: LUNDQUIST, Lita; PICHT, Heribert; QVISTGAARD, Jacques (ed.) *LSP - Identity and Interface Research, Knowledge and Society. The proceedings of LSP Symposium 1997*. Copenhagen: Copenhagen Business School, p. 363-372.

OTMAN, Gabriel (1996a) *Les représentations sémantiques en terminologie*. París: Masson.

OTMAN, Gabriel (1996b) «Le traitement automatique de la relation partie-tout en terminologie». A: *Faits de langues*, 7, p. 43-52.

PICHT, Heribert; DRASKAU, Jennifer (1985) *Terminology: an introduction*. Gran Bretanya: University of Surrey.

PEARSON, Jennifer (1998) *Terms in Context*. Amsterdam / Filadèlfia: John Benjamins Publishing Company.

PÉREZ, M. Chantal (2000) *Explotación de los corpora textuales informatizados paral a creación de bases de datos terminológicas*. Màlaga: Universidad de Málaga,

Facultad de Filosofía y Letras, Departamento de Filología Inglesa, Francesa y Alemana. [Tesi doctoral]

POSNER, Michael I. (1986) «Empirical Studies of Prototypes». A: CRAIG, Colette (ed.) *Noun Classes and Categorization*. Amsterdam/Filadèlfia: John Benjamins Publishing Company, p. 53-61.

POZZI, María (1999) «The Concept of 'Concept' in Terminology: a Need for a new Approach». A: SANDRINI, Peter (ed.) *TKE'99 Terminology and Knowledge Engineering. Proceedings Fifth International Congress on Terminology and Knowledge Engineering 23-27 agost 1999*. Viena: Termnet, p. 28-42.

PUSTEJOVSKY, James (1995) *The Generative Lexicon*. Cambridge (Massachussetts) / Londres: The MIT Press, p. 1-26.

PUSTEJOVSKY, James; RUMSHISKY, Anna; CASTAÑO, José (2002) «Rerendering Semantic Ontologies: automatic Extensions to UMLS through Corpus Analytics». A: *Proceedings of the Ontologies and Lexical Knowledge Bases Workshop (OntoLex 02) de la 3rd International Language Resources and Evaluation Conference, LREC 2002*, p. 60-67.

PUTNAM, Hilary (1981) *Reason, Truth and History*. Cambridge: Cambridge University Press.

RASTIER, François (1991) *Sémantique et recherches cognitives*. París: Presses Universitaires de France.

RASTIER, François (1993) «Catégorisation, typicalité et lexicologie. Préliminaires critiques». A: DUBOIS, Danièle (dir.) *Sémantique et cognition. Catégories, prototypes, typicalité*. París: CNRS Editions, p. 259-277.

REY, Alain (1992) *La terminologie: noms et notions*. 2a ed. corregida. París: Presses Universitaires de France. («Que sais-je?»).

REY, Alain (1995) *Essays on terminology*. Traducció de J. C. Sager. Amsterdam / Filadèlfia: John Benjamins Publishing Company.

RICHARDSON, Stephen D.; DOLAN, William B.; VANDERWENDE, Lucy (1998) «MindNet: acquiring and structuring semantic information from text». A: *Proceedings of ACL-Coling 1998*, p. 1.098-1.102.

ROBINSON, Edward A. (1997) «The cognitive foundations of pragmatic principles: implications for theories of linguistic and cognitive representation». A: NUYTS, Jan; PEDERSON, Eric (ed.) *Language and conceptualization*. Cambridge: Cambridge University Press, p. 253-271.

RONDEAU, Guy; FELBER, Heribert (1981) *I. Fondements théoriques de la terminologie*. Université Laval, Quebec: GISTERM.

ROSARIO, Barbara; HEARST, Marti A.; FILLMORE, Charles (2002) «The Descent of Hierarchy, and Selection in Relational Semantics». A: *Proceedings of ACL'02*. Filadèlfia, p. 417-424.

ROSCH, Eleanor (1975) «Cognitive Representation of Semantic Categories». A: *Cognitive Psycohology*, 7, p. 532-547.

ROSCH, Eleanor (1978) «Principles of Categorization». A: ROSCH, Eleanor; LLOYD, Barbara B. (ed.) *Cognition and Categorization*. New Jersey: Lawrence Erlbaum Associates, Inc., Publishers, p. 27-48.

SAEED, John I. (1997) *Semantics*. Gran Bretanya: Blackwell Publishers Ltd.

SANDRINI, Peter (ed.) (1999) *Terminology and Knowledge Engineering. TKE'99. Proceedings of the Fifth International Congress on Terminology and Knowledge Engineering*. 23-27 d'agost de 1999. Insbruck / Viena: TermNet.

SAGER, Juan-Carlos; DUNGWORTH, David; MCDONALD, Peter F. (ed.) (1980) *English Special Languages. Principles and practice in science and technology*. Wiesbaden: Oscar Branstetter Verlag Kg.

SAGER, Juan-Carlos; KAGEURA, Kyo (1994/95) «Concept Classes and Conceptual Structures: Their Role and Necessity». A: *ALFA* 7/8, p. 191-216.

SALVADOR, Vicent (2000) «L'estil nominalitzat». A: *Caplletra*, 29, p. 69-82.

SINCLAIR, John M. (1997) «Corpus Evidence in Language Description». A: WICHMANN, Anne *et al.* (ed.) *Teaching and Language Corpora*. Londres i Nova York: Longman, p. 27-39.

SKUCE, Douglas; LETHBRIDGE, Timothy C. (1995) *CODE4: A Unified System for Managing Conceptual Knowledge* [en línia]. Ottawa: Universitat d'Ottawa. *http://www.csi.uottawa.ca/~tcl/papers/ijhcs95/cod4jau3.html#RTFToC7* [Consulta: 30 de maig de 1999].

SOLÉ, Elisabet (2002) *Els noms col·lectius catalans. Descripció i reconeixement*. Barcelona: Institut Universitari de Lingüística Aplicada. [Tesi doctoral]

SOLER, Carme (1997) *Desajustes léxicos nominales y su representación en una base de conocimientos léxicos. Valores semánticos del adjetivo*. Barcelona: Universitat Politècnica de Catalunya. [Tesi doctoral]

SOMERS, Harold (ed.) (1996) *Terminology, LSP and Translation. Studies in language engineering in honour of Juan C. Sager*. Amsterdam / Filadèlfia: John Benjamins Publishing Company.

STENGERS, Isabelle (dir.) (1987) *D'une science a l'autre. Des concepts nomades*. París: Editions du Seuil.

STENGERS, Isabelle; SCHLANGER Judith (dir.) (1991) *Les concepts scientifiques. Invention et pourvoir*. París: Éditions Gallimard. (Folio/essais).

TEBÉ, Carles (1996) *Els conceptes en la teoria terminològica: anàlisi i revisió crítica*. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra. [Treball de recerca de doctorat]

VIVALDI, Jorge (2001) *Extracción de candidatos a término mediante combinación de estrategias heterogéneas*. Barcelona: Departament de Llenguatges i Sistemes Informàtics, Universitat Politècnica de Catalunya. [Tesi doctoral]

VIVALDI, Jorge; RODRÍGUEZ, Horacio (2002) «Medical Term Extraction using the EWN ontology». *TKE 2002. Terminology and Knowledge Engineering Proceedings. 6th International Conference 28th-30th August 2002*, p. 137-142.

VIVALDI, Jorge (2003a) *Sistema de reconocimiento de términos* Mercedes. *Manual de utilización*. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra.

VIVALDI, Jorge (2003b) *Sistema de extracción de Candidatos a Término* YATE. *Manual de utilización*. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra.

VOSSEN, P. (ed.) (1999) *EuroWordNet General Document* [en línia]. University of Amsterdam. *http://www.hum.uva.nl/~ewn* [Consulta: 3 de març de 2001]

WINSTON, Morton E.; CHAFFIN, Roger; HERRMANN, Douglas (1987) «A Taxonomy of Part-Whole Relations». A: *Cognitive Science*, 11, 417-444.

YU, Hong; FRIEDMAN, Carol; RHZETSKY, Andrey; KRA, Pauline (1999) «Representing Genomic Knowledge in the UMLS Semantic Network». A: *Proceedings of the AMIA Symposium*, p. 181-185.