Detecció de la relació malaltia-símptoma entre termes de l'àmbit mèdic: una aproximació basada en corpus[1]

Albert Alegrí
Universitat Pompeu Fabra
albertalegri@gmail.com

Iria da Cunha
Universidad Nacional de Educación a Distancia
iriad@flog.uned.es

Albert Alegrí és màster i graduat en Lingüística Aplicada per la Universitat Pompeu Fabra, especialitzat en Terminologia i Ensenyament de Llengües. Ha participat en la confecció i l'etiquetatge sintàctic del Corpus Treebank Català de l'Institut Universitari de Lingüística Aplicada i actualment és col·laborador de la Càtedra Pompeu Fabra. Treballa en els àmbits de la formació i la comunicació, fent així un parèntesi en la seva recerca doctoral com a membre del grup IULATERM.

Iria da Cunha és doctora en Ciències del Llenguatge i Lingüística Aplicada per la Universitat Pompeu Fabra (UPF). Actualment és investigadora Ramón y Cajal en el Departament de Filologies Estrangeres i les seves Lingüístiques de la Universitat Nacional d'Educació a Distància (UNED), on té la seva recerca adscrita al grup ACTUALing. També és investigadora externa del grup IULATERM de l'Institut Universitari de Lingüística Aplicada de la UPF. És especialista en discurs especialitzat, terminologia i lingüística computacional.

Resum

Els especialistes en terminologia afirmen que els textos especialitzats són, quant al contingut, estructures formades per termes relacionats conceptualment (Cabré, 1999). L'objectiu principal d'aquest article és mostrar que existeixen patrons lingüístics que evidencien la relació malaltia-símptoma i que poden servir per detectar semiautomàticament o automàtica aquesta relació, mitjançant una metodologia concreta. Aquesta metodologia es basa en el fet que, un cop obtinguts i analitzats uns determinats contextos, s'extreuen marques lingüístiques que evidenciïn la relació i serveixin per realitzar una generalització de patrons lingüístics. En l'estudi observem que molts dels símptomes no apareixen en les definicions dels recursos terminològics mèdics, però sí en els textos especialitzats del corpus. També detectem que molts dels símptomes apareixen mitjançant col·locacions especialitzades.

PARAULES CLAU: *patrons lingüístics; detecció de relacions; malaltia; símptoma; marques lingüístiques; col·locacions especialitzades*

Abstract

Detection of the disease-symptom relation between terms in the medical sphere: a corpus-based approximation

Experts in terminology state that specialised texts are, with respect to their content, structures formed by terms, and these terms are conceptually connected (Cabré, 1999). The main goal of this article is to prove the existence of linguistic patterns that show the disease–symptom relation. These patterns can be useful to detect the relation semiautomatically or automatically by means of a specific methodology. This methodology is based on the fact that, once we extract and analyse certain contexts, we can obtain linguistic marks that show this relation, allowing a generalisation of patterns. Our results show that most symptoms are not included in the definitions given in medical terminological resources, but they are indeed included in the medical corpus that has been analysed in this project. Moreover, most symptoms are described with specialised collocations.

KEYWORDS: *language patterns; relation detection; disease; symptom; linguistic marks; specialised collocations*

TERMINÀLIA 12 (2015): 7-17 · DOI: 10.2436/20.2503.01.79
Data de recepció: 15/01/2015. Data d'acceptació: 15/04/2015
ISSN: 2013-6692 (impresa); 2013-6706 (electrònica) · http://terminalia.iec.cat

1 Introducció

La terminologia és un àmbit d'estudi en el qual es parteix de la idea que els textos especialitzats estan plens de termes estructurats conceptualment d'una forma determinada (Cabré, 1999). Els termes són unitats lèxiques que adquireixen un sentit especialitzat en un àmbit determinat (Cabré, 1999). En aquesta recerca ens centrem en l'àmbit especialitzat de la medicina, ja que els investigadors d'aquest àmbit detecten amb freqüència noves malalties i nous símptomes de malalties, tant d'existents com d'aparició recent; és a dir, l'estudi de les malalties s'actualitza contínuament. Si es construeix una estructuració conceptual, com per exemple un arbre de camp (figura 1), d'una malaltia, podrien trobar-se diferents branques, com serien les causes de la malaltia en qüestió, els tipus o classificacions, els símptomes que l'evidencien, les tècniques o les estratègies per les quals es pot diagnosticar, i els possibles tractaments. Generalment, l'expressió de les malalties —les causes, la tipologia, els símptomes, el diagnòstic i el tractament— es vehicula mitjançant unitats terminològiques.

La realitat actual és que hi ha pocs treballs sobre la detecció de relacions a partir de textos escrits en català, com per exemple el de Feliu (2004). Tanmateix, no n'hi ha cap que tracti una relació tan específica com la relació en la qual ens centrem en aquesta recerca (la relació malaltia-símptoma entre termes), i aquesta manca de treballs ha estat la principal motivació per dur a terme aquesta recerca. Volem proporcionar una anàlisi lingüística del funcionament de la llengua catalana amb referència a la relació en qüestió. La nostra hipòtesi de partida és que als textos especialitzats de l'àmbit mèdic existeixen determinades marques lingüístiques que permetrien detectar de manera automàtica o semiautomàtica la relació que s'estableix entre un terme que es refereix a una malaltia i els termes que fan referència als símptomes d'aquesta malaltia, tal com indica, per exemple, Hearst (1992). La detecció d'aquest tipus de relacions podria ser de gran utilitat per a la construcció de tesaurus o bases de dades lexicosemàntiques, i per a l'actualització de les definicions de diccionaris terminològics de l'àmbit mèdic, així com també de manuals.

Concretament, doncs, els objectius d'aquesta recerca, emmarcada dins d'una tesi de màster, són els següents:

- Crear patrons lingüístics que evidenciïn la relació malaltia-símptoma entre termes a partir de textos mèdics escrits en català, que puguin servir per detectar de manera semiautomàtica o automàtica aquesta relació.
- Proposar una metodologia concreta per dur a terme la creació de patrons lingüístics de manera semiautomàtica.

Després de la introducció, en l'apartat 2 es presenta el marc teòric, que gira entorn de la *teoria comunicativa de la terminologia* (TCT) de Cabré (1999). En l'apartat 3, on es planteja l'estat de la qüestió, es parla dels treballs existents sobre la detecció automàtica i semiautomàtica de relacions conceptuals i semàntiques. En l'apartat 4 es llisten tots els passos seguits fins a arribar a l'obtenció dels resultats finals, els quals es descriuen i s'analitzen en l'apartat 5. En l'apartat 6 s'inclouen les idees principals que es desprenen del treball realitzat, quines en són les aportacions cap a la terminologia, així com quines en poden ser les possibles futures vies de recerca.

2 Marc teòric

El treball terminològic sobre la base de corpus especialitzats ha demostrat que el nombre i la tipologia de relacions conceptuals establertes per l'enfocament tradicional de la terminologia no són prou representatius (Cabré, 1999; Feliu i Cabré, 2002). Per aquest motiu, Cabré (1999) afirma que, d'una banda, l'objectiu principal de la terminologia teòrica és descriure formalment, semàntica i funcional les unitats que poden adquirir el valor terminològic, subratllar la manera com activen aquest valor i explicar de quina manera es relacionen amb els contextos veïns. De l'altra, l'objectiu de la terminologia aplicada és la recopilació d'aquests elements. L'anàlisi corresponent té finalitats aplicades molt diverses; tanmateix, en totes elles apareix la doble funcionalitat dels termes: la representació del coneixement especialitzat i la seva transferència, de maneres i en situacions diferents. És per aquest motiu que Cabré descriu un conjunt de reflexions que

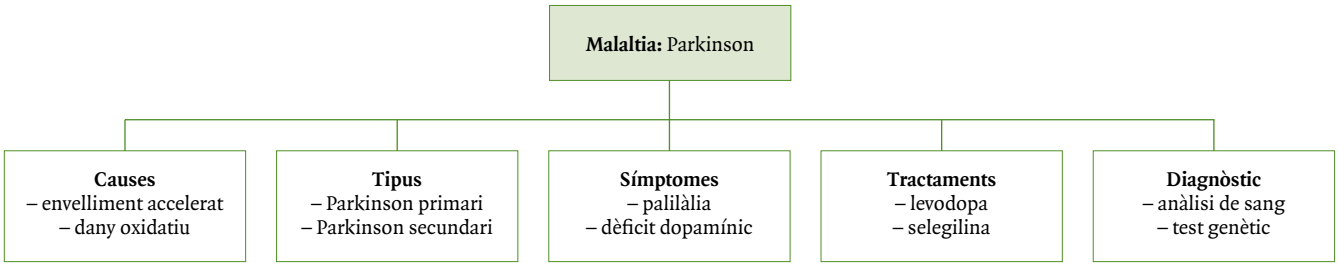

FIGURA 1. Arbre de camp del Parkinson

engloba en una nova teoria, l'anomenada *teoria comunicativa de la terminologia* (TCT), base teòrica per a aquesta recerca.

Cabré (1999) subratlla que cada situació comunicativa és distinta, davant la varietat temàtica i el nivell d'especialització que poden oferir els textos d'especialitat. A causa d'aquesta àmplia varietat possible, la propietat terminològica sempre apareix en situacions de comunicació natural, les quals són potencialment socials, és a dir, en cadascuna d'elles hi ha unes pretensions concretes del parlant.

L'autora defineix les unitats terminològiques (UT) com a unitats que formen part del llenguatge natural i de la gramàtica que descriu cada llengua, i són l'objecte principal de la terminologia. Cada UT té diferents facetes i s'emmarca dins d'un pla concret d'anàlisi de la cognició i de la comunicació alhora, raons per les quals són interdisciplinàries i multidimensionals.

Ateses aquestes descripcions, es pot afirmar que els termes són unitats lèxiques, activades d'una manera determinada per les seves condicions pragmàtiques d'adequació a un tipus de comunicació concreta en cada cas. La importància que té en tot moment el context ha provocat que, en terminologia, d'una banda, es parli de relacions conceptuals i, de l'altra, de relacions semàntiques, discussió oberta des de fa anys. Les relacions conceptuals són imprescindibles per establir sistemes que reflecteixin l'estructuració del coneixement especialitzat, mentre que, segons Evans (1988), les relacions semàntiques són aquelles que connecten conceptes. Murphy (2003), per la seva banda, afirma que les relacions semàntiques són aquelles que poden definir-se a través de paradigmes que fan referència únicament al significat individual, aquells que es troben en el lexicó mental.

3 Estat de la qüestió

En els treballs existents sobre la detecció automàtica o semiautomàtica de relacions entre termes es descriuen principalment dues estratègies: les que són purament lingüístiques, en les quals es duu a terme una descripció de patrons que expressen relacions concretes, segons les estructures sintàctiques; i les estadístiques, que fan servir principalment l'aprenentatge automàtic (*machine learning*). Cal subratllar que aquestes dues estratègies són, en molts casos, complementàries/compatibles. Tot i que oferim una panoràmica dels dos tipus, no és la nostra intenció oferir una descripció exhaustiva dels treballs realitzats, sinó únicament d'alguns dels més representatius.

Sobre l'estudi de patrons que evidencien l'existència de relacions conceptuals en textos, ens centrem en la tesi doctoral de Feliu (2004). L'autora pretén establir les bases per desenvolupar un sistema de detecció semiautomàtica de relacions conceptuals a partir de textos especialitzats. Ho fa detectant marcadors lingüístics verbals, uns de concrets per a cada tipus de relació conceptual; així, per exemple, selecciona els verbs *manifestar*, *mostrar* i *presentar* per tractar la relació de seqüencialitat temporal de simultaneïtat. El pas següent és extreure fragments de textos que continguin unitats de coneixement especialitzat, lligades mitjançant termes i relacions conceptuals.

Feliu estableix estratègies sintacticosemàntiques útils per identificar amb el màxim de precisió els marcadors que expressen cada tipus de relació. Per tant, no només s'estudia el patró sintàctic evidenciat en la relació, sinó que també se subratlla el factor del context, és a dir, l'estratègia semàntica que permet determinar amb fiabilitat si un marcador expressa un tipus de relació concret o un altre, segons les paraules que el formen. L'estratègia semàntica s'aplica amb l'ajuda d'una ontologia de l'àmbit mèdic, per així classificar i estructurar el lèxic especialitzat extret del corpus. Feliu justifica que l'aplicació d'estratègies tant sintàctiques com semàntiques es duu a terme per obtenir fragments textuals especialitzats que continguin, com a mínim, dues unitats terminològiques lligades entre si mitjançant un tipus de relació conceptual.

Quant a les estratègies estadístiques, tot i no ser les estudiades en aquesta recerca, en la taula 1 oferim una lleugera descripció d'alguns dels treballs més representatius, ordenats cronològicament. En la primera columna apareixen els autors de cada recerca. En la segona columna, l'any de la seva realització. En la tercera columna, els tipus de relacions estudiades en cada treball. En la quarta i última columna, les tècniques en què els autors s'han centrat en cada estudi.

La descripció de diversos treballs representatius sobre la detecció de relacions en corpus textuals demostra que n'existeixen molt pocs en català i concretament cap que estudiï la relació malaltia-símptoma en textos de l'àmbit mèdic. En aquest treball posem èmfasi en l'estratègia purament lingüística, atesa la motivació per un estudi i una descripció de la llengua catalana.

4 Metodologia

En aquest apartat s'inclouen, ordenats cronològicament, tots els passos seguits a l'hora de desenvolupar el treball:

1) Seleccionar un corpus de treball adequat per a l'estudi en qüestió. En aquest cas, dins del Corpus Tècnic (CT) de l'Institut Universitari de Lingüística Aplicada (IULA) (Vivaldi, 2009), el subcorpus de medicina, format per textos especialitzats en català, és adient per a aquesta recerca, ja que és un corpus etiquetat pel que fa a marques estructurals, així com validat per professionals de l'àmbit. Aquest subcorpus està format per 236 documents; en total, 2.677.138 paraules.

2) Cercar i consultar tres fonts especialitzades; concretament el *Medline Plus* (*http://www.nlm.nih.gov/medlineplus/spanish*), el manual *Medicina interna* (Farreras i Rozman, 2002) i el *Diccionari enciclopèdic de medicina*

Autor/s	*Any*	*Relació*	*Tècnica*
Hearst	1992	Hiperonímia	Inventari de patrons lexicosintàctics per identificar, amb aprenentatge automàtic, parelles de paraules que exemplifiquin la relació estudiada.
Berland i Charniak	1999	Meronímia	Mètode estadístic amb tres fases: identificació i anotació de patrons, filtratge de paraules acabades morfològicament d'unes maneres determinades i ordenació de les parelles de paraules segons el valor F.
Riloff i Jones	1999	Sis categories semàntiques: construcció, esdeveniments, ésser humà, localització, temps, armes	Algoritme de *bootstrapping* que aprèn trets semàntics de diverses categories.
Snow, Jurafsky i Ng	2004	Hiperonímia i hiponímia	Informació de dependències per a la creació d'un classificador automàtic.
Hasegawa, Sekine i Grishman	2004	Homonímia i sinonímia	Aglomeració de parelles d'entitats (*clustering*).
Turney	2006	Parelles anàlogues	Mètode de mesura de la similitud relacional, *vector space model* (VSM).
Hendrickx *et al.*	2009	Llista de parelles de relacions semàntiques incloses en un inventari	Desenvolupament d'un classificador automàtic que treballa sobre l'inventari de relacions creat per a l'estudi.
Mintz *et al.*	2009	Conjunt de trenta-tres relacions binàries en diferents dominis (persona-nacionalitat, pel·lícula-director, geografia-riu, etc.)	Informació de dependències per a la creació d'un classificador automàtic.
Necsulescu, Mendes i Bel	2014	Homonímia i sinonímia	Model d'aprenentatge automàtic dels trets (*features*) de cada parella per proporcionar tuples.

TAULA 1. Recull de treballs sobre detecció de relacions amb estratègies estadístiques

(*http://www.medic.cat*), per tal de determinar quin grup de malalties s'estudia.

3) Es treballa amb malalties neurodegeneratives perquè són un tipus de patologies en les quals el problema principal se situa en el moment de la seva detecció, ja que, en la major part dels casos, quan s'identifiquen, els malalts ja han perdut més de la meitat de les neurones. Diagnosticar, doncs, aquestes demències de la manera més anticipada possible és imprescindible avui en dia i, per aquest motiu, considerem que la detecció dels símptomes és un tema rellevant. Dins del grup de malalties neurodegeneratives, se selecciona un grup de termes en català que siguin malalties concretes d'aquest grup. Després de consultar les fonts espanyoles comentades en el pas 2, observem que hi ha nou termes que es corresponen amb malalties neurodegeneratives: Parkinson, corea de Huntington, Alzheimer, esclerosi lateral amiotròfica, atròfia muscular, idiòcia amauròtica de Tay-Sachs, esfingolipidosi, atàxia de Friedreich i malaltia de Pick.

4) Mitjançant el *Diccionari enciclopèdic de medicina*, se seleccionen les definicions completes de cadascun d'aquests termes i se n'extreuen només aquells fragments amb símptomes. L'exemple següent inclou una de les malalties, la seva definició i els símptomes marcats en negreta:

> corea de Huntington:
> Procés hereditari, transmès per herència autosòmica dominant, produït per una mutació genètica del braç curt del cromosoma 4, caracteritzat per **discinèsies coreoatetòtiques de la musculatura axial i perifèrica**, **disàrtria**, **disfàgia**, **postures distòniques** i per **trastorns neuropsicològics** variables (**modificacions de la personalitat**, **depressió**, **psicosi**) que evolucionen vers la demència.

5) Mitjançant BwanaNet, una interfície que permet consultar el CT en línia, s'extreuen automàticament tots els contextos oracionals per a cadascun dels sis termes (malalties). Dels nou termes inicials, finalment en seleccionem sis (els sis primers de la llista del punt 3). Un dels criteris per a la selecció de les malalties és que cada terme aparegui en un mínim de deu contextos al corpus de treball.

6) S'analitzen manualment els contextos obtinguts de manera automàtica. Tots els contextos que no contenen termes referents a símptomes són eliminats. Per detectar els símptomes associats a cada malaltia, es consulten els símptomes extrets de les definicions de les malalties al *Diccionari enciclopèdic de medicina* i també s'aplica el coneixement propi, mitjançant evidències lingüístiques que explicitin que es parla de

la malaltia en qüestió. A continuació, mostrem un exemple de context del terme *Parkinson* obtingut del corpus amb els símptomes destacats en negreta:

> En la malaltia de Parkinson, que cursa amb un **dèficit dopamínic** a nivell de sistema nerviós central, s'han descrit **alteracions en el reflex *orbicularis oculi***, concretament del segon component d'aquest reflex.

7) S'analitzen els contextos definitius per tal de detectar totes les marques lingüístiques que evidencien que existeix una relació malaltia-símptoma. Els diferents continguts dins dels contextos es marquen de la manera següent:

- El terme corresponent a la malaltia: entre coixinets.
- El terme o termes corresponents als símptomes: subratllats.
- Les marques lingüístiques explícites: en negreta.

Tot seguit mostrem dos exemples amb les marques comentades:

> <mooo21>: <s>També més excepcionalment poden presentar **crisis epilèptiques** el ##Parkinson##, la corea de Hungtington i la degeneració hepatolenticular.</s>
> < moo623>: <s>La malaltia d'## Alzheimer ## es caracteritza per l'**aparició de dipòsits de plaques amiloides** i la **formació de feixos neurofibril·lars en el cervell**.</s>

8) Es fa una generalització dels patrons lingüístics detectats fent servir lemes i s'utilitzen els símbols següents:

- [S] fa referència al símptoma.
- [M] fa referència a la malaltia.
- Per exemple: poder presentar [S] el (malaltia de / del / de la) [M].
- L'asterisc (*) indica que en aquella posició ha d'aparèixer informació textual, tot i que cap en concret.

5 Anàlisi i resultats

En aquest apartat mostrem les anàlisis que s'han realitzat en aquest treball i els resultats obtinguts després d'aplicar els passos inclosos en la metodologia explicada en l'apartat 4.

En la taula 2 s'aporten dades quantitatives en relació amb el nombre total de contextos obtinguts per a cada malaltia i el nombre de contextos que contenen símptomes d'aquestes. Si s'analitza la tercera columna, s'aprecia que s'han eliminat molts contextos, pel fet que no contenien símptomes.

Malaltia	*Contextos totals*		*Contextos seleccionats amb símptomes*		
	Nombres absoluts	*%*	*Nombres absoluts*	*%*	*% de contextos amb símptomes respecte del nombre absolut de contextos trobats per a cada malaltia*
Parkinson	50	18,31	11	47,82	22,00
(corea de) Huntington	33	12,09	1	4,35	3,03
Alzheimer	144	52,75	9	39,13	6,25
esclerosi lateral amiotròfica	12	4,40	1	4,35	8,30
atròfia muscular	10	3,66	1	4,35	10,00
(idiòcia amauròtica de) Tay-Sachs	24	8,79	0	0	0
TOTAL	273	100	23	100	–

TAULA 2. Taula quantitativa dels contextos analitzats

La primera idea, de cara a l'extracció de contextos productius per a l'estudi, era tractar únicament aquells que incloguessin termes per expressar els símptomes, tant monolèxics com polilèxics. Tanmateix, tal com es pot observar en l'annex I, hem detectat que aquests símptomes també poden ser expressats als textos especialitzats mitjançant fraseologia especialitzada o col·locacions de l'àmbit, en aquest cas el de la medicina. Així, subratllem que en 15 dels 31 contextos extrets, és a dir, en gairebé la meitat del total (48,4 %), els símptomes es reflecteixen mitjançant col·locacions especialitzades formades per:

- un nom deverbal que, en la majoria dels casos, indica algun tipus de gradació (ex. *augment*, *deficiència*);
- una preposició com a lligam, principalment *de*;
- un terme mèdic monolèxic o polilèxic (ex. *dopamina*, *feixos neurofibril·lars*).

Per exemple: *dèficit de dopamina*, *formació de feixos neurofibril·lars* (vegeu-ne més a l'annex 3).

Amb referència a l'anàlisi del contrast entre els símptomes trobats en les definicions i els contextos, han participat en l'estudi un metge i una farmacèutica de la ciutat de Figueres (Alt Empordà). La seva tasca va ser validar les relacions existents entre els símptomes i determinar, així, si hi ha casos de variació terminològica. Aquest contrast s'ha realitzat a partir de l'observació que cap dels símptomes inclosos als contextos del corpus no coincideix amb els símptomes que apareixen en les definicions, tal com es pot veure en l'annex I.

Vegem ara un cas de variació terminològica: en el cas del Parkinson, per exemple, segons la definició del *Diccionari enciclopèdic de medicina*, *amímia* és la «pèrdua més o menys completa de la facultat d'expressió mímica», és a dir, es caracteritza per una manca d'expressió

tant de la cara com del cos, mentre que en el cas de la paràlisi pseudobulbar, la limitació en l'expressió és, de manera més concreta, facial. Per tant, tot i aquest matís, els dos símptomes es poden considerar pràcticament sinònims. Els dos especialistes que han participat en l'estudi ens han corroborat aquesta anàlisi.

Seguidament, en la taula 3, mostrem l'anàlisi quantitativa d'aquest contrast.

El fet que no coincideixi cap símptoma entre els trobats en les definicions i els contextos ens fa pensar en dues possibles causes. D'una banda, la variació terminològica. La falta de consens a l'hora d'expressar aspectes concrets d'un àmbit especialitzat fa que es produeixi aquesta variació, la qual afecta directament els termes. Corbeil (1988, p. 57) afirma que:

> Le résultat le plus apparent et le plus embarrassant de cette variation est l'incertitude terminologique, soit que plusieurs dénominations semblent correspondre plus ou moins à la même notion (concurrence terminologique), soit que la même dénomination semble correspondre à des notions différents, en tout ou en partie (polysémie terminologique).

Malaltia	*Nre. de símptomes al diccionari*	*Nre. de símptomes dels contextos*	
		Variants terminològiques dels símptomes trobats als diccionaris	*Altres símptomes no inclosos als diccionaris*
Parkinson	5	1	16
(corea de) Huntington	8	2	0
Alzheimer	2	3	7
esclerosi lateral amiotròfica	4	0	1
atròfia muscular	11	1	0
(idiòcia amauròtica de) Tay-Sachs	5	0	0
TOTAL	35	7	24

TAULA 3. Taula quantitativa sobre els símptomes en les definicions i els contextos

De l'altra, una altra causa podria ser que als textos especialitzats hi hagi termes que es refereixin a símptomes que no apareguin reflectits als diccionaris especialitzats, a causa dels estudis permanents que es desenvolupen contínuament. La ciència avança molt ràpid i, per consegüent, a diferència dels diccionaris que tarden a actualitzar-se, els textos que tracten noves recerques i nous descobriments són una destacada font d'innovació. Aquest fet indica que els textos especialitzats són molt útils per a la recuperació de símptomes de malalties no inclosos en recursos terminològics i, per tant, la nostra proposta per a la detecció podria servir per a l'actualització tant de diccionaris especialitzats com de manuals de l'àmbit.

Amb referència als patrons finals obtinguts, els quals es poden observar en la taula 4, cal subratllar que apareixen tant verbs generals com d'altres que són purament terminològics. Alguns verbs generals són *patir*, *caracteritzar* o *associar*, mentre que d'altres posseeixen una càrrega semàntica concreta, ja que els seus lexemes i significats estan vinculats a un àmbit d'especialitat; són els que Lorente (2001) cataloga com a *verbs terminològics*. Així, per exemple, *cursar* és un exemple de verb terminològic. De l'estructura sintàctica, destaquem l'ús d'algunes oracions de relatiu, així com l'alta aparició de preposicions, les quals demostren que normalment són un complement indispensable per als noms deverbals.

1. poder presentar [S] el (malaltia de / del / de la) [M]
2. [S]: aparèixer com a símptoma en el (malaltia de / del / de la) [M]
3. [S] associat a / al / a la (malaltia de / del / de la) [M]
4. [M], que cursar amb un [S]
5. en [M] descriure [S]
6. en (la malaltia de / del / de la) [M] interpretar-se (com l'expressió d'un) [S]
7. malaltia en què existeix [S] com en (la malaltia de / del / de la) [M]
8. el disseny de (la malaltia de / del / de la) [M] haver de tenir en compte els símptomes com [S]
9. un dels problemes que presenta (la malaltia de / del / de la) [M] ser [S]
10. en (el cas de) la malaltia de / del / de la [M] la simptomatologia es produeix quan hi ha [S]
11. a * de malalts de [M] detectar [S]
12. el [S] estar implicat en (nombroses malalties neurodegeneratives humanes com són) el M
13. [S] associat a /al / a la (malaltia de / del / de la) [M]
14. [M] * caracteritzar-se per (la presència de) [S]
15. en la malaltia de / del / de la [M] haver-hi un [S]
16. el [S] ser característic de (la demència tipus) [M]
17. el pacient (afectat per la malaltia) de [M] presentar [S]
18. la malaltia de / del / de la [M] caracteritzar-se pel / per la [S]
19. el * més característic del / de la [M] és [S]
20. l'adult amb (la malaltia de / del / de la) [M] presentar [S]
21. la malaltia de / del / de la [M] caracteritzar-se per [S]
22. el [S] estar implicat en (nombroses malalties neurodegeneratives humanes com són) el [M]
23. [S] associat a /al / a la (malaltia de / del / de la) [M]
24. algun pacient patir un [S] denominat [M]

TAULA 4. Llista final de patrons lematitzats

6 Conclusions i treball futur

Un cop arribat al final d'aquesta recerca, respecte del primer objectiu, podem afirmar que la recerca ha servit per confirmar que efectivament hi ha patrons lingüístics que evidencien una relació concreta en català. La llista obtinguda podrà ser alliberada per tots aquells qui la vulguin aplicar a nous textos. Pel que fa al segon objectiu, la metodologia proposada ha servit per representar una llista final de patrons, que pot ser útil per aplicar a futurs treballs més complexos sobre relacions conceptuals.

En la descripció lingüística de les unitats que vehiculen el coneixement especialitzat en l'àmbit de la medicina, l'anàlisi dels resultats ha permès detectar que els símptomes apareixen reflectits dins dels textos especialitzats no només mitjançant termes monolèxics i polilèxics, sinó també mitjançant col·locacions especialitzades el nucli de les quals és un nom deverbal que indica, generalment, algun tipus de gradació. A més, s'evidencia que molts dels símptomes de les malalties no apareixen en les definicions dels recursos terminològics de l'àmbit de la medicina, però sí als textos especialitzats del corpus utilitzat, a causa dels continus estudis que es realitzen en medicina i que serveixen per actualitzar els coneixements sobre aquest àmbit.

Podem considerar que la metodologia concreta proposada per dur a terme l'extracció d'una relació en un corpus textual pot servir per analitzar altres tipus de relacions com les causes, les tipologies, els tractaments i els diagnòstics. És per això que, com a futures línies de recerca, proposaríem determinar la cobertura i la precisió dels patrons de la taula 4, és a dir, quins són més productius, per implementar-los en forma d'expressions regulars; explorar la possibilitat de combinació d'estratègies lingüístiques basades en patrons i estadístiques de *machine learning*; desenvolupar un prototip de sistema que pugui detectar relacions entre termes de l'àmbit mèdic per a la creació automàtica o semiautomàtica d'estructures conceptuals (concretament, d'arbres de camp) de diferents malalties, mitjançant l'explotació de corpus de textos especialitzats i d'estratègies basades en patrons lingüístics, i realitzar experiments per comprovar que la metodologia presentada es pot extrapolar a altres llengües.

Bibliografia

Berland, Matthew; Charniak, Eugene (1999). «Finding parts in very large corpora». A: *Proceedings of ACL, Association for Computational Linguistics*, p. 57-64.

Cabré, Maria Teresa (1999). *La terminología: Representación y comunicación: Una teoría de base comunicativa y otros artículos*. Barcelona: Universitat Pompeu Fabra. Institut Universitari de Lingüística Aplicada. (Sèrie Monografies; 3)

Casassas, Oriol (dir.) (1990). *Diccionari enciclopèdic de medicina* [en línia]. Barcelona: Enciclopèdia Catalana: Acadèmia de Ciències Mèdiques de Catalunya i de Balears. <http://www.medic.cat/> [Consulta: 18 març 2014].

Corbeil, Jean Claude (1988). «Les terminologies devant Babel». A: *Actes du Colloque Terminologie et Technologies Nouvelles. La Défense*. Quebec: Office de la Langue Française, p. 49-62.

Evans, Martha Walton (ed.) (1988). *Relational Models of the Lexicon: Representing Knowledge in Semantic Networks*. Cambridge: Cambridge University Press.

Farreras, Pedro; Rozman, Ciril (2002). *Medicina interna*. 17a ed. Barcelona: Elsevier.

Feliu, Judit (2004). *Relacions conceptuals i terminologia: anàlisi i proposta de detecció semiautomàtica*. Tesi doctoral. Barcelona: Universitat Pompeu Fabra. Institut Universitari de Lingüística Aplicada.

Feliu, Judit; Cabré, Maria Teresa (2002). «Conceptual relations in specalized texts: new typology and an extraction system proposal». A: *TKE2002: Terminology and Knowledge Engineering Proceedings: 6th International Conference: 28th-30th August 2002* (Nancy), p. 45-49.

Hasegawa, Takaaki; Sekine, Satoshi; Grishman, Ralph (2004). «Discovering relations among named entities from large corpora». A: *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*. East Stroudsburg: Association for Computational Linguistics, p. 415-422.

Hearst, Marti A. (1992). «Automatic acquisition of hyponyms from large text corpora». A: *Proceedings of COLING*. Grenoble: IVR Imprimerie, p. 539-545.

Hendrickx, Iris; Nam Kim, Su; Kozareva, Zornista; Nakov, Presalv; Ó Séaghdha, Diarmuid; Padó, Sebastian; Pennacchiotti, Marco; Romano, Lorenza; Szpakowicz, Stan (2009). «SemEval-2010 Task 8: multi-way classification of semantic relations between pairs of nominals». A: *SemEval'10: Proceedings of the 5th International Workshop on Semantic Evaluation*. Stroudsburg: Association for Computational Linguistics, p. 33-38.

Lorente, Mercè (2001). «Teoría e innovación en terminografía: la definición terminográfica». A: *La terminología científico-técnica: Reconocimiento, análisis y extracción de información formal y semántica*. Barcelona: Universitat Pompeu Fabra. Institut Universitari de Lingüística Aplicada, p. 81-112.

MINTZ, Mike; BILLS, Steven; SNOW, Rion; JURAFSKY, Dan (2009). «Distant supervision for relation extraction without labeled data». A: *Proceedings of ACL-CONLL, Association for Computational Linguistics*, p. 1003-1011.
MURPHY, M. Lynne (2003). *Semantic relations and the lexicon: Antonym, synonymy, and other paradigms*. Cambridge: Cambridge University Press.
NECSULESCU, Silvia; MENDES, Sara; BEL, Núria (2014). «Combining dependency information and generalization in a pattern-based approach to the classification of lexical-semantic relation instances». Barcelona: Universitat Pompeu Fabra. Institut Universitari de Lingüística Aplicada; Lisboa: Centro de Lingüística da Universidade de Lisboa.
RILOFF, Ellen; JONES, Rosie (1999). «Learning dictionaries for information extraction by multi-level bootstrapping». A: *Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI-99)*, p. 474-479.
SNOW, Rion; JURAFSKY, Daniel; NG, Andrew Yan-Tak. (2004). «Learning syntactic patterns for automatic hypernym discovery». A: *Proceedings of* NPS. Cambridge, Mass: MIT Press.
TURNEY, Peter D. (2006). «Similarity of semantic relations». *Computational Linguistics*, 32(3), p. 379-416.
VIVALDI, Jordi (2009). «Corpus and exploitation tool: IULACT and bwanaNet». A: CANTOS GÓMEZ, Pascual; SÁNCHEZ PÉREZ, Aquilino (ed.). *A survey on corpus-based research = Panorama de investigaciones basadas en corpus.* [Actas del I Congreso Internacional de Lingüística de Corpus (CICL-09) (Universidad de Murcia, 7-9 de Mayo de 2009)]. Múrcia: Asociación Española de Lingüística del Corpus, p. 224-239.

Nota

1. Aquest treball ha estat finançat parcialment pel projecte de recerca APLE (FFI2009-12188-C05-01), dirigit per la Dra. M. Teresa Cabré, i per un contracte Ramón y Cajal (RYC-2014-16935).

ANNEXOS

ANNEX I. Classificació dels símptomes dels contextos segons si són variants terminològiques o símptomes no inclosos en les definicions del diccionari

Malaltia	*Símptomes diccionari*	*Símptomes dels contextos*	
		Variants terminològiques dels símptomes trobats als diccionaris	*Altres símptomes no inclosos als diccionaris*
Parkinson	tremolor especial		crisis epilèptiques
	moviments de dits		palilàlia
	rigidesa muscular		demència
	amímia	paràlisi pseudobulbar	alteracions en el reflex *orbicularis oculi*
	marxa amb el cos, les cames i els braços en lleugera flexió		augment de l'excitabilitat de les neurones de la substància reticular bulbar lateral
			alteració dels sistemes neuronals dopaminèrgics
			insomni
			ansietat
			depressió
			vòmits
			estrenyiment
			dèficit de dopamina
			deficiència del complex
			increment del dany oxidatiu
			estrès oxidatiu

Malaltia	*Símptomes diccionari*	*Símptomes dels contextos*	
		Variants terminològiques dels símptomes trobats als diccionaris	*Altres símptomes no inclosos als diccionaris*
(corea de) Huntington	discinèsies coreoatetòtiques de la musculatura axial i perifèrica	moviments coreics	
	disàrtria		
	disfàgia		
	postures distòniques		
	trastorns neuropsicològics	trastorns de tipus afectiu amb deteriorament cognitiu	
	modificacions de la personalitat		
	depressió		
	psicosi		
Alzheimer	acumulació intraneuronal d'un material filamentós	placa senil	pèrdua progressiva de les funcions intel·lectuals
		aparició de dipòsits de plaques amiloides	dificultats per a les tasques de la vida quotidiana
	presència de grumolls de dendrites i axons a l'escorça cerebral	neurones caracteritzades per presentar una arborització dendrítica molt pobra	pèrdua progressiva de la memòria
			pèrdua progressiva de capacitats cognitives
			depleció del sistema colinèrgic
			formació de feixos neurofibril·lars en el cervell
			estrès oxidatiu
esclerosi lateral amiotròfica	síndrome piramidal		estrès oxidatiu
	paràlisi perifèrica		
	paràlisi espàstica dels membres inferiors		
	atròfia muscular progressiva		
atròfia muscular	degeneració de les anorexianeurones de les banyes anteriors de la medul·la espinal		
	hipotonia muscular greu	debilitat muscular progressiva	
	síndrome de l'infant pengim-penjam		
	areflèxia		
	plor		
	succió		
	afectació dels músculs respiratoris		

Malaltia	*Símptomes diccionari*	*Símptomes dels contextos*	
		Variants terminològiques dels símptomes trobats als diccionaris	*Altres símptomes no inclosos als diccionaris*
atròfia muscular (cont.)	insuficiència respiratòria infeccions pulmonars iteratives deformitats articulars congènites afectació del nervi facial		
(idiòcia amauròtica de) Tay-Sachs	excés d'esfingosina atròfia amauròtica del nervi òptic hipertonia muscular extrapiramidal insuficiència visual progressiva disfàgia		

Annex II. Símptomes extrets de les definicions i símptomes extrets dels contextos

Malalties	*Símptomes de les definicions del diccionari*	*Símptomes dels contextos del corpus*
Parkinson	• tremolor especial • moviments de dits • rigidesa muscular • amímia • marxa amb el cos, les cames i els braços en lleugera flexió	• crisis epilèptiques • palilàlia • paràlisi pseudobulbar • demència • dèficit dopamínic • alteracions en el reflex orbicularis oculi • augment de l'excitabilitat de les neurones de la substància reticular bulbar lateral • alteració dels sistemes neuronals dopaminèrgics • insomni • ansietat • depressió • vòmits • estrenyiment • dèficit de dopamina • deficiència del complex • increment del dany oxidatiu • estrès oxidatiu
(corea de) Huntington	• discinèsies coreoatetòtiques de la musculatura axial i perifèrica • disàrtria • disfàgia • postures distòniques • trastorns neuropsicològics • modificacions de la personalitat • depressió • psicosi	• moviments coreics • trastorns de tipus afectiu amb deteriorament cognitiu

Malalties	*Símptomes de les definicions del diccionari*	*Símptomes dels contextos del corpus*
Alzheimer	• acumulació intraneutronal d'un material filamentós • presència de grumolls de dendrites i axons a l'escorça cerebral	• pèrdua progressiva de les funcions intel·lectuals • placa senil • dificultats per a les tasques de la vida quotidiana • pèrdua progressiva de la memòria • pèrdua progressiva de capacitats cognitives • depleció del sistema colinèrgic • neurones caracteritzades per presentar una arborització dendrítica molt pobra • aparició de dipòsits de plaques amiloides • formació de feixos neurofibril·lars en el cervell • estrès oxidatiu
esclerosi lateral amiotròfica	• síndrome piramidal • paràlisis perifèriques • paràlisi espàstica dels membres inferiors • atròfia muscular progressiva	• estrès oxidatiu
atròfia muscular	• degeneració de les anorèxianeurones de les banyes anteriors de la medul·la espinal • hipotonia muscular greu • síndrome de l'infant pengim-penjam • areflèxia • plor • succió • afectació dels músculs respiratoris • insuficiència respiratòria • infeccions pulmonars iteratives • deformitats articulars congènites • afectació del nervi facial	• debilitat muscular progressiva
(idiòcia amauròtica de) Tay-Sachs	• excés d'esfingosina • atròfia amauròtica del nervi òptic • hipertonia muscular extrapiramidal • insuficiència visual progressiva • disfàgia	

Annex III. Col·locacions especialitzades que reflecteixen símptomes al nostre corpus

alteracions en el reflex *orbicularis oculi*
augment de l'excitabilitat de les neurones de la substància reticular bulbar lateral
debilitat muscular progressiva
deficiència del complex
dèficit de dopamina
depleció del sistema colinèrgic
dificultats per a les tasques de la vida quotidiana
formació de feixos neurofibrilars en el cervell
increment del dany oxidatiu
pèrdua progressiva de les funcions intel·lectuals
trastorns de tipus afectiu amb deteriorament cognitiu