>>> revista tradumàtica
Traducció i Tecnologies de la Informació i la Comunicació
Número 06: Terminologia i traducció : desembre 2008 : http://www.fti.uab.cat/tradumatica/revista :
ISSN: 1578-7559

L'observatori de terminologia Talaia: mètode i processos

Joaquim Moré, Lluís Rius, Mercè Vázquez, Lluís Villarejo
Universitat Oberta de Catalunya

Resum: L'observatori de terminologia Talaia té com a objectiu aplegar unitats neològiques procedents de revistes acadèmiques fent servir eines d'extracció automàtica de terminologia i tècniques de filtratge de tipus lingüístic i estadístic. La combinació de diferents processos permet disposar d'un cabal continu de propostes terminològiques multilingües d'aparició recent en l'àmbit de la societat del coneixement.
Paraules clau: Projecte Talaia, observatori de terminologia, extracció automàtica de terminologia bilingüe, corpus d'especialitat, filtratge terminològic, unitat LP-CP, neologia, assignació automàtica d'equivalents de traducció.

Resumen: El observatorio de terminología Talaia tiene como objetivo recoger unidades neológicas procedentes de revistas académicas utilizando herramientas de extracción automática de terminología y técnicas de filtrado de tipo lingüístico y estadístico. La combinación de diferentes procesos permite disponer de un caudal continuo de propuestas terminológicas multilingües de aparición reciente en el ámbito de la sociedad del conocimiento.
Palabras clave: Proyecto Talaia, observatorio de terminología, extracción automática de terminología bilingüe, corpus de especialidad, filtrado terminológico, unidad LP-CP, neología, asignación automática de equivalentes de traducción.

Abstract: The objective of the Talaia observatory is to collect and organise neological units published in academic journals using linguistic and statistical automatic terminology extraction tools and filtering techniques. The combination of different processes allows for a continual stream of multi-language terminology that has recently appeared in the information society.
Key words: Talaia Project, terminology observatory, automatic bilingual term extraction, specialized corpus, terminology filtering, LP-CP unity, neology, translation equivalent automatic assignation.

1. Introducció

El reconeixement automàtic d'unitats terminològiques, la detecció precoç de neologismes i l'anàlisi de la implantació de les formes normalitzades són alguns dels reptes que encara té pendents actualment el treball terminològic. Aquests tres aspectes són, justament, a la base del projecte Talaia, un observatori de terminologia interdisciplinari i multilingüe creat en el marc d'un conveni de col·laboració entre el Centre de Terminologia Termcat i la Universitat Oberta de Catalunya.

Talaia té com a finalitat aplegar, catalogar i difondre en línia terminologia pròpia de la societat del coneixement, en català, castellà i anglès, a partir del buidatge periòdic i semiautomatitzat de les revistes acadèmiques de la Universitat Oberta de Catalunya: *Artnodes, Digithum*, *IDP Revista d'Internet, Dret i Política*, *RUSC* i *UOC Papers.*

Aquestes cinc revistes estudien els efectes i les influències de l'ús de les TIC en les persones i la societat des d'àmbits de coneixement diferenciats: l'art, la ciència i la tecnologia; les humanitats; el dret i la ciència política; l'ensenyament superior; i la societat del coneixement (Rius, Sánchez i Vázquez, 2007).

Amb la creació d'aquest observatori es vol posar a disposició pública terminologia neològica de qualitat en català, castellà i anglès, a fi de facilitar les comunicacions especialitzades en els àmbits de la docència i la recerca, i afavorir el treball multilingüe en general i la tasca dels professionals de la redacció, la correcció i la traducció. Les llistes de termes contextualitzats que proporcionen les eines de buidatge també han de permetre elaborar estudis sobre neologia emergent o sobre formes que tenen un ús especialment rellevant en la comunitat d'especialistes. També es preveu que Talaia pugui arribar a ser una eina útil per al seguiment de la implantació de la terminologia normalitzada en textos especialitzats en les tres llengües.

En aquest article presentem el mètode de treball inicial que hem portat a terme (apartat 2), l'evolució del procés de treball inicial per a poder disposar d'un material terminològic més adequat a les necessitats de l'observatori Talaia (apartat 3) i els resultats que hem obtingut després de fer el buidatge dels continguts d'una de les revistes acadèmiques de la Universitat (apartat 4). I, finalment, fem una valoració dels resultats obtinguts.

2. Mètode de treball inicial

El procés de buidatge del corpus de les revistes acadèmiques es porta a terme mitjançant el Lexterm,[1] una eina de codi lliure desenvolupada pel Servei Lingüístic i els Estudis de Llengües i Cultures de la UOC que permet l'extracció automàtica de dades textuals, proporciona una elevada rendibilitat del procés de treball i facilita una explotació àgil dels resultats.

L'extracció de candidats a terme del corpus d'especialitat es fa per parells de llengües, de manera que cada terme disposa des del principi d'un equivalent eventual. En aquest sentit, el focus inicial de Talaia se centra en l'explotació dels parells de llengües català-anglès i català-castellà.

La tasca de buidatge de tot el corpus s'estructura en sis fases de treball: en la primera fase es compilen els continguts de la publicació; en la segona fase s'alineen els continguts de la llengua origen amb els de la llengua de destinació per a fer-ne un buidatge automàtic; en la tercera fase s'extreuen els candidats a terme de manera automàtica; en la quarta fase es fa una cerca automàtica dels equivalents de traducció dels candidats a terme que s'han obtingut en la fase anterior; en la cinquena fase es fa una revisió manual final, tant de la llista dels candidats a terme que s'han obtingut de manera automàtica com dels equivalents de traducció corresponents, i, finalment, els termes validats s'incorporen a Talaia. Seguidament presentem detalladament aquestes sis fases de treball.

2.1. Compilació del corpus bilingüe

El contingut dels articles de les revistes acadèmiques de la Universitat consta de fitxes resum, que es publiquen en format HTML i de les quals es recupera el títol, el resum i les paraules clau; i del cos de l'article, que es publica en PDF i del qual es recupera el conjunt de text, les notes a peu de pàgina i el currículum de l'autor. Les fitxes resum es publiquen sempre en català, castellà i anglès, i els articles en almenys una d'aquestes tres llengües.

Per a poder compilar el corpus bilingüe de textos, els articles en format PDF i HTML són convertits a format TXT. Haver de fer la conversió a format TXT es deu al fet que és més fàcil de preparar la paral·lelització i l'extracció del contingut en aquest format que no pas en PDF o HTML.

2.2. Alineació dels documents

A partir del moment en què tots els continguts de la publicació són en format TXT, el pas següent que cal fer és agrupar-los per parells de llengües per a poder-los paral·lelitzar amb la finalitat de recuperar posteriorment els equivalents de traducció dels candidats extrets. El procés d'alineació consisteix a aparellar un segment de text d'un document escrit en una llengua amb el mateix segment de text de la versió del

[1] http://www.linguoc.cat/

document en una altra llengua. El conjunt de segments alineats, generalment frases, constitueix el corpus paral·lel a partir del qual es fa la cerca automàtica de l'equivalent de traducció mitjançant el Lexterm.

2.3. Extracció automàtica de candidats a terme

En el moment en què s'ha constituït el corpus paral·lel comença el procés d'extracció automàtica dels candidats a terme que hi pugui haver en el corpus d'especialitat. Aquesta tasca la portem a terme amb el Lexterm, eina que permet d'extreure per mitjà de mètodes estadístics totes les paraules o combinacions de paraules consecutives (*n-grams*) que hi ha en el corpus d'especialitat a partir d'un llindar de freqüència establert per l'usuari.

Amb l'objectiu de poder afinar els resultats de la llista de candidats a terme, en el procés d'extracció de candidats s'apliquen tres tipus de filtratges, que són descrits a continuació.

2.3.1. Filtratge per longitud

L'eina d'extracció de candidats permet recuperar la llista de candidats a terme d'un corpus d'especialitat a partir de la longitud que determina l'usuari, és a dir, els candidats a terme poden constar d'una paraula o més. Per al català i el castellà, els candidats consten de tres paraules per a poder trobar els equivalents de traducció de les formes angleses, que generalment consten de dues paraules (per exemple, *tool book* correspon en català a *caixa d'eines*).

2.3.2. Filtratge per llindar de freqüència

A l'hora d'extreure la llista de candidats a terme també es té en compte un llindar mínim a partir del qual s'han de situar els resultats. Per aquest motiu, solament se seleccionen els candidats a terme que tenen una freqüència mínima d'aparició de tres vegades en el corpus d'especialitat, i la resta es desestima.

2.3.3. Filtratge de paraules buides en posicions extremes

Per a evitar que la paraula inicial o final d'un candidat a terme sigui una paraula buida de contingut (articles, pronoms, preposicions, conjuncions, etc.), prèviament es prepara un fitxer amb totes les paraules buides que no es volen recuperar si són en una posició inicial o final de l'*n-gram*, perquè l'eina el faci servir de filtre a l'hora de presentar els candidats a terme. Així, s'eliminen combinacions com ara «de la» o «el banc de» de la llista de resultats per a assegurar que les paraules i les combinacions de paraules tinguin una entitat significativa.

2.4. Cerca d'equivalents de traducció

El pas final d'aquest procés de recuperació de terminologia se centra en la cerca dels equivalents de traducció en el corpus paral·lel que ja s'ha preparat en el procés d'alineació inicial.

La cerca automàtica d'equivalents de traducció es duu a terme per a cada un dels candidats a terme que han estat extrets de manera automàtica amb el Lexterm. Aquesta eina cerca en el corpus paral·lel l'*n-gram* en la llengua L2 que és més probable que correspongui a un candidat de la llista en la llengua L1. Els *n-grams* presentats són propostes, ja que en la fase següent els resultats d'aquesta llista són validats per un terminòleg.

2.5. Selecció manual dels candidats a terme i els equivalents de traducció

En la fase final del procés d'extracció de candidats a terme, un terminòleg revisa manualment els resultats obtinguts amb l'objectiu, primerament, de validar el candidat a terme en català i, en segon lloc, de revisar l'equivalent o equivalents més probables de traducció proposats pel Lexterm de manera automàtica. Per a facilitar aquesta tasca, l'eina permet veure els contextos en què apareix el terme original i els contextos en els quals apareix la proposta de traducció. A mesura que es validen els equivalents, es confecciona el glossari bilingüe.

2.6. Incorporació de les propostes terminològiques a l'observatori

En el moment en què ja es disposa de la llista final revisada dels candidats a terme i els equivalents de traducció, el material es prepara en format XML per a poder ser incorporat a l'observatori Talaia.

Les propostes terminològiques que s'incorporen a Talaia tenen com a informació bàsica, a més de l'entrada, la referència de la revista a la qual pertany i, concretament, el número on ha estat publicada; l'àrea temàtica; el context o contextos d'ús de l'entrada, i l'autor que l'ha usada. Un exemple de l'entrada tipus que s'incorpora a Talaia és la que es mostra seguidament.

cibertextualitat f [*UOC Papers*, núm. 4] Àrea temàtica: art, ciència i tecnologia Context: "La *cibertextualitat* és un terme paraigua per a diferents tipus de textos digitals, com ara hipertextos, textos cinètics, textos generats, textos que utilitzen tecnologies agents, etc." (Raine Koskimaa, *UOC Papers,* núm. 4)	**cybertextuality** n [*UOC Papers*, núm. 4] Àrea temàtica: art, science and technology Context: "Cybertextuality is an umbrella term for different types of digital texts, such as hypertexts, kinetic texts, generated texts, texts employing agent technologies, etc." (Raine Koskimaa, *UOC Papers*, núm. 4)

En l'apartat següent veurem com s'ha anat millorant la tasca de filtratge dels candidats a terme a fi de poder disposar de terminologia neològica adequada per a ser incorporada a Talaia.

3. Evolució del mètode de treball

A partir dels primers resultats obtinguts amb el mètode que hem explicat en l'apartat anterior, ens hem adonat que moltes combinacions no tenen res a veure amb el domini temàtic del corpus d'especialitat amb què treballem. Així mateix, hem vist que hi ha combinacions sintàctiques que mai no podran arribar a ser un terme (combinacions verb-adverbi, per exemple). A més, hem observat que si tractem les diferents formes d'un terme (singular o plural) com a unitats independents l'una de l'altra no és possible tenir una dada objectiva de la rellevància d'aquest terme.

Per aquest motiu, el mètode de confecció de llistes de candidats a terme bilingües ha evolucionat. Els canvis i les millores s'han concentrat a aconseguir llistes de candidats a terme cada cop més acurades i significatives pel que fa a la rellevància dels termes, independentment de les seves formes. En canvi, no hi ha hagut variacions en el mètode d'obtenció dels equivalents de traducció. A més, l'ampliació de l'ús de filtres ha estat una prioritat, ja que els corpus a partir dels quals extraiem els candidats són molt grans i, per aquest motiu, les llistes de combinacions de paraules poden ser molt llargues.

3.1. Canvi de formes a lemes

L'evolució del mètode es basa en un canvi d'orientació de les primeres extraccions. Ara no ens centrem en les formes sinó en els lemes de les paraules. Aquest canvi metodològic es deu al fet que el treball amb formes presenta una sèrie d'inconvenients que queden resolts fent servir lemes. Un dels inconvenients que presenta l'ús de formes és d'eficiència a l'hora de triar els candidats a terme a partir de la freqüència d'aparició en el corpus. Per exemple, si el terme *cinemàtic* apareix en la llista de candidats a terme amb les formes *cinemàtic*, *cinemàtica* i *cinemàtics*, cada una té una posició diferent segons la freqüència d'aparició en el corpus. A l'hora de fer la revisió manual dels resultats s'han de tenir en compte totes les possibles

variants de cada candidat per a evitar incoherències, tasca que és feixuga i que introdueix un risc d'error. També és un inconvenient afegit que hi hagi formes que siguin homònimes d'altres que tenen una categoria gramatical diferent i que siguin terminològicament irrellevants. Per exemple, el substantiu *net*, manlleu que fa referència a la xarxa, és un terme rellevant, mentre que *net* com a adjectiu no ho és.

Un altre inconvenient té a veure amb la precisió de la llista de candidats, ja que, segons el llindar de freqüència que s'apliqui, es perden formes que corresponen a candidats a terme que són interessants de recollir. Per exemple, si establim un llindar de freqüència de quatre aparicions, i el candidat *imatge vectorial* apareix a la llista tres vegades i *imatges vectorials* una sola vegada, aquest candidat es perd, ja que cap de les dues formes supera el llindar de freqüència. En canvi, si treballem amb lemes i fem servir el mateix llindar de freqüència, aquest candidat no el perdrem, perquè apareix quatre vegades en el corpus.

3.1.1. Una nova orientació: unitats lema-categoria gramatical

Tenint en compte els inconvenients que hem assenyalat, ens hem plantejat canviar d'orientació fent servir uns triplets que contenen la informació següent: <forma d'una paraula, lema de la paraula, categoria gramatical de la paraula>. Gràcies a aquests triplets podem crear unitats <lema de la paraula-categoria gramatical de la paraula> (unitat LP-CP), i així el càlcul d'*n-grams* i la freqüència d'aparició en el corpus es fa sobre aquestes unitats.

3.1.2. Avantatges de la nova orientació

Gràcies al càlcul d'*n-grams* sobre unitats LP-CP, l'extracció parteix d'una llista més curta que no pas la que s'obté fent servir el mètode de treball inicial. La raó d'això es deu al fet que les diferents formes variants d'una paraula i d'una combinació de paraules no apareixen a la llista per separat, ja que comparteixen la mateixa combinació d'una unitat LP-CP o més d'una. Només surten els *n-grams* d'unitats LP-CP amb la dada de la seva freqüència, que és la suma de les freqüències de les formes variants. Això ens permet no sols treballar amb llistes més curtes, sinó recuperar candidats que hem perdut en la primera creació del glossari, sense haver de canviar el llindar de freqüència. Per exemple, tornant a l'exemple d'*imatge vectorial*, ara es comptabilitzen quatre aparicions de la unitat <<'imatge'-nom><'vectorial'-adjectiu>> i, per tant, *imatge vectorial* s'accepta perquè supera el llindar de freqüència.

Altres avantatges d'aquest enfocament és que tenim un alt grau de certesa que no sumem la freqüència de formes homògrafes que tenen una categoria gramatical i un sentit diferent del que té el terme. A més, tot i que l'assignació de la categoria gramatical es fa de manera automàtica, i que per tant s'ha de revisar, aquesta informació pot ser d'utilitat al terminòleg a l'hora de decidir si dóna per bo el candidat o no. Seria el cas, per exemple, d'*art net* i de *net art*: *net* és un adjectiu a *art net*, i probablement no seria escollit com a terme. En canvi, a *net art* és un substantiu i sí que s'acceptaria.

De tota manera, l'assignació automàtica de lema i categoria gramatical no està lliure d'error. Per aquest motiu, convé verificar que les formes d'una sola paraula no hagin estat considerades formes de dues paraules o més, amb lemes i categories gramaticals diferents. Per exemple, convé revisar que *net art* no tingui dues combinacions LP-CP diferents, una amb *net* com a adjectiu i l'altra amb *net* com a substantiu.

3.1.3. Recursos necessaris

Hem dit que les unitats LP-CP es creen a partir d'uns triplets <forma de la paraula, lema de la paraula, categoria gramatical de la paraula>. Necessitem, per tant, una llista de triplets que continguin la forma, el lema i la categoria gramatical de cada una de les paraules i símbols (signes de puntuació, etc.) del corpus. Els triplets han d'aparèixer en el mateix ordre en què apareixen les paraules i els símbols. Per exemple, si en el corpus tenim la seqüència "*la Bauhaus propicià un ensenyament [...]*" hem de tenir la seqüència de triplets següent:

<la, el, determinant>

<Bauhaus, bauhaus, nom propi>
<propicià, propiciar, verb>
<un, un, determinant>
<ensenyament, ensenyament, nom comú>
[...]

La tasca de crear manualment la llista de triplets a partir dels corpus de gran volum amb què treballem seria gairebé impossible, per això fem servir l'etiquetador de codi lliure i de lliure distribució FreeLing,[2] que permet crear aquesta llista de manera automàtica. Un cop hi posem un document d'entrada en format de text pla, el FreeLing crea un fitxer en què en cada línia hi ha la forma, el lema i la categoria gramatical de cada paraula. Com ja hem comentat, hi pot haver errors en l'assignació automàtica de lema i categoria gramatical. Amb tot, l'assignació incorrecta de lema és inferior al 5% i l'assignació incorrecta de categoria gramatical és inferior al 2%.

3.2. Disseny de nous filtres

A banda dels filtres descrits en l'apartat 2 pel que fa a la longitud de l'*n-gram*, el llindar de freqüència i l'aparició de paraules buides en les posicions extremes, la informació lingüística que conté cada *n-gram* ens permet ampliar els filtres aplicant-hi criteris més acurats. Presentem seguidament aquests filtres.

3.2.1. Filtre d'*n-grams* amb una CP que no pot ser terme

Fent servir aquest filtre podem filtrar *n-grams* que contenen la unitat LP-CP d'una conjunció, un pronom, un adverbi, un determinant demostratiu, etc. D'aquesta manera no caldrà revisar una llista de candidats que, des d'un punt de vista morfosintàctic, no poden ser termes. Per exemple, es desestimen candidats com ara *autor que crea* o *parlen molt fort*. És inevitable, però, que es perdi algun terme que pugui ser interessant, com ara *llenguatge no verbal*, pel fet que l'etiquetador ha etiquetat la partícula negativa 'no' com a adverbi.

3.2.2. Filtre d'*n-grams* LP-CP dependents d'un *n-gram* més llarg

Aquest filtre s'aplica als *n-grams* d'unitats LP-CP que estan continguts en un *n-gram* més llarg que té la mateixa freqüència. Això obeeix a l'assumpció que la freqüència idèntica és deguda al fet que l'*n-gram* més curt depèn de l'*n-gram* més llarg. Per aquest motiu, preferim presentar al revisor la forma completa. Amb l'aplicació d'aquest filtre, si *dinàmica de fluids* i *dinàmica* apareixen amb la mateixa freqüència, s'acceptaria *dinàmica de fluids* i es desestimaria *dinàmica*. Evidentment, la coincidència en les freqüències en alguns casos pot ser casual i és possible que l'*n-gram* contingut en l'*n-gram* més llarg sigui un terme independent d'aquest. De tota manera, la consulta dels contextos pot ajudar el terminòleg a esbrinar-ho.

3.2.3. Filtre d'*n-grams* LP-CP del vocabulari general

Després d'haver aplicat els filtres que tenen una base lingüística, en el procés de filtratge fem servir un filtre que té una base estadística. Aquest filtre es basa en el càlcul estadístic de la rellevància de cada *n-gram* LP-CP del corpus d'especialitat en contrast amb un corpus de llengua general de diferents temàtiques. Així es consideren candidats a terme els *n-grams* que són rellevants en el corpus d'especialitat però que no ho són en el de llengua general.

El càlcul es basa en un corpus de textos representatius de l'ús general de la llengua, els quals estan organitzats en dominis temàtics diferents. Sobre aquest corpus es pondera la freqüència d'aparició de cada *n-gram* i es contrasta amb la seva distribució en els diferents textos. Si un *n-gram* freqüent en el corpus d'especialitat no apareix en cap text o apareix en pocs textos de dominis temàtics diferents, llavors serà un *n-gram* rellevant del corpus d'especialitat. En canvi, si apareix en molts textos de dominis temàtics diferents, es considera que és un *n-gram* d'ús general i es desestima. Amb aquest criteri, en la prova pilot d'aquest filtratge, es van desestimar *arquitectura* i *investigació científica*, que tenen un ús força estès, i es van acceptar *artista multimèdia* i *hacktivisme*.

[2] www.lsi.upc.edu/~nlp/freeling/

El corpus representatiu de llengua general que fem servir actualment és un corpus del català, construït a partir de les edicions web del diari *Avui* de l'any 2003 al 2007. Els *n-grams* estan distribuïts segons la secció del diari en què apareixen (esports, espectacles, etc.). El llindar de rellevància segons el corpus d'especialitat és un valor numèric que es pot modificar i ajustar.

3.2.4. Filtre d'*n-grams* LP-CP amb noms propis

Finalment, també filtrem combinacions de paraules que contenen un nom propi, amb la qual cosa eliminem de la llista de candidats noms amb cognoms o sense, els quals són molt freqüents en els articles de les revistes. Per a fer la combinació de noms propis hem recopilat els noms de pila que són més habituals a Catalunya, la resta d'Espanya, els Estats Units, el Regne Unit i Alemanya segons diferents organismes oficials, com ara el Departament de Justícia de la Generalitat,[3] l'Institut Nacional d'Estadística[4] i les llistes de noms propis més populars de cada país que es poden trobar a la Viquipèdia.[5]

4. Resultats

Del conjunt de fonts que formaran part de l'observatori Talaia i que hem esmentat més amunt, hem començat a explotar els continguts d'*Artnodes*, revista que conté un nombre elevat de termes nous que són d'ús freqüent en l'àmbit de l'art i la tecnologia. Aquesta revista es començà a publicar l'any 2002 i el corpus amb què hem treballat consta de 71.584 paraules en català i de 69.436 paraules en anglès.

El nombre d'*n-grams* de longitud 3 sobre el qual treballem, sense aplicar cap filtre ni fixar cap llindar de freqüència (fins a freqüència 1), és de 72.893. En establir el llindar de freqüència de 3, que ja hem comentat, el nombre *d'n*-grams es redueix a 5.176. D'aquests 5.176 *n-grams*, 2.676 superen el filtratge de paraules buides i, finalment, un cop aplicats els filtres que hem explicat en l'apartat anterior, hem obtingut 672 candidats a terme. Aquests candidats han estat revisats per un terminòleg, el qual ha escollit els candidats que, a criteri seu, poden ser termes del domini temàtic de la societat de la informació. Per a facilitar la tasca de selecció, el terminòleg pot consultar un fitxer en format XML amb els contextos en què apareixen les formes de cada candidat. Dels 672 candidats, 176 (26%) han estat considerats termes específics del domini temàtic de la revista. En una selecció prèvia feta pel terminòleg, en què prescindí de si els candidats a terme eren específics del domini temàtic de la revista, dels 672 candidats en va seleccionar 362 (54%).

Un cop seleccionats manualment els termes, s'ha construït un corpus paral·lel català-anglès amb els sis números publicats de la revista, i així s'han obtingut de manera automàtica els possibles equivalents de traducció en anglès de les formes de cada un dels termes seleccionats.

5. Conclusions

En aquest article hem descrit el procés de treball que es duu a terme en el si del projecte Talaia i també com s'ha anat afinant el mètode de treball inicial amb l'objectiu de facilitar la detecció i validació final de la terminologia.

Una de les prioritats del projecte ha estat aconseguir que l'extracció automàtica de terminologia es dugui a terme fonamentalment mitjançant eines que permetin l'extracció multilingüe de candidats a terme, per a donar així al mètode una certa independència de la llengua de treball.

[3] http://www20.gencat.cat/portal/site/Justicia/
[4] http://www.ine.es/
[5] http://en.wikipedia.org/wiki/Category:Lists_of_popular_names

Els filtres que hem presentat, sobretot els que tenen en compte les categories gramaticals dels candidats i la seva rellevància segons el domini temàtic, ofereixen un enfocament flexible en l'extracció automàtica de termes. Una flexibilitat que contrasta amb altres mètodes d'extracció automàtica amb plantejaments més rígids, que solament consideren combinacions a partir de dues paraules i a partir de patrons morfosintàctics predefinits i dependents de la llengua. Així mateix, l'aplicació de filtres a partir de combinacions de categories gramaticals buides tampoc també es pot aplicar independentment de la llengua del corpus. A més, tenint en compte que les unitats terminològiques no tenen patrons fixos, permetem que surtin a la llum estructures morfosintàctiques variades, com ara adjectius i, fins i tot, combinacions verb-substantiu, que tenen una entitat terminològica.

Els filtres que apliquem també permeten seleccionar candidats monoparaula sense que representi allargar innecessàriament la llista de resultats. El filtre de pertinència al domini temàtic del corpus és prou potent perquè la presència de candidats monoparaula no introdueixin soroll, ja que es rebutgen directament

candidats que pertanyen a la llengua general i que no són interessants. Pels resultats obtinguts, podem dir que els candidats monoparaula tenen un pes rellevant des del punt de vista terminològic, com queda palès en paraules com ara *supercomputació* o *hipermèdia*, que han estat seleccionades candidates a terme. A més, com que l'etiquetador treballa sobre un formari representatiu de la llengua general, les unitats monoparaula que no són en aquest formari poden marcar-se com a possibles manlleus. Encara que aquest criteri pugui desestimar alguns termes monoparaula interessants perquè coincideixen amb paraules de la llengua general, com és el cas de *galetes* en el domini d'internet, el guany en temps i en rapidesa de producció de nous glossaris compensen a bastament aquests problemes puntuals de cobertura terminològica.

Finalment, cal dir que els filtres de pertinència al domini temàtic i la detecció de manlleus s'adeqüen molt bé a un dels propòsits del projecte, que és enriquir l'observatori Talaia amb unitats terminològiques de nova aparició i, per tant, d'ús encara no gaire estès.

Bibliografia

Daille B., Gaussier E., Lange J. M., (1994). "Towards automatic extraction of monolingual and bilingual terminology", *Proceedings of International Conference on Computational Linguistics* (COLING 1994), Kyoto, Japó.

Daille, B. (1996). "Study and implementation of combined techniques for automatic extraction of terminology", a Resnik, P.; Klavans, J. (ed.), *The Balancing Act : Combining Symbolic and Statistical Approaches to Language*, Cambridge: MIT Press, pàg. 49-66.

Bourigault, D. (1992). "Surface grammatical analysis for the extraction of terminological noun phrases", *Proceedings of the XIV Conference on Computational linguistics*, Nantes, França, pàg. 23-28.

Manning, C.D.; Schütze, H. (2003). *Foundations of statistical natural language processing.* MIT Press.

Oliver, A.; Vàzquez, M.; Moré, J. (2007). "Linguoc LexTerm: una herramienta gratuita de extracción automática de terminología", *Translation Journal*, vol. 11, núm. 4.
ISSN: 1536-7207.

Rius, Ll.; Sánchez, I.; Vàzquez, M. (2007). "Projecte Talaia: cap a un observatori de la societat del coneixement", a: *II Jornada de l'Acaterm. Nous reptes dels professionals en la comunicació especialitzada*.

Actes de la II Jornada de Terminologia i Serveis Lingüístics. Palma: Universitat de les Illes Balears. Servei Lingüístic, pàg. 69-77.
ISSN 978-84-8384-026-9

Vàzquez, M.; Oliver, A. (2007). "A Free Terminology Extraction Suite", a *Translating and the computer 29*, ASLIB Information Management, Londres.
ISBN 0 85142 485 6