Lluc Potrony i Joan Maria Romaní (cur.)
Indexació, terminologia i llenguatge jurídic
Barcelona: Institut d'Estudis Catalans, 2011, p. 149-162 DOI: 10.2436/15.2503.02.23

La terminologia jurídica catalana dins de WordNet 3.0

Merce Lorente
Jordi Vivaldi
Cristian Zanotti
Grup IULATERM, Institut Universitari de Lingüística Aplicada (UPF)
Barcelona

Resum

WordNet és una base de dades lèxica desenvolupada inicialment per a la llengua anglesa; amb el temps, s'ha convertit en un estàndard per a la representació de la informació lèxica arreu. Les seves múltiples aplicacions en el camp de l'enginyeria lingüística van fer que l'Institut Universitari de Lingüística Aplicada (IULA) de la Universitat Pompeu Fabra l'escollís per treballar amb l'extractor de terminologia YATE, desenvolupat en el mateix centre. Des de la versió 3.0, WordNet és d'accés lliure, fet que ha facilitat l'acord entre els grups que hi treballen a l'Estat espanyol.

L'IULA van començar a enriquir WordNet amb termes del dret —i també dels àmbits de la informàtica i del medi ambient— l'any 2007. Actualment, s'ocupa de l'ampliació terminològica de WordNet (hi introdueix termes en català i, si no han estat recollits abans, també en espanyol i en anglès) i de la migració d'informació de versions anteriors en què havien treballat diverses universitats catalanes. Aquest enriquiment amb termes jurídics es veu alentit pels problemes específics d'aquest llenguatge i per les limitacions de WordNet.

Paraules clau: bases de dades lèxiques, dret, terminologia, WordNet, YATE.

Abstract: *Catalan legal terminology in WordNet 3.0*

WordNet is a lexical data base initially developed for English but which, as time has gone by, has become a standard for the representation of lexical information everywhere. Its multiple applications in the field of linguistic engineering prompted the Institut Universitari de Lingüística Aplicada [University Institute of Applied Linguistics] (IULA) of the Pompeu Fabra University to select it to work with the YATE terminology extractor developed in the same centre. WordNet is free since version 3.0, which has facilitated the reaching of agreements among the groups that work with it in the Spanish State.

The IULA began to feed WordNet with legal terms — as well as IT and environmental terminology — in 2007. At the moment it is expanding WordNet's terminology (it adds terms in Catalan and, if they have not already been collected, in Spanish and English as well) and is in charge of migrating information from previous versions that different Catalan universities had worked on. The feeding-in of legal terms has been hampered somewhat by the specific problems of this type of language and by WordNet's own limitations.

KEY WORDS: lexical data bases, law, terminology, WordNet, YATE.

1. INTRODUCCIÓ

WordNet® és una base de dades lèxica d'ampli abast de la llengua anglesa, desenvolupada sota la direcció del professor George A. Miller a la Universitat de Princeton —a la figura 1 es pot veure la pàgina principal del web de WordNet. S'hi recullen noms, verbs, adjectius i adverbis agrupats en conjunts de sinònims que expressen significats, anomenats *synsets*. Els *synsets* estan interconnectats mitjançant relacions semàntiques, bàsicament de caràcter jeràrquic: hiperonímia i hiponímia. La xarxa de paraules relacionades significativament es pot explorar mitjançant un navegador intern. Des de la versió 3.0 (2006) és un recurs de lliure accés.

FIGURA 1. Web del projecte WordNet

Convé destacar que, com que s'utilitza molt en activitats de tipus diversos de processament del llenguatge natural i de recuperació d'informació, el recurs WordNet ha esdevingut pràcticament un estàndard de representació de la informació lèxica. Per les seves múltiples aplicacions, ha rebut diferents denominacions: base de dades, jerarquia lèxica, diccionari en línia, ontologia.[1] A més, són moltes les llengües per a les quals s'han desenvolupat o s'estan desenvolupant WordNets específics o multilingües (alemany, francès, italià, japonès, castellà, basc, català, gallec, entre d'altres).

Precisament perquè és un recurs molt utilitzat i adaptat a moltes llengües per a una gran diversitat d'aplicacions de l'enginyeria lingüística, va ser escollit per interactuar amb l'extractor automàtic de terminologia YATE (Yet Another Terminology Extractor).

L'extractor YATE (figura 2) va ser desenvolupat a l'Institut Universitari de Lingüística Aplicada (IULA) de la Universitat Pompeu Fabra l'any 2001 per Jordi Vivaldi, del grup IULATERM. Es tracta d'un extractor automàtic de terminologia híbrid, és a dir que utilitza estratègies estadístiques i estratègies lingüístiques combinades per a identificar les paraules d'un text especialitzat que tenen més probabilitats de ser unitats terminològiques. YATE té una estructura modular en la qual cada mòdul puntua el grau de terminologicitat d'un candidat a terme, com si fos el membre d'un jurat. Els mòduls de naturalesa lingüística són el morfològic (o de formants cultes), el morfosintàctic (o de patrons categorials i sintagmàtics) i el semàntic (la consulta de WordNet). Els mòduls lingüístics han d'adaptar-se per a cada llengua, mentre que els mòduls estadístics són d'ús comú. Les feines d'adaptació per a cada llengua i per a cada àmbit temàtic es concentren fonamentalment en l'enriquiment terminològic de WordNet.

L'objectiu d'aquesta comunicació és presentar la problemàtica diversa que suposa incorporar terminologia jurídica en un recurs ontològic com aquest, que sempre s'ha dissenyat partint de la llengua anglesa, que es preveu com un recurs multilingüe i que no està pensat inicialment per al treball terminològic. Veurem casos tipus i presentarem possibles solucions per al debat. Seria interessant d'obrir una línia de col·laboració entre juristes i terminòlegs per tal que la terminologia jurídica en llengua catalana estigués ben representada a WordNet.

1. Cal recordar, però, que no se la pot considerar una «ontologia» en sentit estricte, perquè és un recurs de representació d'unitats lèxiques (forma i significat) i no de conceptes.

FIGURA 2. Web de YATE

2. ELS ANTECEDENTS DE L'ENRIQUIMENT TERMINOLÒGIC DE WORDNET

El mòdul semàntic de l'extractor YATE utilitzava una còpia amb llicència de la jerarquia lèxica WordNet 1.5, concretament d'EuroWordNet, amb informació enriquida per a l'espanyol i el català, que són les llengües per a les quals s'ha adaptat fins ara aquest extractor. En la primera versió de YATE (2001) els mòduls lingüístics van ser adaptats per a l'extracció de terminologia de textos mèdics. L'any 2003, arran de la construcció del *Banc de Genoma Humà* (*http://genoma.iula.upf.edu:8080/genoma/index.jsp*), es va realitzar una segona adaptació per a la genòmica. Durant el període 2004-2007 el projecte RICOTERM2 va assumir l'adaptació per a l'economia i es va realitzar un protocol de treball per a futures adaptacions a altres àmbits.

La tasca d'enriquiment de WordNet per als tres àmbits ja complets (medicina, genòmica i economia) queda reflectida en la taula següent:

TAULA 1. *Enriquiment de WordNet per a l'adaptació de YATE*

Noves entrades	*Medicina*	*Genòmica*	*Economia*
Synsets	1.370	137	15
Variants	1.286	163	445
Relacions	526	11	16

Com es pot observar, el volum d'informació introduïda per a la medicina és molt més gran que per als altres camps. Per a explicar per què no ha calgut entrar tants *synsets* terminològics per a la genòmica o per a l'economia hem de remetre a causes diferents: *a*) la genòmica comparteix molts dels recursos lèxics i semàntics de la medicina, i *b*) l'economia ho fa amb el llenguatge comú. La dificultat de la tasca en economia no rau, doncs, en la inclusió de noves dades, sinó en el disseny d'estratègies complementàries que permetin millorar els resultats de l'extracció, procediment extensible a tots els àmbits especialitzats propers a la llengua comuna (ciències socials, professions, arts).

Després de l'enriquiment del mòdul semàntic de WordNet, les avaluacions de l'extractor YATE han estat molt positives: un 95 % de precisió en medicina i genòmica i un 75 % en economia, sempre per a una cobertura del 30 %.

Del 2007 al 2010, en el marc del projecte RICOTERM3, s'han iniciat les adaptacions de domini per al dret, el medi ambient i la informàtica. Amb aquestes adaptacions, l'extractor YATE cobrirà els mateixos àmbits temàtics que té el Corpus Tècnic del IULA (*http://bwananet.iula.upf.edu*). Com que la versió 3.0 del WordNet anglès, de lliure accés, es va publicar l'any 2006, es va optar per deixar d'enriquir la versió 1.5 per a l'extractor YATE i participar en un nou projecte interuniversitari de desenvolupament del WorNet 3.0 per a les llengües oficials de l'Estat espanyol.

2.1. *L'ampliació terminològica de WordNet*

El treball d'enriquiment de WordNet amb informació especialitzada per a l'extractor YATE consisteix, bàsicament, en la detecció de nodes (*synsets*) que puguin funcionar com a fronteres de domini, o sigui, que permetin assegurar que sota aquesta frontera totes les unitats relacionades seran terminològiques (lèxic especialitzat). Una altra tasca, més àrdua, se'ns presenta quan no existeixen aquestes fronteres i cal declarar una quantitat d'entrades noves mínima i suficient. En qualsevol cas, l'enriquiment de WordNet sempre es fa amb informació lèxica en anglès, en espanyol i en català.

Per tant ens trobem davant de dos escenaris possibles:

a) El concepte que volem entrar ja és a la versió anglesa i hi introduïm les variants catalanes i les castellanes, amb les glosses corresponents (o quasidefinicions) i els exemples d'ús.

b) El concepte que volem entrar no és a la versió anglesa i hi introduïm totes les dades noves corresponents a l'anglès, al català i al castellà, i les pengem de la jerarquia.

2.2. *La migració de la informació terminològica a WordNet 3.0*

Com apuntàvem abans, WordNet s'ha anat desenvolupant en centres de recerca diferents, en projectes diferents, per a objectius diferents i en èpoques diferents. El resultat ha estat que hi ha molts WordNets escampats pel món amb informacions diferents i en versions diferents (1.5, 1.6, 2.0). Pel que fa a la llengua catalana, hi havien treballat investigadors de les universitats Politècnica de Catalunya, de Barcelona, Oberta de Catalunya i Pompeu Fabra. Les dades es trobaven esparses, eren parcials i potser estaven indesitjablement repetides. Era un neguit antic de molts de nosaltres aconseguir posar-nos d'acord i compartir aquest recurs, fet entre tots i amb diners públics.

L'aparició l'any 2006 de la versió 3.0 de l'anglès de lliure accés ha fet més fàcil l'acord entre els grups que encara hi treballàvem a l'Estat espanyol per al basc, el gallec, el castellà i el català. Així, el Multilingual Central Repository 2.0 (MCR 2.0) és un projecte interuniversitari, coordinat per la Universitat del País Basc, en el qual participen, a més d'aquesta universitat, la Universitat de Vigo, la Universitat de Barcelona i la Universitat Pompeu Fabra. L'objectiu principal d'aquest projecte és establir una infraestructura cientificotecnològica bàsica que faciliti la coordinació, integració i transferència de tecnologia i recursos entre els diferents grups d'investigació que actualment desenvolupen bases de coneixement semàntic d'àmplia cobertura (WordNets).

S'ha desenvolupat una nova versió del Multilingual Central Repository (MCR)[2] —vegeu la figura 3— que integrarà el coneixement de nous WordNets per al castellà, el basc, el català i el gallec enllaçats a l'última versió del WordNet anglès. A més, també s'estan desenvolupant i actualitzant les eines i aplicacions necessàries per a la gestió i el manteniment del MCR, de manera que els diferents recursos que l'integren es desenvolupin distribuïdament però sincronitzada. Tots els recursos i les aplicacions que s'obtinguin com a resultat d'aquesta acció seran lliures i accessibles per a qualsevol grup d'investigació.

2. Plataforma que ja existia per a l'edició i la consulta d'EuroWordNet en versions anteriors, desenvolupada per la Universitat Politècnica de Catalunya.

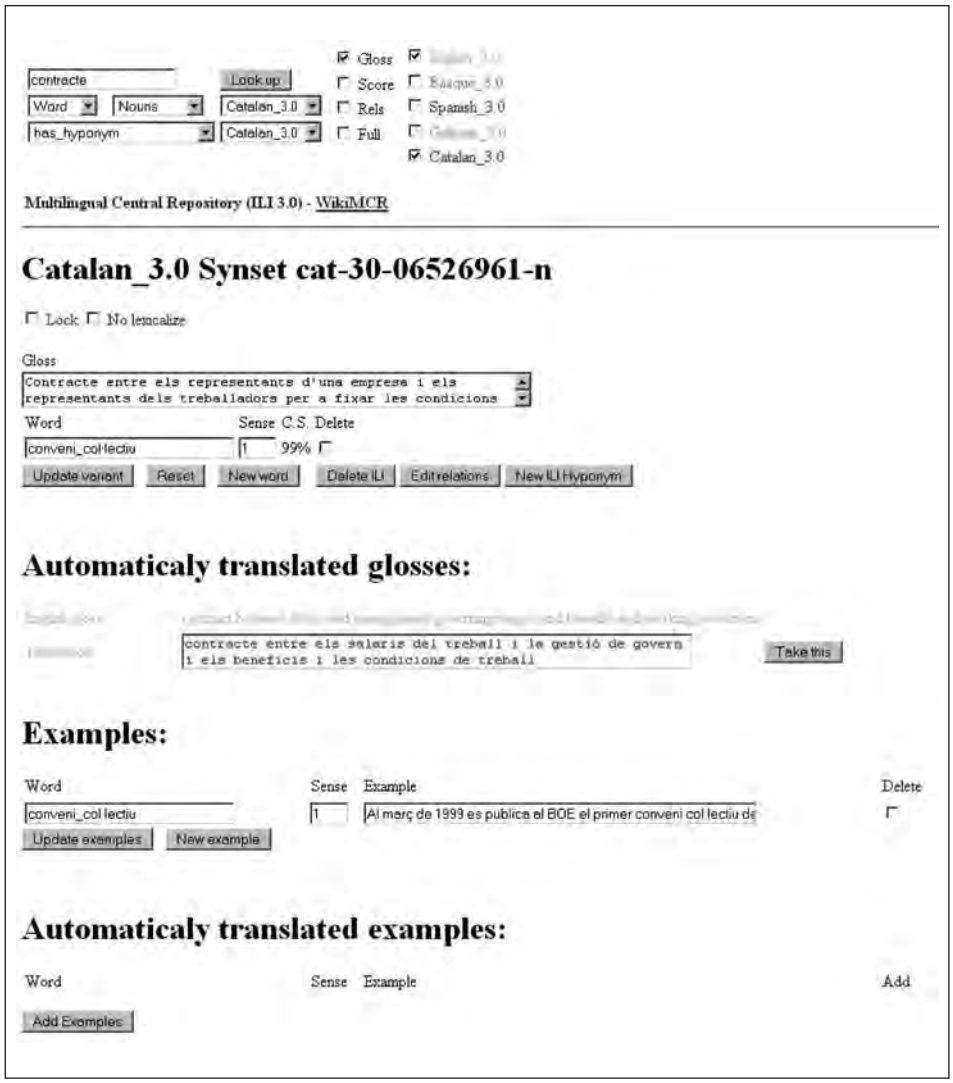

FIGURA 3. Interfície de la plataforma d'edició MCR 2.0

Concretament, el nostre grup d'investigació participa en el projecte amb l'actualització de dades per a la llengua catalana, amb la migració de les dades terminològiques que havíem desenvolupat a WordNet 1.5 i amb noves incorporacions terminològiques per a adaptar l'extractor YATE als dominis de la informàtica, el dret i el medi ambient.

L'objectiu del projecte d'enriquiment del nou WordNet 3.0 per mitjà de la plataforma MCR 2.0 és aconseguir que l'avaluació dels resultats de l'extractor YATE en els àmbits de la informàtica, el dret i el medi ambient s'acosti als resultats obtinguts en economia (un 75 % de precisió per a un 30 % de cobertura).

3. La problemàtica específica de la terminologia jurídica a WordNet

Cal destacar que l'estructura de WordNet 3.0 anglès ha estat modificada respecte de les versions anteriors, cosa que ha provocat problemes en la migració de dades però també ha mostrat com evoluciona la visió que es té del lèxic especialitzat.

Si bé, en les versions anteriors de WordNet, la presència d'unitats terminològiques era molt minoritària, en la versió actual àmbits com la informàtica i el medi ambient es troben més representats per la seva relació amb les tendències culturals i de consum. En canvi, en altres àmbits com el dret, la nova versió es manté allunyada de la terminologia amb molt poques entrades representatives, ni tan sols per a conceptes jurídics bàsics i d'ús comú en contextos distints, com la distinció entre *dret nacional* i *dret internacional* o entre *persona física* i *persona jurídica*.

Actualment les novetats terminològiques introduïdes que pertanyen a l'àmbit jurídic són les que es recullen en la taula 2.

Taula 2. *Novetats de dret a WordNet*

Novetats 2011 MCR 2.0	*Dret*
Synsets adaptats al SP	137
Variants relacionades	378
Synsets adaptats al CAT	216
Variants relacionades	362
Nous synsets SP-CAT-EN	52

Hem vist més amunt que se'ns mostren dos escenaris possibles:

— Escenari 1. El significat que volem entrar ja és en la versió anglesa: hi introduïm, doncs, les variants del català i del castellà, amb glosses i exemples d'ús.

— Escenari 2. El significat que volem entrar no és en cap llengua: aleshores hi introduïm totes les dades noves corresponents a l'anglès, al català i al castellà.

Podem il·lustrar l'escenari 1 amb exemples de l'àmbit jurídic, com els relacionats amb principis jurídics derivats del dret anglosaxó, com el següent:

EN *due process of law*: (law) the administration of justice according to established rules and principles; based on the principle that a person cannot be deprived of life or liberty or property without appropriate legal procedures and safeguards.

CAT *tutela judicial, dret a la tutela judicial*: (dret) l'administració de la justícia d'acord amb les normes i principis establerts, basats en el principi que una persona no pot ser privada de la vida, o la llibertat, o la propietat, sense tenir garanties ni procediments jurídics.

Per exemplificar l'escenari 2 amb dades reals del discurs jurídic, podem esmentar l'entrada *ex novo* del terme *persona jurídica* (figura 4):

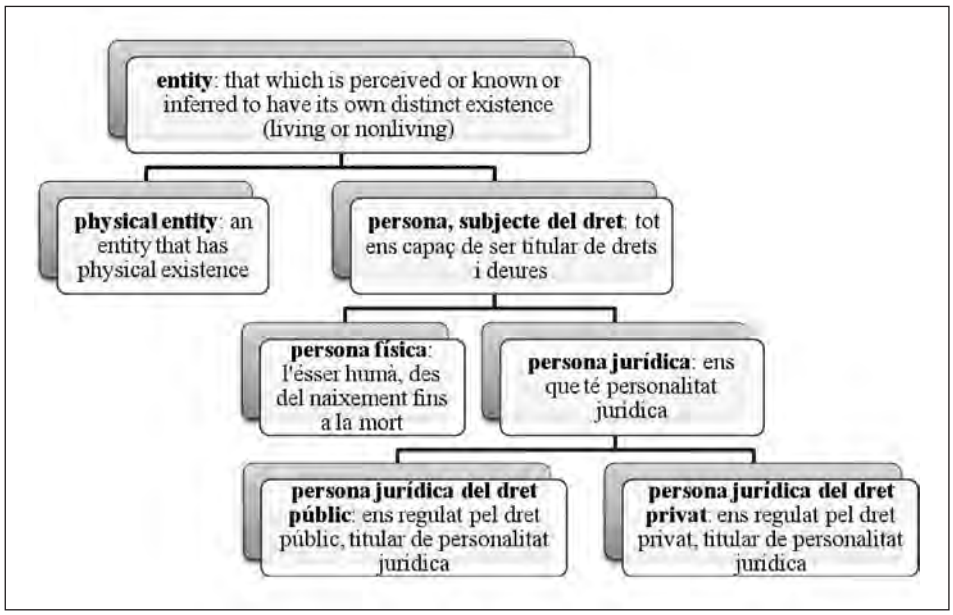

FIGURA 4. Ubicació de *persona jurídica* en la jerarquia general

A banda de les dues situacions més habituals en l'enriquiment de terminologia a WordNet, la tasca aplicada a la terminologia jurídica comporta una sèrie de dificultats afegides, que podem classificar en dos grans blocs d'obstacles en la representació: problemes semàntics i limitacions de WordNet.

3.1. *Problemes semàntics*

Els problemes semàntics que detectem en la tasca de representar unitats terminològiques del dret en català, castellà i anglès es poden sintetitzar en la noció de

anisomorfisme. No sempre hi ha equivalència conceptual entre dues llengües amb matisos diferents.

Un dels casos més habituals —relacionats sobretot amb la terminologia referida a professions, competències, càrrecs o organismes, i, per tant, molt vinculat a l'administració de qualsevol branca del dret— és la delimitació conceptual no coincident. Ho podem il·lustrar amb els termes anglesos *barrister, solicitor* i *attorney* en front del terme català *advocat* (figura 5).

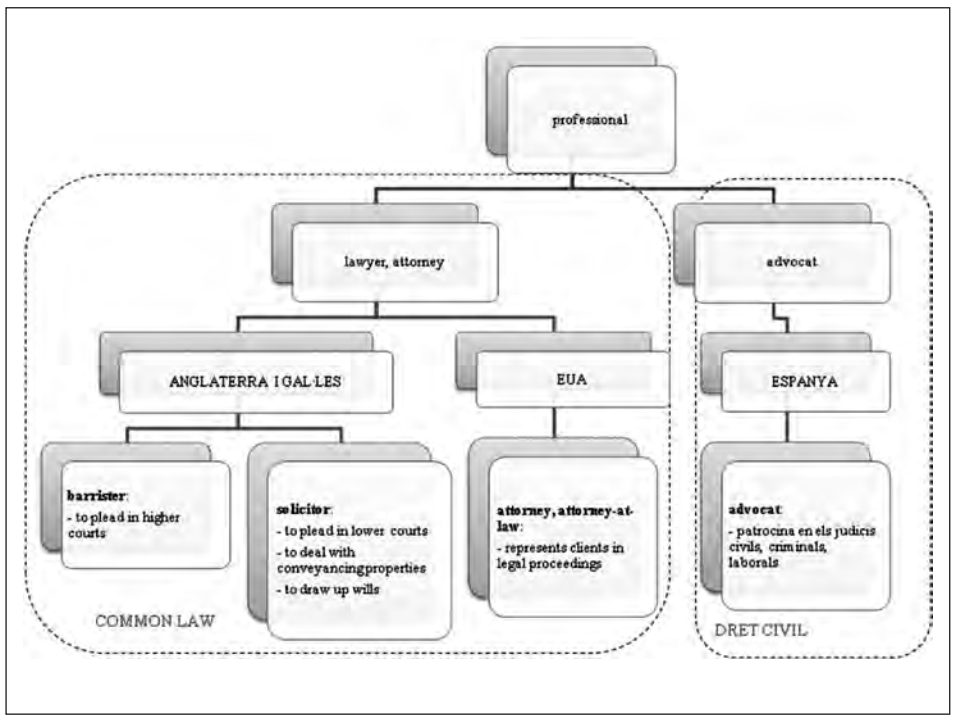

FIGURA 5. Exemple de delimitació conceptual no coincident

El segon dels casos de no-coincidència és més complex. Es tracta d'un problema d'ubicació diversa en l'estructura de representació semàntica. Dos termes que poden ser considerats equivalents des del punt de vista de la traducció, com per exemple *ticket* i *multa* (figura 6), han de ser representats semànticament en espais diferents de la jerarquia lèxica per tal com responen a procediments jurídics absolutament diferents i, lògicament, han de ser descrits mitjançant glosses o definicions ben diferenciades.

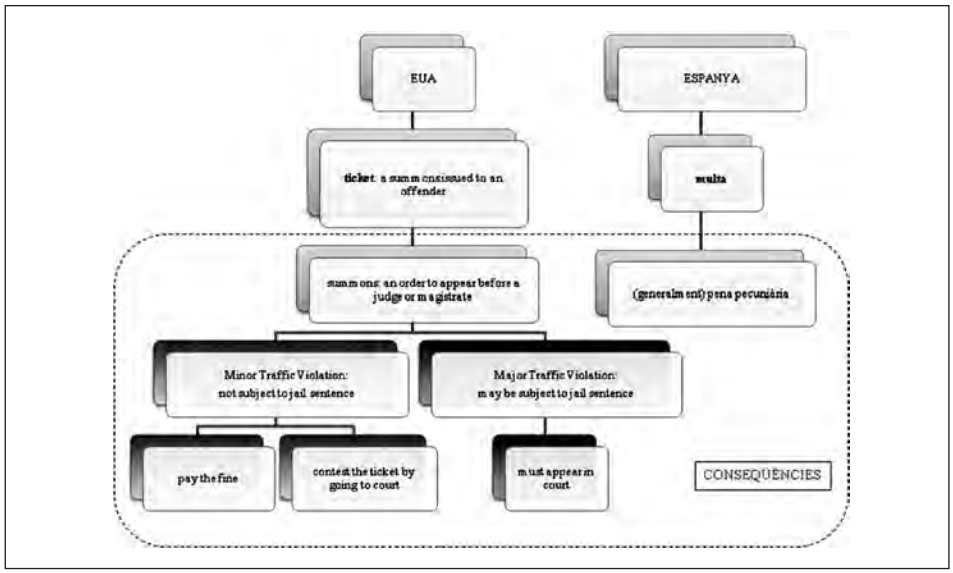

FIGURA 6. Exemple d'ubicació en l'estructura no coincident

El tercer cas, segurament el més descrit en la bibliografia sobre traducció jurídica, és la manca d'equivalència, per efecte de sistemes conceptuals divergents. Així, trobem casos com els termes *capital offence* o *statutory offence*, propis de la *common law*, que no disposen d'equivalents propis en el discurs jurídic català (figura 7).

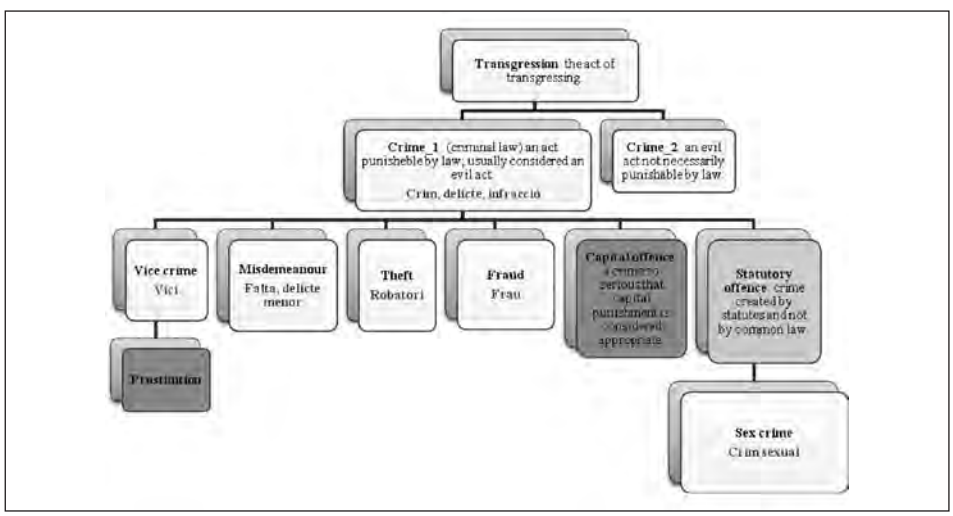

FIGURA 7. Manca d'equivalència conceptual

Una altra de les dificultats de representació semàntica de termes jurídics en diverses llengües és la que es presenta quan per a cada llengua disposem d'una estructura distinta de cohipònims que penja del mateix concepte o node. Fixem-nos, per exemple, en algunes de les branques en què s'organitza la jurisprudència. Mentre que en anglès, al costat de *contract law*, trobem el cohipònim *patent law*, en català (i en castellà) reconeixem el *dret contractual*, però no el **dret de patents* com a branca diferenciada de la jurisprudència.

> *Jurisprudence_1: the branch of philosophy concerned with the law and the principles that lead courts to make the decisions they do*
>
> Jurisprudència: la branca de la filosofia que s'ocupa del dret i dels principis que porten els tribunals a prendre les seves decisions

> *Contract law: that branch of jurisprudence that studies the rights and obligations of parties entering into contracts*
>
> Dret contractual: la branca de la jurisprudència que estudia els drets i les obligacions de les parts d'un contracte

> *Patent law: that branch of jurisprudence that studies the laws governing patents*
>
> *Dret de patents: la branca de la jurisprudència (?) que estudia les lleis que governen les patents

Un darrer exemple d'aquest últim cas de no-coincidència ens l'aporta l'intent de representació de termes com *senate, congress, house, chamber, legislative assembly* dins de l'anglès mateix, ja que hi ha divergències importants entre el sistema jurídic britànic i el nord-americà. Evidentment, aquest problema s'agreuja quan volem representar els termes equivalents d'altres llengües, com ara el català, que presenten relacions d'hiperonímia i cohiponímia diferents.

3.2. *Limitacions de WordNet*

De fet, els problemes de representació semàntica que hem exposat quan encarem el contrast de terminologia jurídica en diverses llengües no són problemes lingüístics. Per a la lingüística, i per a una terminologia d'orientació comunicativa, aquests exemples reflecteixen la realitat. Cada llengua configura de manera autònoma conceptes, categories, relacions, denominacions, variació, tipus de discurs. I la llengua reflecteix també de manera natural els diversos sistemes d'organització social de les comunitats.

La no-coincidència, doncs, és natural. La dificultat rau en l'assaig de representar el coneixement lèxic (formal, semàntic i pragmàtic) en una única estructura per a totes les llengües. Per als enginyers seria més pràctic que les estructures in-

terlingüístiques fossin isomòrfiques, però no ho són. És feina, doncs, de la cooperació entre lingüistes i enginyers resoldre l'anisomorfisme per tal que recursos com WordNet continuïn sent útils en traducció automàtica, en recuperació d'informació i en gestió del coneixement, sense trair la variació lingüística.

Les noves versions de WordNet ja comencen a aplicar algunes d'aquestes millores i algunes plataformes de consulta i edició multilingüe com la MCR 2.0 també ens hi ajuden.

4. A MODE DE CONCLUSIÓ

Acabem de referir-nos a la cooperació en marxa entre lingüistes i enginyers, però no voldríem cloure aquesta presentació sense apel·lar a la cooperació necessària entre lingüistes i experts en dret per a mirar de resoldre els casos exposats o d'altres que van sorgint quan s'enriqueix de WordNet 3.0 amb terminologia jurídica. De vegades, hem pogut copsar que aquesta mena de recursos generalistes, d'ampli abast, no tenen prou cura de la precisió amb què cal tractar les unitats terminològiques si volem usar-los per a tasques de mediació en discurs especialitzat. No en tenim prou d'entrar denominacions localitzades en textos jurídics; cal entrar unitats terminològiques del dret rellevants i fer-ho amb rigor, amb glosses i exemples adequats, indicant sempre que calgui la variació entre llengües i representant les delimitacions semàntiques i les relacions lèxiques de manera precisa. Només així podrem disposar de recursos lèxics prou potents per a treballar eficaçment en les aplicacions del discurs especialitzat.

5. BIBLIOGRAFIA

CABRÉ, M. Teresa; ESTOPÀ, Rosa; VIVALDI, Jordi (2001). «Automatic term detection: A review of current systems». A: BOURIGAULT, Didier; JACQUEMIN, Christian; L'HOMME, Marie-Claude (ed.). *Recent Advances in Computational Terminology.* Amsterdam: John Benjamins, p. 53-87.

JOAN, Anna; VIVALDI, Jordi; LORENTE, Mercè (2008). «Turning a term extractor into a new domain: first experiences». A: *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)* [en línia]. Marràqueix: ELRA. <http://www.lrec-conf.org/proceedings/lrec2008> [Consulta: 24 novembre 2011].

LORENTE, Mercè (2006). «Expansió de consultes multilingüe per a la recuperació d'informació en economia». A: JUAN, Maria; AMENGUAL, Marian; SALAZAR. Joana (ed.). *Lingüística aplicada en la sociedad de la información y la comunicación*. Palma de Mallorca: Universitat de les Illes Balears.

— (2009). «Algunas experiencias de la integración de ontologías en proyectos de terminología». *Puntoycoma* [Brussel·les; Luxemburg: Dirección General de Traducción de la Comisión Europea], núm. 115-S (novembre-desembre), p. 34-37.

VIVALDI, Jordi (2001). *Extracción de candidatos a término mediante combinación de estrategias heterogéneas* [tesi de doctorat]. Barcelona: Universitat Politècnica de Catalunya. [També a «Sèrie Tesis», núm. 9, Institut Universitari de Lingüística Aplicada de la Universitat Pompeu Fabra]

— (2006). *Sistema de extracción de candidatos a término YATE*: *Manual de utilización*. Barcelona: Institut Universitari de Lingüística Aplicada. (Papers de l'IULA. Sèrie Informes; 43)

— (2009). «Extracción de información conceptual basada en ontologías». A: *Terminología y sociedad del conocimiento.* Berna: Peter Lang, p. 375-404.