REVISTA DE LLENGUA I DRET // JOURNAL OF LANGUAGE AND LAW #69

LA TERMINOLOGIA JURÍDICA DEL IATE EN CATALÀ

Mercè Vàzquez*

Antoni Oliver**

Georgina Ubide***

Resum

Les institucions europees elaboren una gran diversitat de lleis i documentació que pertanyen a diferents àmbits de coneixement i que són traduïdes a totes les llengües oficials de la Unió Europea. Aquest conjunt de lleis i disposicions constitueix el cabal comunitari (*acquis communautaire*) i conté el conjunt del dret comunitari vigent. La terminologia que és present en el cabal comunitari es troba recollida en la base de dades terminològica multilingüe IATE (InterActive Terminology for Europe). Actualment les lleis i la documentació que constitueixen el cabal comunitari no estan disponibles en llengua catalana i, en conseqüència, la terminologia corresponent tampoc no forma part de la base de dades terminològica IATE. Per a disposar del recull més ampli possible de terminologia jurídica present en el IATE en llengua catalana aprofitant repertoris terminològics existents, en aquest article presentem un mètode de treball que permet seleccionar la terminologia jurídica present en el IATE i localitzar de manera automàtica els corresponents equivalents de traducció en català. Els resultats obtinguts confirmen que l'aprofitament de recursos terminològics ja disponibles permet completar la base de dades IATE amb les corresponents denominacions en llengua catalana de manera ràpida i eficaç. Amb aquest mètode de treball s'obre la possibilitat de crear nous repertoris terminològics en llengua catalana que serveixin per a completar els diferents àmbits temàtics del IATE i que puguin estar a disposició d'especialistes, traductors, correctors i usuaris de la llengua en general.

Paraules clau: Terminologia; processament del llenguatge natural; traducció; corpus lingüístics; lingüística.

IATE'S LEGAL TERMINOLOGY IN CATALAN

Abstract

Europe's institutions draw up a wide variety of laws and documentation belonging to different fields of knowledge, which are translated into all the official languages of the European Union. This set of laws and provisions constitutes the body of European law (the acquis communautaire) and encompasses all applicable community legislation. The terminology contained in this body of European law is gathered together in the IATE (InterActive Terminology for Europe) multilingual terminology database. As of today, the legislation and other documentation constituting this body of law are not available in Catalan, and so neither does the associated terminology form part of the IATE database. To possess the broadest possible collection of legal terminology present in IATE in Catalan, taking advantage of existing terminological directories, in this article, we present a working method that permits selection of the legal terminology present in IATE and the automatic location of the associated equivalents translated into Catalan. The results obtained confirm that taking advantage of pre-existing terminological resources permits quick and efficient supplementing of the IATE database with the associated terms in Catalan. This approach opens up the possibility of creating new terminological directories in Catalan to complement IATE's thematic areas and that may be made available to specialists, translators, correctors and language users in general.

Keywords: Terminology; natural language processing; translation; text corpora; linguistics.

* Mercè Vàzquez, professora dels Estudis d'Arts i Humanitats de la Universitat Oberta de Catalunya. Les seves principals línies de recerca són l'extracció automàtica de terminologia, la recuperació d'informació i la gestió d'informació especialitzada.

** Antoni Oliver, professor dels Estudis d'Arts i Humanitats de la Universitat Oberta de Catalunya i director del màster de Traducció Especialitzada de la mateixa universitat. La seva àrea de recerca és el processament del llenguatge natural i especialment la traducció automàtica i la generació de recursos lingüístics.

*** Georgina Ubide, graduada en Traducció i Interpretació per la Universitat de Vic - Universitat Central de Catalunya i traductora jurada espanyol-català. La seva àrea d'especialitat és la traducció jurídica.

Article rebut el 27.07.2017. Avaluació cega: 11.08.2017. Data d'acceptació de la versió final: 19.12.2017.

Citació recomanada: Vàzquez, Mercè; Oliver, Antoni; Ubide, Georgina. «La terminologia jurídica del IATE en català». *Revista de Llengua i Dret, Journal of Language and Law*, núm. 69, (juny 2018), p. 139-153. DOI: 10.2436/rld.i69.2018.3010.

Sumari

1 Introducció

2 Materials i mètodes

2.1 La base de dades terminològica IATE

2.2 La Terminologia Oberta del TERMCAT

2.3 El corpus del DOGC

2.4. Part experimental

2.4.1 Cerca d'equivalents de traducció a Terminologia Oberta

2.4.2 Cerca d'equivalents de traducció al corpus del DOGC

3 Resultats i discussió

4 Conclusions i treball futur

5 Referències bibliogràfiques

1 Introducció

Les institucions i els organismes oficials de la Unió Europea s'encarreguen de redactar un nombre important de lleis i un gran volum de documentació que pertanyen a diferents àrees de coneixement i que són traduïdes a totes les llengües oficials. Aquesta legislació constitueix el que s'anomena *cabal comunitari* (*acquis communitaire*), que és el conjunt del dret comunitari vigent que accepta un estat en entrar a formar part de la Unió Europea. Per tal de compilar la terminologia que és present en el cabal comunitari s'ha creat una base de dades terminològica multilingüe i interinstitucional de la Unió Europa, el IATE[1] (InterActive Terminology for Europe) (Johnson, Macphail, 2000), que compta amb les aportacions permanents que fan els gairebé 5.000 traductors de les institucions europees. Una base de dades terminològica és una recopilació sistematitzada i generalment presentada en un format informàtic (Oliver, 2016). Les bases de dades terminològiques recullen termes, és a dir, representacions superficials d'un concepte d'un domini especialitzat (Pazienza, 2005). Quan parlem de representacions superficials ens referim a la denominació, és a dir, a la paraula o conjunt de paraules que s'utilitzen per referir-se a aquest concepte. El IATE està disponible en les 24 llengües oficials de la Unió Europea i disposa de 552 combinacions lingüístiques. Les incorporacions terminològiques que fan els traductors al IATE són coordinades des de la Unitat de Coordinació Terminològica (TermCoord) del Parlament Europeu, que és una unitat que va ser creada pel Parlament Europeu l'any 2008 per a coordinar, harmonitzar i donar suport a la recerca terminològica i a l'emmagatzematge de noves dades al IATE, i també per a cooperar amb la resta d'unitats de treball i institucions en la titànica tasca de depurar i actualitzar una base de dades que conté uns 8 milions de termes. La terminologia que conté el IATE és disponible per a fer-hi consultes específiques; ara bé, per a poder dur a terme estudis de recerca relacionats amb el contingut d'aquesta base de dades, des de l'any 2014 se'n pot descarregar informació per llengües i àrees temàtiques. Actualment, la versió que es pot consultar del IATE consta solament de les llengües oficials de la Unió, i també el llatí. En la taula 1 es mostra el nombre de termes que hi ha disponible per cada llengua de la Unió Europea. Es pot observar que el nombre de termes que tenen algunes llengües com ara el croat, el txec o el letó és força baix si el comparem amb el nombre de termes que tenen llengües amb molts parlants, com ara l'anglès o el francès. Tenint en compte aquestes dades, la representació de termes en llengua catalana en el IATE també seria superior a aquestes tres llengües.

Llengua	Nombre de termes	Llengua	Nombre de termes
Búlgar	33.462	Italià	654.819
Txec	30.109	Lituà	40.456
Danès	568.069	Letó	32.273
Alemany	980.033	Maltès	43,817
Grec	498.728	Holandès	642.833
Anglès	1.282.712	Polonès	59.024
Espanyol	577.955	Portuguès	480.058
Estonià	38.783	Romanès	38.817
Finès	307.485	Eslovac	38.416
Francès	1.240.606	Eslovè	44.432
Irlandès	60.320	Suec	291.302
Croat	11.030	Llatí	61.316
Hongarès	34.290	Multilingüe	5.039

Taula 1. Nombre de termes per llengua al IATE (2017). Font: IATE

En aquest context, la llengua catalana no és una llengua oficial de la Unió Europea, malgrat les peticions que s'han fet perquè pugui tenir l'estatus de llengua oficial. La llengua catalana és parlada per uns deu milions de persones, moltes de les quals viuen a Espanya però també n'hi ha a Andorra, França i Itàlia. Si comparem la llengua catalana amb les llengües oficials de la Unió Europea, observem que és la novena llengua de la Unió pel que fa al nombre d'habitants i que té més parlants que quinze de les vint-i-quatre llengües oficials de la Unió Europea (grec, portuguès, txec, hongarès, suec, búlgar, irlandès, danès, eslovè, finès, lituà, letó, eslovè, estonià i maltès).

1 IATE, the EU's multilingual term base (http://iate.europa.eu/UserSettings.do?method=remove).

Els estudis que es duen a terme per a compilar terminologia multilingüe en diferents àmbits de coneixement, se centren en les llengües que tenen presència en la Unió Europea (Gaizauskas i altres, 2015). Com que la llengua catalana és reconeguda solament com a llengua de comunicació, però no com a llengua de treball dins la Unió, els documents oficials que configuren el cabal comunitari no són traduïts en aquesta llengua i, en conseqüència, la terminologia corresponent no és recollida al IATE. No obstant això, la Unitat de Coordinació de Terminologia del Parlament Europeu TERMCOORD i el Centre de Terminologia TERMCAT han signat recentment un acord de col·laboració amb l'objectiu d'ampliar, i enriquir així, el contingut del IATE amb terminologia en llengua catalana. Mitjançant aquest acord de col·laboració, el TERMCAT proporciona les designacions catalanes dels conceptes ja presents al IATE, que es van incorporant en la versió interna de la base de dades.

Així, doncs, per tal de dotar la llengua catalana d'aquests recursos lingüístics, hem dissenyat un mètode de treball que es basa en l'aprofitament de recursos lingüístics ja existents i que permeti compilar terminologia present al IATE en llengua catalana. Per a fer-ho, en aquest article s'identifiquen de manera automàtica els termes de l'àmbit del dret que són presents en el IATE en anglès i espanyol, ja que són les llengües que disposen d'un major nombre de termes; així mateix, se seleccionen automàticament els termes en català d'aquest àmbit que hi ha disponibles tant en l'espai Terminologia Oberta (TO) del TERMCAT[2] com en el *Diari Oficial de la Generalitat de Catalunya*[3] (DOGC), i, finalment, s'identifiquen manualment els termes jurídics en llengua catalana que són equivalents als termes en anglès i espanyol que hi ha al IATE.

L'article està estructurat en les parts següents: en el segon apartat descrivim els recursos lingüístics que han servit de base per a la part experimental de la nostra recerca, com són la base de dades terminològica IATE, els repertoris terminològics de l'àmbit del dret presents a Terminologia Oberta del TERMCAT i el corpus del *Diari Oficial de la Generalitat de Catalunya* (DOGC), i també presentem la part experimental de la nostra recerca en la qual descrivim de quina manera hem compilat la terminologia jurídica en llengua catalana; en el tercer apartat, analitzem els resultats obtinguts i la rendibilitat i aplicació de la recerca que hem dut a terme, i en el quart i darrer apartat presentem les conclusions i les línies de treball futur.

2 Materials i mètodes

Per tal de compilar terminologia jurídica en llengua catalana que és present en la base de dades IATE, hem dut a terme la tasca de selecció d'un conjunt de recursos lingüístics que actualment ja estan disponibles i hem implementat de manera experimental el procés de compilació d'aquesta terminologia. Així, primer hem recopilat de manera automàtica el conjunt de termes corresponent a l'àmbit jurídic que hi ha en el IATE. A continuació, hem fet un buidatge automàtic dels repertoris terminològics procedents de la Terminologia Oberta del TERMCAT. Seguidament, hem constituït el corpus del *Diari Oficial de la Generalitat de Catalunya* entre els anys 1977 i 2015 en castellà i català. I, finalment, hem implementat el procés d'identificació dels equivalents de traducció en català de la terminologia jurídica del IATE. En aquesta darrera fase de treball un equip d'experts lingüistes han revisat manualment els termes en anglès procedents del IATE i els corresponents equivalents de traducció extrets automàticament de la Terminologia Oberta i del DOGC per a comprovar tant si els termes són propis de l'àmbit jurídic com si les correspondències en català i castellà són pertinents.

2.1 La base de dades terminològica IATE

El IATE[4] és la base de dades terminològica de la Unió Europea i està disponible en les 24 llengües oficials de la Unió Europea, encara que no totes les llengües hi estan representades de la mateixa manera. Els termes que formen part del IATE estan organitzats en una sèrie d'àrees i subàrees temàtiques. En la taula 2 podem observar el primer nivell d'aquesta classificació.

2 TERMCAT, Centre de Terminologia (<http://www.termcat.cat/>).

3 *Diari Oficial de la Generalitat de Catalunya* (<http://dogc.gencat.cat/ca>).

4 http://iate.europa.eu/

Codi	Àrea temàtica
0	NO SUBJECT DOMAIN
4	POLITICS
8	INTERNATIONAL RELATIONS
10	EUROPEAN COMMUNITIES
12	LAW
16	ECONOMICS
20	TRADE
24	FINANCE
28	SOCIAL QUESTIONS
32	EDUCATION AND COMMUNICATIONS
36	SCIENCE
40	BUSINESS AND COMPETITION
44	EMPLOYMENT AND WORKING CONDITIONS
48	TRANSPORT
52	ENVIRONMENT
56	AGRICULTURE, FORESTRY AND FISHERIES
60	AGRI-FOODSTUFFS
64	PRODUCTION, TECHNOLOGY AND RESEARCH
66	ENERGY
68	INDUSTRY
72	GEOGRAPHY
76	INTERNATIONAL ORGANISATIONS

Taula 2. Àrees temàtiques de primer nivell del IATE. Font: IATE

El IATE es pot consultar en línia i permet recuperar les fitxes terminològiques completes de cada terme amb les denominacions en diverses de les llengües disponibles. Moltes de les entrades contenen definicions i exemples en diverses llengües. Les denominacions en les diferents llengües tenen associat un índex de fiabilitat, que és una xifra que va de l'1 (fiabilitat no verificada) al 5 (molt fiable), i també tenen associada una categoria, que pot ser *terme* (*fullForm*), *abreviatura* (*abbreviation*), *unitat fraseològica* (*phraseological unit*), *fórmula* (*formula*) i *forma curta* (*shortForm*).

Aquesta base de dades també està disponible per a la seva descàrrega en format TBX (*Term Base eXchange*), que és el format estàndard d'intercanvi de memòries de traducció. En aquest cas es poden descarregar les denominacions en els diversos idiomes disponibles, però no les definicions ni els exemples. La descàrrega es fa en un únic fitxer traduït d'una mida molt gran. Com que el maneig d'un fitxer tan gran no és gens fàcil, en la mateixa pàgina de descàrrega del IATE es pot obtenir un programa que permet fer subconjunts més petits per a determinades llengües i àrees temàtiques.

Per a dur a terme el nostre objectiu de treball, hem fet servir els termes corresponents a l'àrea temàtica de dret del IATE (12-*Law*), que són un total de 15.200 termes amb denominació en anglès i espanyol i que compleixen les condicions següents:

- Que la denominació anglesa estigui formada com a màxim per quatre paraules.
- Que les denominacions anglesa i espanyola tinguin un grau de fiabilitat mínim de 3 (fiable).
- Que les denominacions anglesa i espanyola siguin del tipus terme (*fullForm*).

2.2 La Terminologia Oberta del TERMCAT

El Centre de Terminologia TERMCAT va ser creat l'any 1985 per a garantir el desenvolupament de la terminologia catalana i la seva integració en els diferents sectors especialitzats. El TERMCAT està constituït per la Generalitat de Catalunya, l'Institut d'Estudis Catalans i el Consorci per a la Normalització Lingüística, que s'ocupen respectivament del finançament, la normalització i la difusió de la llengua catalana. El

TERMCAT s'encarrega de normalitzar els mots nous i crear productes terminològics, i té com a principal objectiu garantir la disponibilitat de la terminologia catalana en tots els sectors i activitats, així com també d'afavorir-ne l'ús.

El TERMCAT ofereix un servei de consultes terminològiques en línia (Cercaterm) amb el qual es poden fer cerques a les seves bases de dades de terminologia. A més, moltes de les seves dades terminològiques estan disponibles per a la descàrrega en la col·lecció Terminologia Oberta. Es tracta d'un conjunt de repertoris terminològics procedents de diferents àmbits temàtics que el TERMCAT ofereix des del 2005 i que posa a disposició de la societat. La terminologia que hi ha disponible es pot copiar, distribuir i difondre, ja que es presenta en format estàndard XML, que ajuda a estructurar, interpretar i reutilitzar les dades. La informació es presenta en forma de fitxa conceptual i conté la denominació catalana, espanyola i anglesa, i en algun cas es facilita la definició del terme. Aquesta col·lecció de repertoris temàtics té la funció específica d'enriquir les memòries de traducció, els correctors automàtics i altres aplicacions o programes que ho necessitin.

Hem aprofitat l'accés obert als recursos terminològics existents en la col·lecció Terminologia Oberta per a disposar del conjunt de termes procedents de l'àmbit del dret que actualment hi ha disponibles. A continuació (taula 3) indiquem els diferents repertoris terminològics que han estat utilitzats en el nostre estudi i també el nombre d'entrades a les quals hem tingut accés i de les quals hem pogut disposar.

Repertoris terminològics	**Entrades**
Diccionari de dret administratiu	1.130
Diccionari de dret civil	1.010
Vocabulari de dret penal i penitenciari	2.705
Diccionari de la negociació col·lectiva	590
Diccionari notarial	305
Diccionari de la renda	122
Vocabulari de la responsabilitat social	36
Terminologia electoral bàsica	129
Vocabulari terminològic LGBT (lèsbic, gai, bisexual i transgènere)	190
Consulteca (àmbit dret)	119
Neoloteca (àmbit dret)	223
Total	**6.559**

Taula 3. Repertoris de terminologia oberta de dret. Font: elaboració pròpia.

Cal parar especial atenció a les llicències amb les quals es publiquen els continguts de Terminologia Oberta del TERMCAT, ja que aquestes llicències depenen del format dels fitxers. Els fitxers XML tenen una llicència Creative Commons Reconeixement-Sense obra derivada; en canvi, els fitxers HTML i els PDF a la carta tenen una llicència Creative Commons Reconeixement 3.0. Així, si fem servir les dades en format XML, com és el nostre cas, no en podrem fer una obra derivada. En canvi, si fem ús de les dades en format HTML, sí que en podrem fer una obra derivada. Com que el nostre treball es duu a terme en col·laboració amb el propietari de les dades, el TERMCAT, i els fitxers XML han estat proporcionats per a aquesta tasca, no caldrà aplicar la restricció de la llicència i en podrem fer una obra derivada.

2.3 El corpus del DOGC

El *Diari Oficial de la Generalitat de Catalunya* (DOGC) és el mitjà de publicació oficial de les lleis, les normes, els acords, les resolucions, els edictes, les notificacions i els anuncis de l'Administració i el Govern de Catalunya. El DOGC es va començar a publicar en paper, i des del 2007 està disponible íntegrament en una versió electrònica d'accés lliure.[5] Els textos provinents dels documents del DOGC tenen una llicència lliure i es poden distribuir i processar sense cap limitació legal.[6] Molts dels textos del DOGC apareixen publicats tant en català com en espanyol.

5 Es pot accedir al DOGC des de l'enllaç http://dogc.gencat.cat.

6 El corpus es pot descarregar de https://sourceforge.net/projects/corpus-dogc/.

Com que el contingut del DOGC tracta específicament temes de l'àmbit jurídic, hem considerat oportú incorporar-lo com a recurs lingüístic essencial per a identificar terminologia en llengua catalana d'aquest àmbit. A més, tal com afirma Isabelle (1992), "les traduccions existents contenen més solucions a més problemes de traducció que qualsevol altre recurs". Així, doncs, hem constituït el corpus del DOGC a partir de la descàrrega de la totalitat dels documents des de 1977 fins a 2015. Per *documents* entenem cada una de les lleis, acords, resolucions, edictes, notificacions i anuncis que es publiquen en el DOGC. A continuació, aquest corpus l'hem processat fins a obtenir un corpus paral·lel català-castellà amb el contingut publicat en totes dues llengües. El procés de descàrrega i processament del corpus l'hem dut a terme amb programes desenvolupats en llenguatge de programació Python[7] i fent servir programari lliure. A continuació descrivim les diferents etapes de processament del corpus i el resultat que n'hem obtingut.

1. Descàrrega dels continguts. S'han descarregat els arxius HTML corresponents a tots els documents del DOGC des de 1977 fins a 2015. Els arxius HTML que s'han descarregat es caracteritzen per tenir un codi numèric que comparteixen tant la versió catalana com castellana.

2. Classificació automàtica per anys. No hi ha una correspondència exacta entre el codi numèric d'un determinat document i l'any de publicació. En aquesta etapa classifiquem automàticament els documents per any de publicació per a identificar en l'estructura de l'HTML la data de publicació del document.

3. Conversió d'HTML a text. Els fitxers HTML es converteixen en format de text pla.

4. Verificació de la llengua. La majoria de documents s'ofereixen en català i castellà, però alguns documents de la versió catalana estan disponibles només en castellà i viceversa. Un cop convertits a text pla tots els fitxers, es comprova que els que s'han descarregat com a català realment ho siguin i que els que s'han descarregat en castellà també ho siguin. Per a fer aquesta comprovació, es fa servir un algorisme de detecció automàtica de llengua (Dunning, 1994) en tots els fitxers en format text.

5. Segmentació dels textos amb llengua verificada. Tots els fitxers que tenen la llengua correcta són segmentats, és a dir, es divideixen per unitats de tipus oració.

6. Alineació. Les versions catalana i castellana de cada document, una vegada han estat segmentades, són alineades automàticament fent servir Hunalign (Varga, 2005).

7. Neteja. A partir dels fitxers alineats, es duu a terme un procés de neteja que consisteix a eliminar els segments que contenen només xifres o símbols, els segments llargs que siguin idèntics en català i castellà, etc.

8. Eliminació de segments repetits. El corpus es distribueix en versió íntegra i per anys; per aquest motiu, també s'ha creat una versió del corpus on s'eliminen els segments repetits. A partir de la versió sense segments repetits, també es generen els formats de sortida Moses i memòria de traducció en format TMX (*Translation Memory eXchange*).

El corpus en català i castellà del DOGC, una vegada constituït, consta d'un elevat nombre de segments (frases) en català i castellà. En la taula 4 recollim el nombre de segments que té la totalitat del corpus amb segments repetits i també el corpus sense repeticions.

	Segments	**Paraules català**	**Paraules castellà**
Corpus amb repeticions	8.074.284	188.908.522	197.991.183
Corpus sense repeticions	5.026.847	142.502.123	149.339.268

Taula 4. Mida del corpus del DOGC (1997-2015). Font: elaboració pròpia.

Una vegada s'ha dut a terme tot aquest processament, hem obtingut un corpus lingüístic, és a dir, una recopilació de fragments d'una llengua que se seleccionen i s'ordenen segons un criteri lingüístic amb la finalitat de ser utilitzats com a mostra de la llengua o d'una varietat de llengua (Sinclair, 1996). En el nostre cas, es tracta d'un corpus especialitzat que pot servir de mostra del llenguatge legal i administratiu. El corpus DOGC és un corpus paral·lel, és a dir, un corpus en què els textos apareixen en dues llengües o més i, a més,

7 www.python.org

els segments o oracions originals i traduïts estan alineats, és a dir, hi ha una correspondència entre l'oració traduïda i l'oració original.

Els corpus lingüístics han esdevingut la font principal per al tractament automàtic de la llengua pel fet de poder tractar gran quantitat de dades textuals en suport electrònic (Condamines, 2005) i permeten compilar terminologia d'un àmbit d'especialitat. Quan els corpus són bilingües (Teubert, 2007), aleshores és possible identificar els termes en una llengua i també els corresponents equivalents de traducció. D'aquesta manera, és possible saber quin terme s'ha utilitzat com a equivalent en català, per exemple, de *responsabilidad solidaria* fent una simple consulta en els segments en castellà del corpus. En la taula 5 hi ha una mostra de 3 dels 220 resultats obtinguts d'aquesta cerca. Aquesta mateixa cerca també es pot dur a terme fent servir eines d'extracció automàtica de terminologia, com ara TBXTools (Oliver, 2015).

Segments en castellà	**Segments en català**
2.3 El incumplimiento de la obligación de convocar asamblea general determinará la responsabilidad solidaria de los administradores por las obligaciones sociales nacidas a partir del momento en que expira el plazo para solicitar la disolución.	2.3 L'incompliment de l'obligació de convocar assemblea general determinarà la responsabilitat solidària dels administradors per les obligacions socials nascudes a partir del moment en que expira el termini per a sol·licitar la dissolució.
2) dejar sin efecto la responsabilidad solidaria efectuada per la resolución objeto de recurso	2) deixar sense efecte la responsabilitat solidària efectuada per la resolució objecte de recurs
La segunda es una responsabilidad solidaria de los administradores, puesto que se trata de una responsabilidad por deuda ajena y no por actos propios	La segona és una responsabilitat solidària dels administradors, degut a que es tracta d'una responsabilitat per deute aliè i no per actes propis

Taula 5. Fragments del corpus paral·lel. Font: elaboració pròpia.

Ara bé, quan la mida del corpus paral·lel és molt gran, com és el cas del corpus del DOGC, cal fer servir algorismes d'indexació eficients per a ser capaços de fer les cerques i disposar dels resultats en molt poc temps. En el present treball fem servir models de traducció automàtica estadística calculats amb Moses (Koehn, 2007) per a poder identificar els equivalents del IATE en català i castellà d'una manera eficient. Moses[8] és un sistema de traducció automàtica estadística que proporciona totes les eines necessàries per a entrenar sistemes de traducció i dur a terme traduccions amb els sistemes entrenats. Moses es distribueix amb una llicència de programari lliure.[9] Encara que la construcció dels models de traducció demana molt de temps de càlcul, una vegada han estat calculats els models, aquests ja estan disponibles per a fer tantes cerques com calgui. Les taules de traducció de frases creades per Moses ofereixen totes les possibles traduccions per cada combinació de paraules en la llengua de partida. A continuació podem observar un fragment de la taula de traducció que conté el terme castellà *responsabilidad solidaria*, l'equivalent de traducció i la probabilitat que té de ser un equivalent adequat (taula 6).

responsabilidad solidaria ||| responsabilitat solidària d'; ||| 0.02361172 0.3819126 0.0197653 0.00998172 ||| 0-0 1-1 ||| 36 50 2 ||| |||
responsabilidad solidaria ||| responsabilitat solidària de ||| 0.0285722 0.190684 0.0060614 0.048065 ||| 0-0 1-1 ||| 7 33 1 ||| |||
responsabilidad solidaria ||| responsabilitat solidària ||| 0.9640546 0.9515062 0.9652966 0.951715 ||| 0-0 1-1 ||| 48 50 47 ||| |||

Taula 6. Mostra de la taula de traducció. Font: elaboració pròpia.

El conjunt d'eines distribuïdes amb Moses és actualment un dels més emprats per a la creació de sistemes de traducció estadístics. Es distribueixen eines específiques per a la identificació d'equivalents de traducció de paraules o combinacions de paraules que són molt eficients, encara que es facin servir amb corpus paral·lels de gran mida. Així, doncs, en el nostre treball hem fet servir aquestes eines per a poder disposar de manera

8 http://www.statmt.org/moses/

9 LGPL - GNU Lesser General Public License

eficient de tots els equivalents de traducció del corpus del DOGC, juntament amb el valor de probabilitat que té cada equivalent de traducció de ser-ho. En la taula 6 observem que hi ha tres equivalents de traducció per al terme castellà *responsabilidad solidaria*. Per a identificar quin és l'equivalent idoni d'aquest terme en català, prenem com a referència el valor de probabilitat que ofereix el model de traducció de Moses. En aquest cas, el terme en català que té una major probabilitat de ser l'equivalent de traducció és *responsabilitat solidària* (0.9640546).

La taula de traducció obtinguda a partir del corpus del DOGC és d'una mida molt gran (192.835.512 línies) i, per tant, una consulta directa d'aquesta taula per cada un dels termes a cercar també seria molt lenta. Per a reduir el temps de cerca amb una eina d'extracció automàtica de terminologia com TBXTools, hem implementat dues estratègies:

a. L'ús de taules compactes (Junczys-Dowmunt, 2012) fent servir els algorismes que proporciona la distribució de Moses. Un cop compactada la taula es poden fer consultes de manera molt eficient fent servir el programa *queryPhraseTableMin*, que també es distribueix amb Moses.

b. L'ús de taules de traducció indexades fent servir estructures de dades de tipus diccionari de Python. Aquestes estructures es poden emmagatzemar a disc i consultar posteriorment d'una manera molt eficient.

La implementació d'aquestes dues estratègies ens ha permès recuperar del corpus del DOGC els equivalents de traducció en català i castellà de l'àmbit jurídic del IATE d'una manera ràpida i exhaustiva.

2.4 Part experimental

Per tal de poder desenvolupar el IATE en llengua catalana corresponent a l'àrea temàtica de dret, la tasca que hem dut a terme ha estat compilar els termes en anglès d'aquesta base de dades que disposen d'equivalent de traducció en castellà i que són de l'àmbit de dret. Així, a partir de la denominació en castellà hem buscat de manera automàtica la corresponent denominació catalana fent servir dues estratègies:

a. Cercar els equivalents de traducció en el conjunt de repertoris terminològics de Terminologia Oberta que tenim disponibles de l'àmbit temàtic del dret.

b. Cercar els equivalents de traducció en el corpus paral·lel del DOGC.

Per a dur a terme aquesta tasca hem fet servir la versió descarregable del IATE disponible l'11 d'octubre del 2016. En aquesta versió, i per a l'àrea de dret (codi 12-*Law*), hem recuperat un total de 15.200 termes en anglès que tenen equivalent en castellà.

2.4.1 Cerca d'equivalents de traducció a Terminologia Oberta

La cerca d'equivalents de traducció en els repertoris terminològics de Terminologia Oberta ens permet obtenir els equivalents catalans corresponents als termes del IATE que tenen denominació en anglès i castellà. Per a fer aquesta cerca, prenem com a referència la denominació castellana del terme en el IATE i comprovem si hi ha una entrada terminològica coincident en algun dels repertoris disponibles de Terminologia Oberta de l'àrea de dret. Si n'hi ha, relacionem l'entrada terminològica del IATE amb la denominació catalana i, en cas que l'entrada terminològica disposi de definició, també recuperem aquesta informació. El procés d'assignació de l'equivalent català es duu a terme de manera automàtica, mitjançant una sèrie d'algorismes programats en llenguatge Python. A continuació podem observar un exemple d'aquesta relació (taula 7).

eng (IATE)	stateless person
spa (IATE - TO)	apátrida
cat (TO)	apàtrida
POS (cat) (TO)	n m, f
Definició (TO)	Persona que no té cap nacionalitat o que l'ha perduda sense adquirir-ne una altra, la situació jurídica de la qual depèn de la legislació de l'estat en què resideix.
Tema (TERMCAT)	Dret internacional

Codi (TERMCAT)	Neoloteca-336
Codi (IATE)	IATE-775760
Tema (IATE)	1231, 1236003
Tipus (IATE)	fullForm
Fiabilitat (IATE)	3

Taula 7. Correspondència IATE i Terminologia Oberta. Font: elaboració pròpia.

L'avantatge de fer servir aquest mètode és que, una vegada hem identificat la denominació catalana d'un determinat terme, a partir del codi IATE també podem localitzar les denominacions en les altres llengües presents en aquesta base de dades. Per a l'exemple, el terme IATE-775760 (*stateless person*) té les denominacions següents: *лице без гражданство* (búlgar), *statsløs* i *statsløs person* (danès), *Staatenlose* (n.f.) i *Staatenloser* (n.m.) (alemany), *ανιθαγενής* (grec), *stateless person* (anglès), *apátrida* (castellà), *kodakondsuseta isik* (estonià), *kansalaisuudeton henkilö* (finès), *apatride* (francès), *duine gan stát* (gaèlic), *osoba bez državljanstva* (croat), *hontalan személy* (hongarès), *apolide* (italià), *asmuo be pilietybės* (lituà), *apatrīds* i *bezvalstnieks* (letó), *apolidi* i *persuna apolida* (maltès), *staatloze* i *staatloze persoon* (holandès), *apatryda* i *bezpaństwowiec* (polonès), *apátrida* (portuguès), *apatrid* (romanès), *osoba bez štátnej príslušnosti* (eslovac), *apatrid* i *oseba brez državljanstva* (eslovè), *statslös* i *statslös person* (suec). I ara també la denominació catalana *apàtrida*. Així, doncs, el resultat que obtenim fent servir aquest mètode és la constitució d'un repertori terminològic de l'àmbit del dret en català i 24 llengües més.

Una vegada hem recuperat automàticament els equivalents en català dels repertoris terminològics, hem dut a terme una revisió lingüística dels resultats. Per a poder fer aquesta revisió hem recollit les denominacions en anglès i espanyol extretes del IATE que pertanyen a l'àrea temàtica de dret del IATE (12-*Law*) i les denominacions en català extretes de Terminologia Oberta, la definició, l'àrea temàtica recollida pel TERMCAT, les obres terminològiques del TERMCAT de les quals prové cada denominació, el codi IATE que correspon a la denominació en anglès, l'àrea temàtica del IATE i el nivell de fiabilitat de la denominació en anglès en el IATE (taula 8).

Codi IATE:	IATE-914286
Denominació anglesa	legal aid
Denominació espanyola	asistencia de letrado, asistencia letrada
Denominació catalana	assistència lletrada, assistència de lletrat
Definició	Dret d'un detingut a sol·licitar un advocat per a la supervisió de totes les diligències declaratòries i d'identificació que el tinguin com a objecte i a entrevistar-s'hi reservadament.
Àrea temàtica TERMCAT	Dret processal penal
Obra terminològica TERMCAT	Vocabulari de dret penal i penitenciari
Àrea temàtica IATE	12
Tipus IATE	fullForm
Fiabilitat IATE	4

Taula 8. Exemple d'equivalent en català de Terminologia Oberta. Font: elaboració pròpia.

Per tal d'avaluar la pertinença de les denominacions hem tingut en compte que en la fitxa IATE hi hagi la definició completa del terme, que la fitxa IATE faci referència al mateix concepte que recull el TERMCAT i que el nivell de fiabilitat del terme a IATE sigui preferentment de nivell 3 o 4, és a dir, que tingui un nivell de fiabilitat alt. Tenint en compte aquests criteris, hem considerat com a incorrectes les denominacions que en la fitxa IATE no tenen definició (*proclamation* or *public notice*) o bé que tenen una definició incompleta (*entail*), i també les denominacions que en les obres del TERMCAT fan referència a un concepte diferent del que recull la fitxa IATE (*business establishment*) o bé a un concepte parcialment diferent (*Council of State*).

Tenint en compte aquests criteris, hem pogut calcular els valors de precisió (nombre d'extraccions correctes respecte el total d'extraccions) i cobertura (nombre d'extraccions correctes respecte el nombre de termes en anglès) d'aquests resultats. Així, el nombre de termes amb denominació anglesa i castellana és de 15.200, i

hem pogut obtenir 2.099 denominacions en català, 2.066 de les quals són correctes. Això ens permet obtenir els valors de precisió i cobertura següents (taula 9):

Precisió	98,42 %
Cobertura	13,81 %

Taula 9. Resultats de la cerca d'equivalents de traducció a Terminologia Oberta. Font: elaboració pròpia

Com podem observar, la precisió és molt elevada, però la cobertura és baixa. El motiu d'aquesta baixa cobertura es deu simplement al fet que molts termes amb denominació espanyola al IATE no són presents a la Terminologia Oberta del TERMCAT.

2.4.2 Cerca d'equivalents de traducció al corpus del DOGC

La identificació d'equivalents de traducció en català al corpus del DOGC es duu a terme de manera automàtica, mitjançant una sèrie d'algorismes programats en llenguatge Python que fan servir l'eina TBXTools d'extracció automàtica de terminologia. Concretament fem servir la tècnica d'extracció d'equivalents a partir de les taules de traducció calculades amb Moses i indexades en una estructura de dades de tipus diccionari de Python. Així, el sistema cerca automàticament si cada terme en castellà és present en la taula de traducció indexada i, si hi és, ofereix els possibles equivalents de traducció endreçats per probabilitat. Al final del procés, els resultats obtinguts són revisats manualment. En la taula 10 podem observar un exemple d'aquesta identificació d'equivalents en català.

Codi IATE:	IATE-126975
Denominació anglesa	employment contract
Denominació espanyola	contrato de trabajo
Denominació catalana correcta	contracte de treball
Opcions automàtiques	contracte de treball: contracte de feina: contractes de treball: contracte de Treball: contracte de treball: contracte de traball: contracte de treball: Contracte de Treball: contracte de treball i amb possibilitat: pertinent contracte de treball
Àrees temàtiques IATE	12, 44
Tipus IATE	fullForm
Fiabilitat IATE	3

Taula 10. Exemple d'extracció d'equivalent en català del corpus del DOGC. Font: elaboració pròpia.

Dels resultats obtinguts amb aquest mètode hem calculat el valor de precisió i de cobertura obtinguts. El nombre de termes amb denominació anglesa i castellana al IATE és de 15.200, i hem pogut obtenir 2.684 denominacions en català del corpus del DOGC, 2.547 de les quals és correcta la primera de les opcions. Si tenim en compte quantes denominacions tenen alguna de les opcions de traducció correcta, n'obtenim 2.592. D'aquesta manera, disposem de dos valors de precisió i cobertura, que recollim en la taula 11. Tal com es pot observar, la majoria de les extraccions correctes s'obtenen com a primera opció.

Precisió (primera posició):	94,90 %
Cobertura (primera posició):	16,76 %
Precisió (totes les posicions)	96,57 %
Cobertura (totes les posicions)	17,05 %

Taula 11. Resultats de precisió i cobertura de la cerca d'equivalents de traducció en el corpus del DOGC. Font: elaboració pròpia.

Malgrat obtenir una cobertura més alta fent servir el corpus del DOGC que no pas fent servir els repertoris terminològics de la Terminologia Oberta, aquesta continua essent força baixa. El motiu és que solament un petit percentatge de les denominacions espanyoles dels termes del IATE de l'àmbit del dret són presents en el corpus del DOGC.

3 Resultats i discussió

Els resultats que hem obtingut en aquest estudi constaten la capacitat de compilar terminologia jurídica del IATE en llengua catalana fent servir eines de tractament de corpus multilingües. Així mateix, es mostra la possibilitat de completar la terminologia d'aquesta base de dades amb llengües que no són presents en la Unió Europea. Aquesta tasca d'ampliació del IATE ha estat possible dur-la a terme amb l'aprofitament d'un conjunt de recursos terminològics que actualment són disponibles en català i que han estat revisats per experts i validats per terminòlegs.

El procés de compilació de la terminologia jurídica en llengua catalana s'ha realitzat en dues fases. En una primera fase, hem identificat de manera automàtica els equivalents de traducció de l'àmbit jurídic presents a Terminologia Oberta i en el corpus del DOGC. En el cas de Terminologia Oberta, aquesta cerca s'ha fet a partir de l'àrea temàtica i l'equivalent en espanyol del terme en anglès i, pel que fa al corpus del DOGC, s'ha aplicat el mètode de les taules de traducció indexades per a localitzar l'equivalent en català corresponent al terme en castellà del IATE. En aquest procés automàtic de compilació de la terminologia jurídica en llengua catalana, hem identificat un total de 2.099 denominacions en català procedents dels repertoris terminològics de terminologia oberta del TERMCAT d'un total de 6.559 denominacions disponibles i 2.684 denominacions en català corresponents al corpus del DOGC.

En una segona fase, hem procedit a revisar manualment el conjunt d'equivalents en català identificats automàticament en la fase anterior. Amb referència a les denominacions de Terminologia Oberta, n'hem validat com a correctes un total de 2.066 (98,42 %), és a dir, pràcticament la totalitat de les denominacions identificades automàticament. I pel que fa a les denominacions del corpus del DOGC, n'hem seleccionat 2.592 (96,57 %) com a correctes, de manera semblant al grup anterior, gairebé la totalitat de les denominacions que han estat identificades automàticament corresponien a la denominació en anglès recollida al IATE. En els dos conjunts de resultats hem identificat un total de 523 entrades coincidents, és a dir, entrades del IATE en català que s'han obtingut tant a partir de Terminologia Oberta com a partir del DOGC. Així, amb els resultats que hem obtingut en aquest estudi hem pogut extreure un total de 4.658 denominacions catalanes de l'àmbit del dret i relacionar-les amb les corresponents entrades terminològiques del IATE. En la taula 12 podem observar que, per a l'àmbit del dret, les extraccions obtingudes situarien el català com a onzena llengua pel que fa al nombre de denominacions.

Llengua	**Nombre de denominacions**
Francès	85.886
Italià	35.546
Anglès	28.989
Holandès	24.996
Espanyol	19.659
Portuguès	19.120
Alemany	16.415
Finès	15.991
Grec	15.945
Suec	14.978
Català	**4.658**
Irlandès	4.457
Polonès	4.375
Eslovè	3.146
Maltès	3.137
Lituà	2.987
Romanès	2.979
Eslovac	2.943
Danès	2.672
Hongarès	2.632
Búlgar	2.588

Letó	2.588
Estonià	2.506
Txec	1.809
Croat	867
Llatí	758

Taula 12. Denominacions de l'àmbit del dret en les llengües del IATE i el català. Font: elaboració pròpia.

En el procés de revisió manual de les denominacions en anglès presents en el IATE i els equivalents de traducció en català procedents de la Terminologia Oberta del TERMCAT i el corpus del DOGC, hem identificat una sèrie de correspondències anòmales que ha calgut esmenar. A continuació mostrem la tipologia de casos erronis o incomplets que hem recollit i que hem esmenat.

- Posar en infinitiu algunes formes verbals, ja que els equivalents identificats automàticament corresponien a formes verbals conjugades (*to make a report* - *informar*).
- Normalitzar majúscules i minúscules quan els substantius o adjectius són amb majúscules (*Sub-Saharan Africa* - *Àfrica subsahariana*).
- Localitzar equivalents de traducció en sintagmes preposicionals, ja que en català s'han identificat estructures oracionals (*without prejudice to* - *sense perjudici de*).
- Lematitzar adequadament alguns termes (*terrorist offence* [s.] - *delictes de terrorisme* [pl.]).
- Desestimar algun terme per manca d'equivalent de traducció en català (*incest*, *range*).
- Substituir l'equivalent de traducció en català per presentar una correspondència errònia (*to kill* - *òptica* substituït per *to kill* - *matar*).
- Completar la correspondència d'equivalents en català d'un terme anglès quan hi ha més d'una opció correcta (*member* - *associat* i *soci*).
- Seleccionar l'equivalent de traducció adequat per a un terme en anglès quan hi ha més d'un equivalent disponible (*to rank* - *rang, tipus, categoria*; *termination*, *lay off* - *acomiadament*).
- Esmenar els equivalents de traducció en català que no corresponen al terme en anglès (*Corruption Perceptions Index* - *índex de preus de consum*; *continuity check indicator* - *índex de preus de consum*).

El mètode de treball que hem aplicat en el nostre estudi ens permet confirmar, d'una banda, que l'aprofitament dels recursos terminològics revisats per especialistes i validats pel TERMCAT ens permet compilar la terminologia jurídica en català corresponent a la base de dades IATE d'una manera àgil i eficaç. Així mateix, la revisió manual final de les correspondències en català assignades automàticament fa possible assegurar la bondat dels equivalents identificats. I, a més, ens ofereix l'oportunitat de crear nous repertoris terminològics multilingües que incloguin la llengua catalana, gràcies a tenir accés al conjunt de dades del IATE en obert. D'altra banda, la compilació de terminologia en llengua catalana de manera automatitzada representa un avenç important en l'ampliació de la base de dades IATE en termes econòmics i d'estalvi de temps. En aquest sentit, únicament intervenen els especialistes en la matèria en l'última fase de treball, que correspon a la validació de les denominacions identificades automàticament o bé que no han estat localitzades en cap dels glossaris o publicacions especialitzades de l'àmbit.

Quant a la identificació d'equivalents de traducció en català a partir del corpus del DOGC, constatem que l'ús de taules de traducció indexades ha permès agilitar el procés de cerca dels equivalents de traducció en el corpus respecte dels mètodes tradicionals. La diferència més significativa que introdueixen les taules de traducció indexades respecte dels mètodes tradicionals és que permeten calcular la totalitat dels equivalents de traducció d'un corpus multilingüe, cosa que en facilita la identificació posterior en poder fer cerques recursives de termes. Els mètodes tradicionals, en canvi, permeten identificar solament els equivalents de

traducció d'una selecció prèvia de termes, fet que n'alenteix la cerca i per a cada nova cerca d'equivalent s'ha de preparar una nova selecció de termes.

Finalment, cal dir que la participació del TERMCAT en el desenvolupament d'aquest projecte garanteix la correcta aplicació de la metodologia del treball terminològic i també fa possible la intervenció de terminòlegs especialitzats en l'àmbit del dret per a verificar l'adequació dels resultats finals.

4 Conclusions i treball futur

En el present article hem descrit un nou mètode de treball que té com a objectiu aplicar diferents tècniques de processament de llenguatge natural en la creació de la terminologia jurídica en llengua catalana de la base de dades IATE. Concretament, es fan servir taules de traducció procedents de sistemes de traducció automàtica de tipus estadístic per a calcular els equivalents de traducció en català corresponents als termes de l'àmbit del dret presents en la base de dades IATE. Per a fer-ho, aquest mètode es basa en l'aprofitament de recursos lingüístics i l'ús de corpus multilingües ja existents amb la finalitat d'identificar denominacions catalanes de l'àmbit del dret i relacionar-les amb les corresponents entrades terminològiques de la base de dades IATE.

L'aplicació d'aquest mètode de treball ens ha permès constatar que és viable la creació de la terminologia jurídica en llengua catalana del IATE a partir de recursos lingüístics ja existents. En aquest sentit, de manera automatitzada s'han pogut identificar un 30,64% de denominacions en català extretes de Terminologia Oberta i el DOGC del total de denominacions en anglès i espanyol presents en el IATE. Per a completar la identificació del conjunt de denominacions en català cal ampliar el mètode de treball amb nous recursos lingüístics que hi hagi disponibles.

L'accés obert a les dades del IATE permet crear nous repertoris terminològics en què també s'inclogui la llengua catalana i posar-los a disposició d'especialistes, traductors, correctors i usuaris en general. La creació d'aquests repertoris és possible gràcies a l'aprofitament de recursos lingüístics ja existents com és la Terminologia Oberta del TERMCAT, institució que també participa en el desenvolupament d'aquest projecte i que és el fòrum idoni per fer la difusió dels equivalents en llengua catalana dels termes presents en el IATE que es vagin compilant en el si d'aquest projecte. D'aquesta manera, es posa en valor la tasca de normalització terminològica que duu a terme el centre i l'alt nivell d'aprofitament que tenen les seves bases de dades terminològiques en el si de projectes en què intervingui el processament del llenguatge natural. I també és possible fent ús de corpus lingüístics que estiguin disponibles en obert, els quals esdevenen una font d'informació essencial en la tasca de creació automatitzada de nous repertoris terminològics. En la present recerca s'ha fet servir concretament el corpus paral·lel del *Diari Oficial de la Generalitat de Catalunya*, que es pot anar ampliant periòdicament amb les noves publicacions que hi hagi disponibles.

Pel que fa concretament a l'àmbit jurídic, el fet de disposar de repertoris terminològics específics en llengua catalana que es puguin anar ampliant progressivament representa una contribució rellevant a la tasca de normalització en aquest àmbit. A més, permet completar el contingut que actualment ja hi ha disponible en portals específics de terminologia jurídica.

La necessitat de disposar de la terminologia que hi ha present en el IATE en llengua catalana fa plantejar-nos com a tasca de futur l'aplicació del mètode de treball que hem presentat a altres àmbits d'especialitat, per a poder disposar de repertoris terminològics en llengua catalana corresponents a les diferents àrees temàtiques del IATE. Per a fer-ho, incorporarem en l'estudi altres recursos lingüístics que actualment ja hi ha disponibles com ara materials docents de la Universitat Oberta de Catalunya.

Finalment, també com a treball futur ens plantegem d'actualitzar anualment el corpus del DOGC que actualment tenim disponible, a fi de poder ampliar el nombre d'equivalents de traducció en llengua catalana de diferents àmbits d'especialitat.

5 Referències bibliogràfiques

Condamines, Anne. «Linguistique de corpus et terminologie». *Langages*, Vol. 39, núm. 157 (2005), p. 36–47.

Dunning, Ted. *Statistical identification of language*. Computing Research Laboratory: New Mexico State University, 1994.

Gaizauskas, Robert; Paramita, Monica Lestari; Barker, Emma; Pinnis, Marcis; Aker, Ahmet; Solé, Marta Pahisa. «Extracting bilingual terms from the Web». *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*. Vol. 2, núm. 21 (2015), p. 205-236.

Isabelle, Pierre. «Bi-textual aids for translators». A: *Proceedings of the Annual Conference of the UW Center for the New OED and Text Research*, 1992.

Johnson, Ian; Macphail, Alastair. «IATE-Inter-Agency Terminology Exchange: development of a single central terminology database for the institutions and agencies of the European Union». A: *Workshop on Terminology resources and computation,* 2000.

Junczys-Dowmunt, Marcin. «Phrasal rank-encoding: Exploiting phrase redundancy and translational relations for phrase table compression». *The Prague Bulletin of Mathematical Linguistics*. Vol. 98 (2012), p. 63-74.

Koehn, Philipp; Hoang, Hieu; Birch, Alexandra; Callison-Burch, Chris; Federico, Marcello; Bertoldi, Nicola; Cowan, Brooke; Shen, Wade; Moran, Christine; Zens, Richard; Dyer, Chris; Bojar, Ondrej: Constantin, Alexandra; Herbst, Evan. «Moses: Open Source Toolkit for Statistical Machine Translation». A: *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*. Praga (República Txeca): Association for Computational Linguistics, 2007, p. 177-180.

Macken, Lieve; Lefever, Els; Hoste, Véronique. «TExSIS: bilingual terminology extraction from parallel corpora using chunk-based alignment». *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*. Vol. 19, núm. 1 (2013), p. 1-30.

Oliver, Antoni; Vàzquez, Mercè. «TBXTools: A Free, Fast and Flexible Tool for Automatic Terminology Extraction». A: *International Conference on Recent Advances in Natural Language Processing* (RANLP 2015), 2015, p.473-479.

Oliver, Antoni. *Herramientas tecnológicas para traductores*. Editorial UOC, 2016.

Pazienza, Maria Teresa; Pennacchiotti, Marco; Zanzotto, Fabio. «Terminology Extraction: an Analysis of Linguistic and Statistical Approaches». A: *Knowledge Mining. Studies in Fuzziness and Soft Computing*. Vol. 185. Heidelberg, Berlín: Springer, 2005, p. 255-279.

Kilyeni, Annamaria, Ciobanu, Georgeta; Palea, Audina. «Designing a database for landscape architecture terminology». *Procedia-Social and Behavioral Sciences*. Vol. 46 (2012), p. 4.666-4.671.

Teubert, Wolfgang (ed.). *Text Corpora and Multilingual Lexicography*. Amsterdam: John Benjamins, 2007.

Varga, Dániel; Németh, Lásló; Halácsy, Peter; Kornai, Andras; Trón, Viktor; Nagy, Vitkor. «Parallel corpora for medium density languages». A: *Proceedings of the Recent Advances in Natural Language Processing* (RANLP 2005), p. 590-596.

Sinclair, John. «Preliminary recommendations on corpus typology». *EAGLES Document TCWG-CTYP/P* (1996). <http://www.ilc.cnr.it/EAGLES/corpustyp/corpustyp.html>