

La terminologia jurídica en català. Aproximació descriptiva dels seus patrons morfosintàctics

ALFRED ISERTE BUSQUETA
Universitat Oberta de Catalunya
aiserter@uoc.edu

Alfred Iserte Busqueta és llicenciat

en Ciències Econòmiques i Empresarials (1990) per la Universitat de Barcelona (UB), màster d'Especialització Tributària (1991) i llicenciat en Filologia Catalana (2015) per la Universitat Oberta de Catalunya (UOC).

Els seus interessos principals són la lexicologia en general i l'estudi de la terminologia jurídica i econòmica en particular. Exerceix l'activitat professional d'assessor fiscal i és membre del Registre d'Economistes Assessors Fiscals d'Espanya.



Resum

En aquest treball es descriuen els patrons morfosintàctics que segueix la terminologia jurídica en català amb l'objectiu de contribuir a establir estratègies lingüístiques útils en l'extracció automatitzada de terminologia i es descriu un cas pràctic d'extracció terminològica de termes jurídics amb el programa TBXTools.

PARAULES CLAU: terminologia catalana; terminologia jurídica; patrons morfosintàctics; extracció terminològica automatitzada

Abstract

Legal terminology in Catalan. A descriptive approach to its morphosyntactic patterns

This study describes the morphosyntactic patterns of legal terminology in the Catalan language to help to establish linguistic strategies which are useful in automatic terminology extraction, presenting a case study of the extraction of legal terms with the TBXTools program.

KEYWORDS: Catalan terminology; legal terminology; morphosyntactic patterns; automatic terminology extraction

TERMINÀLIA 14 (2016): 24-30 · DOI: 10.2436/20.2503.01.96
Data de recepció: 08/09/2015. Data d'acceptació: 12/04/16
ISSN: 2013-6692 (impresa); 2013-6706 (electrònica) · <http://terminalia.iec.cat>

1 Introducció

Malgrat les divergències existents en la conceptualització dels llenguatges d'especialitat, hi ha una coincidència mínima a entendre que presenten unes característiques que els fan diferents del llenguatge general perquè vehiculen l'expressió i la transmissió d'àrees específiques de coneixement. El domini conceptual d'una especialitat s'estructura en categories conceptuais dinàmiques i prototípiques a partir de les quals es poden extraure patrons sintàctics i semàntics recurrents. Així, un nombre limitat de categories específiques de cada domini serveix per estructurar els conceptes, que, al seu torn, remetent a unitats terminològiques de cada llengua. Un terme és l'expressió d'un concepte.

L'objectiu de la majoria d'aproximacions a l'extracció terminològica ha estat obtenir el conjunt de termes més significatiu d'un corpus, és a dir, aquell que representi millor el domini conceptual per a un expert.

Segons la metodologia aplicada, aquestes aproximacions es poden agrupar en tres blocs: les aproximacions basades en l'aplicació de tècniques estadístiques, les basades en anàlisis lingüístiques i les que podríem anomenar mixtes, que apliquen una combinació de metodologies estadístiques i lingüístiques. Aquestes últimes són les que s'han demostrat més eficaces. Normalment, l'extracció comença per una fase lingüística, que actua a tall de filtratge, i posteriorment es passa a la fase estadística.

Les anàlisis lingüístiques cerquen la identificació de termes a partir de les seves estructures sintàctiques característiques, perquè es considera que la seva funció de representació semàntica imposa unes restriccions en la seva forma que obeeixen a unes regles de correcta formació sintàctica anomenades *composició sinàptica*.

L'experiència de treballs realitzats fins ara permet concloure que els termes que generalment apareixen en forma de sintagmes nominals curts, formats per dos elements amb significat, que alguns autors anomenen *termes base* (Pazienza et al., 2005; Daille, 1994), són les estructures més rellevants. A partir d'aquesta idea d'associació entre dos elements, Daille proposa una definició complementària de *terme* com a coocurrència. Els termes serien, doncs, les coocurrències de dos elements amb les propietats següents:

- Poden ser descrites a partir de la seva estructura morfosintàctica.
- Estan formades per dos ítems principals.
- Accepten modificacions que poden generar nous termes que poden estar formats per més de dos ítems.
- Accepten variants.

L'objectiu primer d'aquest treball és determinar aquestes estructures, que nosaltres denominarem aquí *patrons morfosintàctics*, de les unitats terminològiques de

l'àmbit de coneixement jurídic, vàlides per contribuir a establir estratègies lingüístiques útils en el disseny d'aplicacions d'extracció automatitzada de terminologia jurídica.

La metodologia i el punt de vista aplicats en el seu desenvolupament es basen en el marc teòric de la teoria comunicativa de la terminologia (TCT) formulat per M. Teresa Cabré (1999).

L'estudi es fa sobre les unitats terminològiques contingudes en el corpus terminogràfic *Justiterm* (<http://justicia.gencat.cat/ca/serveis/justiterm>), publicat pel Departament de Justícia de la Generalitat de Catalunya. Les entrades actuals de *Justiterm* són 8.719 unitats, de les quals 6.830 (el 78,33 % del total) són sintagmes nominals.

Es descriuen tres aspectes essencials d'aquestes unitats terminològiques:

- La seva categoria gramatical.
- Les estructures o patrons morfosintàctics que segueixen.
- La freqüència d'aquestes estructures per determinar-ne la productivitat.

Com a segon objectiu del treball, ens proposem realitzar un experiment d'extracció terminològica sobre un corpus textual constituït amb els números del *Diari Oficial de la Generalitat de Catalunya* (DOGC) publicats l'any 2013, mitjançant l'eina TBXTools.

2 Anàlisi dels sintagmes nominals

Aquest treball es centra en l'anàlisi dels sintagmes nominals perquè, com és comú a tots els llenguatges d'especialitat, la categoria nominal és la prototípica dels termes com a nodes de coneixement especialitzat (Cabré, 1999). També es farà una descripció dels adjectius perquè, com es veurà més endavant, formen part del patró morfosintàctic més productiu (N A) d'unitats terminològiques polilèxiques.

2.1 Estructures formals i freqüències

Els sintagmes nominals adopten els patrons morfosintàctics que apareixen en la taula 1, ordenats per freqüència de casos.

Patrons morfosintàctics	Formes	Casos	%
N (nom)	1	3.102	45,42 %
N SA (nom + sintagma adjectival)	62	2.052	30,04 %
N SP (nom + sintagma preposicional)	128	1.520	22,25 %
Altres 10 patrons	21	156	2,29 %
Totals	212	6.830	100 %

TAULA 1. Patrons morfosintàctics de les entrades del *Justiterm*

Els patrons dels sintagmes nominals presents en aquest corpus adopten fins a 212 formes diferents. Per formes entenem les diferents composicions que pot adoptar un patró morfosintàctic. Per exemple, el patró N SP pot adoptar diverses formes, com N de N, N de NA, N a N, etc. La major part de formes i ocurrencies es concentren en tres patrons. La unitat nominal monolèxica és la més freqüent amb gairebé la meitat de les ocurrencies. Aquesta unitat N i les unitats polilèxiques que segueixen els patrons N SA i N SP representen el 98 % del total.

El patró d'estructura N SP és el que presenta més varietat de formes, seguit de l'estructura N SA. Entre els dos reuneixen el 90 % del total de formes.

Aquests resultats són semblants als obtinguts per Estopà (1996). En el seu treball dedicat exclusivament a l'estudi de les unitats terminològiques polilèxiques (UTP), Estopà va analitzar una mostra de 690 unitats (d'un total de 1.099 UTP) contingudes en el *Diccionari jurídic català*, que en aquell moment disposava de 3.045 entrades. Els patrons morfosintàctics més productius són els mateixos en els dos estudis.

Els patrons morfosintàctics més freqüents per àrees i subàrees d'especialitat del corpus *Justiterm* són els que es mostren a la taula 2.

2.2 Anàlisi de les unitats nominals monolèxiques

Per tal de tenir una aproximació a les característiques dels noms continguts en aquesta base de dades, hem fet l'anàlisi de les primeres 1.000 unitats nominals monolèxiques amb els resultats següents:

- La major part dels noms són cultismes o semicultismes (66,20 %) i noms patrimonials d'origen llatí (23,90 %).
- La sufixació és el principal mètode de formació (51 % dels casos). Dels 47 sufixos diferents, el més productiu és *-ció* (39 %) seguit de *-ment* (19 %). Amb 9 sufixos es deriven el 89 % dels mots sufixats.
- La major part dels noms tenen origen verbal.

No es detecta cap característica específica dels termes nominals monolèxics. El tret més freqüent és la construcció per sufixació amb *-ció* d'un mot culte deverbal, però tot just representa el 19 % del total de termes i, a més, el sufix *-ció* és el més comú en la formació de mots verbals: al DIEC trobem exactament 1.000 mots amb aquesta terminació.

Àrees/subàrees	Total		Dret administratiu		Dret civil		Dret penal penitenciari		Vocabulari general jurídic i administratiu	
Total d'entrades	6.830	100 %	1.285	19 %	1.087	16 %	2.535	37 %	1.923	28 %
N	3.102	45 %	208	16 %	435	40 %	770	30 %	1.689	88 %
N A	1.639	24 %	394	31 %	332	31 %	809	32 %	104	5 %
N A A	179	3 %	88	7 %	17	2 %	70	3 %	4	0 %
N de N	821	12 %	209	16 %	151	14 %	394	16 %	67	3 %
N de N A	142	2 %	74	6 %	7	1 %	59	2 %	2	0 %
N de art N	109	2 %	14	1 %	11	1 %	76	3 %	8	0 %
Total	5.992	88 %	987	77 %	953	88 %	2.178	86 %	1.874	97 %

TAULA 2. Patrons morfosintàctics per àrees i subàrees d'especialitat

2.3 Anàlisi de les unitats adjectivals monolèxiques

L'anàlisi dels sintagmes adjectivals independents presents al *Justiterm* proporciona les dades següents:

- La pràctica totalitat d'entrades són monolèxiques (631).
- Hi ha 14 casos de lemes amb més d'una entrada (polisèmics) i 64 casos de conceptes amb més d'un lema (sinònims).
- La majoria pertanyen al vocabulari general jurídic i administratiu (552).
- La seva estructura morfològica és diversa amb predomini dels derivats (43 %).
- Els derivats per sufixació són 274 amb 22 sufixos diferents. Els més freqüents són *-al* (28 %) i *-ble* (15 %).
- Els derivats prefixats són 86, amb un clar predomini (78 % dels casos) del prefix *in-* i els seus diversos *-lomorfs*.

Alguns dels adjectius que disposen d'entrada específica en aquest diccionari són d'ús habitual en el llenguatge comú (*lleu, greu, imminent, natural* o *parcial*) i adquireixen significació especialitzada pel seu ús en context, per exemple com a complement de substantius especialitzats com *delicte, falta* o *prova*.

Tampoc en les unitats monolèxiques adjectives trobem cap tret distintiu que les caracteritzi i que sigui útil per aplicar-lo a l'extracció per mitjans informàtics.

2.4 Anàlisi de les unitats nominals polilèxiques que segueixen el patró N SA

El patró més habitual és N A, que representa el 80 % dels patrons N SA. La resta de patrons són, en la major part, extensions d'aquest terme base que generalment es formen per postposició d'un altre adjectiu, N A A, que representa un 10 % dels patrons N SA. El 10 % restant de termes es reparteix entre 60 formes diferents.

Les 1.639 ocurrences amb estructura N A estan formades per 793 adjectius i 571 noms diferents.

Per analitzar el grau d'especialització dels noms d'aquestes unitats he seleccionat els 70 noms més utilitzats (un 12 % del total de noms), que representen més del 43 % de les ocurrences, amb els resultats següents:

- Un 74 % (52) dels noms són especialitzats.
- D'aquests, només 12 (17 %) són termes estrictament jurídics, és a dir, que no són presents amb la mateixa forma en llenguatges d'altres àrees de coneixement (*delicte, contracte, jurisdicció, presó, heretament, fideïcomís, hereu, reglament, expropiació, sanció, tribunal* i *vis-a-vis*).
- La resta té un grau de polisèmia elevat: 25 noms tenen tres o més accepcions diferents de la jurídica (el terme *acció* té fins a 13 significats diferents del jurídic).

Amb el mateix objectiu, he seleccionat 100 adjectius (el 12 % del total), que apareixen en 730 ocurrences (44 % del total), i he observat un menor grau d'especialització: només el 38 % dels adjectius es pot considerar especialitzat.

Els adjectius tenen generalment una funció classificadora; creen subclasses del nom al qual acompanyen. En els casos que l'adjectiu té, a priori, un component valoratiu, el seu valor no és qualificatiu sinó que classifica i subespecifica el substantiu associat i identifica un node (o subnode) conceptual jurídic concret. Així passa amb els adjectius *greu* i *lleu* aplicats als noms *delicte, pena, falta, infracció, injúria* i *imprudència*, tots ells delimitats amb precisió per la legislació penal. És interessant apreciar el fet que en una unitat com *abús sexual* cap dels dos components té un valor especialitzat, que només adquireix el conjunt en representar un concepte definit per la legislació penal. Aquest fet és corrent en dret penal: *activitat subversiva, abús deshonest, accés carnal, adequació social, anomalia psíquica, banda armada, batalla/batussa tumultuària, clau falsa, clonació humana* o *escàndol públic* presenten aquesta característica.

2.5 Anàlisi de les unitats polilèxiques que segueixen el patró N SP

Els sintagmes preposicionals es formen a partir de fins a 11 preposicions diferents i són els que presenten més varietat de formes (128 sobre un total de 212 formes diferents detectades). La preposició *de* és la que concentra la major part de les ocurrences (86 %). Aquest tret no és exclusiu de l'àmbit jurídic perquè la preposició preponderant de les estructures N SP és *de* en tots els llenguatges d'especialitat.

El patró amb SP regit per *de* no presenta regularitats remarcables. La forma més corrent és N de N. Les formes més extenses a partir d'aquesta poden ser hipònims com: *cèdula de citació a termini*, a partir de la forma reduïda *cèdula de citació*, o les formes *bé de domini públic accidental, bé de domini públic artificial, bé de domini públic necessari*, a partir de la forma reduïda *bé de domini públic*. Però el més corrent és que les formes més complexes no siguin extensions de formes més simples com, per exemple, *llibre de peces de convicció, motivació d'un acte administratiu, obligació de servei públic* o *peça de situació personal*.

Com assenyala Estopà (1996), es comprova que el nucli del patró N de N és el primer nom (per exemple: *acció d'ofici*). En canvi, en el patró N de art N, el nucli és el segon nom (per exemple: *accessorietat de la participació*).

En general, el sintagma preposicional té una funció classificatòria i crea una subclasse del nom que complementa.

3 Extracció automatitzada de terminologia sobre un corpus textual

L'extracció automatitzada s'ha fet sobre un corpus textual format per tots els DOGC de l'any 2013, que té una mida de 2.262.547 paraules. Aquest corpus ha estat prèviament etiquetat amb l'analitzador FreeLing.

Hem fet l'extracció automatitzada amb TBXTools, una eina creada amb el llenguatge Python que permet dur a terme tasques relacionades amb l'extracció de terminologia i altres utilitats de gestió de la terminologia. Aquesta eina permet l'aplicació de mètodes estadístics i lingüístics (TBXTools.py, copyright 2014, d'Antoni Oliver, està disponible a <https://sourceforge.net/projects/tbxtools>).

Per a l'extracció de tipus lingüístic, el programa cerca patrons morfosintàctics determinats per l'usuari sobre un corpus etiquetat morfosintàcticament.

Aquest programa té altres utilitats que permeten l'aprenentatge de patrons d'una llista de termes de referència i la detecció d'equivalents de traducció en corpus paral·lels.

3.1 Comprovació de patrons

Amb el mètode d'aprenentatge automàtic de patrons, observarem si es poden obtenir patrons morfosintàctics nous d'aquest àmbit d'especialitat i verificarem els patrons obtinguts amb l'anàlisi manual.

Apliquem aquest mètode sobre el corpus legislatiu DOGC 2013, prenent com a llista de termes de referència totes les unitats polilèxiques de Justiterm excepte les que contenen termes en llatí. En total, 3.653 unitats.

El resultat obtingut és una llista de 163 patrons apresos, expressats amb les etiquetes completes assignades per FreeLing (per exemple: NP00000 SPS00 NCMS000 SPS00 NCMS000) en el procés d'etiquetatge del corpus.

Si es classifiquen aquests patrons amb les etiquetes generals utilitzades en l'estudi del corpus terminogràfic Justiterm, els patrons apresos són 26, i es comprova que segueixen una distribució molt semblant a l'obtinguda en l'anàlisi manual i que no hi ha patrons nous.

N A	(386)	56 %
N prep N	(177)	26 %
N A A	(30)	4 %
N prep N A	(26)	4 %
N prep art N	(14)	2 %
21 patrons restants	(53)	8 %
		100 %

TAULA 3. Patrons apresos amb aprenentatge automàtic

En aplicar aquest procés, a més d'identificar els patrons que segueix el corpus terminogràfic estudiat, també hem detectat automàticament que 686 dels 3.653 termes del Justiterm són presents en el corpus legislatiu DOGC 2013.

En l'aplicació pràctica d'aquest procés s'han d'ajustar algunes qüestions relacionades amb l'etiquetatge com, per exemple:

- En els termes de referència que continguin algun apòstrof, s'ha de deixar un espai entre l'apòstrof

i la primera lletra de la paraula següent perquè sigui adequadament identificat.

- FreeLing assigna la categoria de nom propi a tot nom començat per majúscula. Per tant, quan un terme té algun component que apareix en el corpus en minúscula i en majúscula, se li assignen etiquetes diferents.
- Hi ha termes amb la mateixa grafia als quals s'assigna més d'una etiqueta, de les quals, generalment, només una és correcta.
- Hi ha molts casos en què cal canviar la classificació del patró per problemes d'etiquetatge: assignació de l'etiqueta de verb a participis verbals amb funcions d'adjectiu, noms comuns etiquetats com a verb incorrectament o adjectius etiquetats solament com a nom.

3.2 Extracció automatitzada de termes

Per fer l'extracció de termes amb TBXTools, fem servir les etiquetes de la taula 4 per expressar els patrons en el programa.

Patró morfosintàctic	Etiqueta
N A	NC.* AQ.*
N A A	NC.* AQ.* AQ.*
N de N	[NC.*] /de/ NC.*
N de N A	[NC.*] /de/ NC.* AQ.*
N de art N	[NC.*] /de/ DA.* NC.*

TAULA 4. Etiquetes per expressar els patrons en el programa

Les etiquetes aplicables al català es poden trobar a <https://talp-upc.gitbooks.io/freeling-user-manual/content/tagsets/tagset-ca.html>.

Amb els claudàtors [] s'agrupen les formes singulars i plurals i el terme detectat s'expressa amb el lema. En aquest cas, solament apliquem aquesta utilitat als noms seguits de preposició. Si l'apliquem al nom seguit d'adjectiu ens dona resultats poc reeixits com *basa reguladores per bases reguladores o finança públiques per finances públiques*. Tampoc l'apliquem als adjectius perquè en molts casos deixen de concordar amb el nom, per exemple *partida pressupostari o persona beneficiari*.

Els resultats obtinguts en l'extracció són les que es mostren a la taula 5.

Patró morfosintàctic	Candidats	Ocurrences
N A	36.233	241.262
N A A	4.586	17.350
N de N	21.187	97.650
N de N A	6.871	21.066
N de art N	30.236	94.367

TAULA 5. Resultats de l'extracció

No hem fet extracció d'unitats monolèxiques perquè, com hem vist en l'anàlisi del corpus terminogràfic, no s'ha trobat cap criteri que distingeixi les unitats especialitzades de les comunes.

Arribats a aquest punt, després d'haver aplicat tècniques d'extracció de tipus lingüístic, resultaria útil implementar una aproximació estadística que capturess automàticament la noció de *termhood*, és a dir que detectés quins dels candidats obtinguts són realment termes. Això queda fora de l'objecte d'aquest treball, que es limita a l'estudi dels patrons morfosintàctics útils en l'extracció.

3.3 Cerca d'equivalents de traducció

La cerca d'equivalents de traducció permet crear glosaris terminològics bilingües d'una manera molt ràpida. Per fer la cerca d'equivalents s'han de tenir dos elements: un corpus paral·lel en les diferents llengües i una selecció de termes a cercar.

En aquest cas tenim el corpus paral·lel català-castellà del DOGC de 2013 i els candidats obtinguts en l'extracció de termes sobre el DOGC 2013 en català descrita en l'apartat anterior. Per obtenir una selecció de termes sobre els quals cercar l'equivalent de traducció hem seguit el següent procés amb el programa TBX-Tools:

- Hem seleccionat els candidats que tenen com a mínim un element clarament especialitzat (per exemple, *dret, jurídic, judicial, processal*, etc.).
- D'aquests candidats hem eliminat els que tenen una freqüència inferior a 5 perquè amb freqüències petites és més difícil obtenir l'equivalent.
- Hem canviat *de el* per *del* en els candidats amb patró N de art N perquè la contracció s'ha desfet en el procés d'extracció.

Els candidats seleccionats com a termes obtinguts amb l'aplicació d'aquests criteris són 880 i serviran de base per fer la detecció d'equivalents de traducció.

Hem indicat al cercador que retorni, per cada terme, els 6 candidats de traducció que apareguin amb més freqüència. Amb aquesta restricció, el programa ha trobat equivalents de traducció correctes de 708 termes (80,45 %) del total de 880. Si haguéssim demanat més candidats hauríem trobat alguns equivalents de traducció vàlids més, però la major part de resultats bons es concentra en els primers candidats.

En el procés de cerca d'equivalents es detecta una sèrie de problemes que s'han d'anar solucionant. Alguns dels problemes són producte de les inconsistències en els criteris tipogràfics del corpus:

- Problemes per reconèixer el punt volat.
- Ús de diferents caràcters per representar l'apòstrof.
- Resulta problemàtica l'aparició de símbols que fan que el tokenitzador funcioni malament. Per exemple, l'ús de cometes o altres signes: «*legislació processal*».
- També s'ha donat la circumstància que el tokenitzador separava d'una manera incorrecta l'apòstrof, per exemple: l' *assignació tributària*.

4 Conclusions

Les conclusions d'aquest estudi que poden ser útils per establir estratègies lingüístiques en l'extracció automatitzada de terminologia jurídica són les següents:

- La categoria gramatical bàsica de la terminologia jurídica és la nominal.
- Les estructures morfològiques de les unitats polilèxiques presenten una gran varietat de formes. Els patrons més productius són N SA i N SP. Les formes particulars més freqüents són N A i N de N.
- Les unitats monolèxiques nominals i adjectivals no presenten cap tret distintiu que les caracteritzi d'unitats de llengua general i que sigui útil per a l'extracció informàtica.

La tasca d'identificació d'unitats terminològiques en un corpus textual resulta enormement beneficiada de l'aplicació de processos informàtics. Es facilita molt la feina de l'especialista perquè aconsegueix un nombre de candidats relativament reduït. L'aplicació d'estratègies lingüístiques en el procés informàtic d'extracció millora els resultats i, en particular, resulta molt eficient la cerca basada en patrons morfosintàctics. Malgrat tot, els resultats són manifestament millorables perquè s'obté un gran nombre de candidats que no són termes i també hi ha un gran nombre de termes no detectats, com els monolèxics. Els processos de millora de resultats preveuen diferents aspectes, com aplicar instruments estadístics més elaborats que la simple freqüència d'aparició i tenir en compte aspectes semàntics i pragmàtics. ✿

Bibliografia

- CABRÉ, M. Teresa (1999). «Hacia una teoría comunicativa de la terminología: aspectos metodológicos». *Revista Argentina de Lingüística*, 11.
- DAILLE, Beatrice (1994). «Combined approach for terminology extraction: lexical statistics and linguistic filtering». *UCREL Technical Papers*, 5. Universitat de Lancaster.
- ESTOPÀ, Rosa (1996). *Les unitats terminològiques polilexemàtiques en els lèxics especialitzats: dret i medicina*. Treball de recerca de doctorat. Barcelona: IULA.
- PAZIENZA, M. Teresa; PENNACCHIOTTI, Marco; ZANZOTTO, Fabio Massimo (2005). «Terminology extraction: An analysis of linguistic and statistical approaches». A: SIRMAKESSIS, Spiros (ed.). *Knowledge mining*. Springer Verlag, p. 255-279. (*Studies in fuzziness and soft computing*; 185).