L'anàlisi lexicomètrica dels textos especialitzats: un terreny d'intersecció entre terminologia, documentació i traducció*

Lluís de Yzaguirre, Carles Tebé, Araceli Alonso, Rosa-Ana Folguerà
Institut Universitari de Lingüística Aplicada
Universitat Pompeu Fabra

1. Introducció i objectius

Els autors d'aquesta comunicació partim del supòsit que terminòlegs, documentalistes i traductors realitzen habitualment tasques que tenen molts punts en comú, ja que totes tres disciplines tenen un mateix entorn natural de treball, que són els llenguatges especialitzats, i analitzen unes mateixes manifestacions, que són els textos especialitzats. En contrast, uns i altres s'acosten als textos i hi treballen amb estratègies, recursos i eines diferents, per tal com els resultats aplicats que han de produir són també diferents.

Aquestes similaritats i divergències es poden il·lustrar si examinem l'enfocament amb què cada col·lectiu aborda els textos. Els terminòlegs se serveixen dels textos especialitzats per detectar-ne i extreure'n els termes (el procés que s'anomena *buidatge terminològic*); els documentalistes s'enfronten amb els textos especialitzats amb la finalitat de descriure'n el contingut per mitjà de descriptors o paraules-clau (*l'anàlisi documental*); i els traductors descomponen els textos de partida en unitats de traducció, a fi de poder-les transvasar en una altra llengua respectant al màxim totes les característiques del text original (*l'anàlisi del text de partida*).[1]

El propòsit d'aquesta comunicació és demostrar que l'anàlisi del text de partida amb eines lexicomètriques és una estratègia de treball que ajuda a determinar el pes o gradient temàtic que presenten les unitats lèxiques d'un text especialitzat, i que aquesta informació, convenientment elaborada i tractada, ofereix elements que considerem valuosos per als diversos enfocaments que hem considerat:

a) Per als terminòlegs, creiem que és una eina útil que pot servir per determinar el grau d'especialització d'un text, alhora que ajudarà a establir la representativitat temàtica d'un corpus de buidatge.

---

* Comunicació presentada a la I Jornada de Terminologia i Documentació, celebrada el 14 de maig del 2000 a la Universitat Pompeu Fabra.

[1] Per a una descripció de teories, mètodes i aplicacions de cada matèria, vegeu tres obres clàssiques: en terminologia, Cabré (1992); en documentació, Lancaster (1995); i en traducció, Newmark (1992).

b) Per als traductors, pot ser important per valorar la competència cognitiva necessària per traduir el text, així com per determinar els recursos lexicogràfics i documentals que caldrà utilitzar per resoldre els problemes d'equivalència que es plantejaran en la traducció.

c) Per als documentalistes, aquesta anàlisi temàtica aplicada a un únic document permet obtenir un conjunt d'indicadors dels termes més rellevants d'aquest document, així com un conjunt de termes candidats a descriptors. Aplicat a un conjunt de documents del mateix domini temàtic, pot proporcionar un llistat inicial de termes per a construir un tesaurus.

2. Textos especialitzats i unitats d'anàlisi

D'acord amb Cabré (1999) considerem que «La terminologia és el factor privilegiat, tot i que no pas l'únic, de representació del coneixement especialitzat. Una de les característiques lingüístiques més destacables dels textos cientificotècnics és la presència d'unitats específiques d'un àmbit determinat». En aquest treball hem partit del supòsit que terminòlegs, documentalistes i traductors treballen sobre uns mateixos materials i sobre unes mateixes unitats de base, que en la introducció hem anomenat *termes*, *descriptors* i *unitats de traducció* segons la disciplina a què fèiem referència.

Però l'afirmació que es tracta d'unes mateixes unitats de base s'ha de matisar. Com han notat diversos autors, tota disciplina que intenta consolidar-se al costat d'altres matèries que ja es troben legitimades socialment, acadèmicament o professionalment, intenta buscar uns fonaments epistemològics propis per a la seva matèria. En síntesi, aquesta fonamentació consisteix a delimitar un *espai o àmbit de treball propi*, definir un *objecte o perspectiva d'estudi original*, i descriure una *unitat d'anàlisi* també pròpia i diferent. Cada disciplina mira aleshores de definir i caracteritzar els límits de la seva *unitat de base* de manera distinta a les unitats ja reconegudes per altres disciplines, remarcant per sobre de tot suposades diferències essencials, en comptes de remodelar, redefinir i reutilitzar unitats que estan molt ben descrites i avalades per una llarga tradició científica, i que pel seu caràcter polimòrfic acceptarien de ser analitzades des d'altres perspectives.[2]

En aquest sentit, considerem que la diversitat de denominacions a l'entorn de les unitats que es troben en els textos especialitzats s'explica, d'una banda, per tradicions

[2] Aquesta és la raó de fons per la qual molts teòrics de la terminologia han destinat grans esforços a descriure el *terme* (i el *concepte*) com una unitat clarament diferenciada de la *paraula (*i el seu *significat)*. Per a una anàlisi en profunditat de les conseqüències que ha tingut en el pla científic un procés de legitimació d'aquestes característiques per a la disciplina de la terminologia, vegeu Cabré (1999).

acadèmiques i enquadraments disciplinars diferents, i de l'altra, per necessitats aplicades (elaboració de diccionaris, construcció de tesaures, preparació de traduccions) que certament presenten característiques i metodologies de treball específiques, però que sobretot reflecteixen punts de vista diferents sobre un mateix objecte. Així, el *terme* és descrit sobretot a partir de supòsits semàntics i cognitius; el tret distintiu del *descriptor* es troba en el seu potencial per a identificar conceptes d'una banda, i d'etiquetar-los de manera unívoca per altre; i *la unitat de traducció* gira a l'entorn de la idea d'equivalència interlingüística. Sovint les seves fronteres i denominacions coincideixen, però no sempre.[3]

Abans de proposar una denominació comuna que englobi totes aquestes unitats, caracteritzarem globalment les nostres unitats d'anàlisi:

a) Les unitats que tractem són semànticament específiques, és a dir, vehiculen un significat especialitzat en el text o discurs en què es produeixen; aquest tret és compartit per totes les unitats.

b) Les unitats que tractem són generalment unitats lèxiques, per bé que presenten una gran varietat de processos de formació; poden ser monolexemàtiques o plurilexemàtiques, i entre aquestes s'hi troben gran varietat d'estructures sintagmàtiques.[4]

c) Les unitats que tractem presenten graus de fixació en la llengua molt desiguals, de manera que no sempre és fàcil delimitar el segment que les representa. En la pròpia terminologia aquest és un problema descrit i estudiat, específic de les unitats terminològiques polilexemàtiques i de les unitats fraseològiques. En documentació i en traducció, les mateixes denominacions de *descriptor* i d'*unitat de traducció* també vehiculen aquesta problemàtica.

Amb la intenció d'utilitzar una denominació que, més enllà dels límits tradicionals del terme, abraci una concepció més àmplia de les unitats especialitzades del text, utilitzarem el concepte d'Unitat de Significació Especialitzada. En paraules d'Estopà (1999: 286), «la unitat que és objecte d'estudi de la terminologia no pot reduir-se a la unitat terminològica, sinó que ha d'abastar totes les unitats que anomenem Unitats

[3] Val a dir que hi ha una denominació comuna que s'utilitza regularment en els textos de les tres matèries, i és la de *terme*, que conviu amb les altres esmentades més amunt. Ara bé, no sempre és utilitzada amb el mateix valor, sinó que de vegades s'emfasitzen els trets específics de cada enfocament que hem comentat.

[4] En els textos especialitzats també podem trobar conceptes vehiculats per mitjans no lingüístics, caracteritzats per la presència d'elements procedents d'altres sistemes de notació: icònics, numèrics, gràfics... o bé els sistemes de nomenclatura (com els de la química), que són en realitat llenguatges artificials, establerts per consens internacional, però construïts manllevant algunes característiques del llenguatge natural. En el nostre treball hem descartat aquest tipus d'unitats, però pensem que podrien ser tractades amb els mateixos principis metodològics que hem seguit.

de Significació Especialitzada (USE) que inclouen tant les unitats especialitzades de categories gramaticals diferents que formen part del llenguatge natural, com les unitats que formen part de llenguatges artificials: i dins de les unitats que són llenguatge natural, abraça des de les unitats terminològiques simples a les complexes, des dels noms als verbs, adjectius i adverbis, des de les unitats lèxiques a les unitats fraseològiques especialitzades».

Per bé que en aquest treball analitzem sobretot les USE més prototípiques, que són les unitats terminològiques de categoria nominal, d'ara endavant utilitzarem de manera preferent la denominació d'USE en lloc de la de termes.

Hipòtesis de partida, metodologia i treball experimental

La nostra hipòtesi de partida és que en l'anàlisi de la distribució temàtica de les USE d'un text especialitzat podien donar-se tres situacions diferents:

a) En l'anàlisi del text predomina una única àrea temàtica ben delimitada: és quan parlem de *consistència temàtica.*

b) El text conté USE que estan associades a dues o més àrees temàtiques, però cap d'aquestes àrees és predominant ni es concentra clarament en una part del text, de manera que es produeix una situació que qualifiquem d'*intersecció temàtica*. Subratllem que en aquest cas ens trobem davant d'un text especialitzat com els altres, però inespecífic des d'un punt de vista temàtic.

c) El text conté USE que pertanyen a dues o més àrees temàtiques clarament diferenciades en el text: aleshores considerem que es planteja una situació de *complementarietat temàtica*.

En l'apartat metodològic, hem pres les següents decisions:

Per al corpus d'anàlisi, hem seleccionat textos disponibles en suport electrònic que per la seva naturalesa podien presentar una situació de variació temàtica, és a dir d'*intersecció* o de *complementarietat* temàtiques. El segon criteri utilitzat ha estat de seleccionar textos que, en la mesura del possible, presentessin versions paral·leles en altres llengües, a fi de poder verificar posteriorment si els resultats eren generalitzables quantitativament en altres sistemes lingüístics, o per contra hi havia variacions significatives. El text pilot seleccionat ha estat un text de legislació sobre medi ambient: *Conveni Marc de les Nacions Unides sobre el Canvi Climàtic*.[5]

[5] L'adreça on es troba el text complet en català és: http://www.gencat.es/mediamb/sosten/cnucc.htm. El mateix text es pot trobar en altres idiomes a les següents adreces:
castellà: http://www.gencat.es/mediamb/cast/sosten/enucc.htm
francès: http://www.un.org/french/ecosocdev/geninfo/environ/climcon.htm
anglès: http://www.unfccc.de/resource/conv/index.html

El tractament lexicomètric del text s'ha fet amb el programa d'anàlisi textual TACT, desenvolupat al Center for Computing and Humanities de la Universitat de Toronto.[6] Aquest programa ha estat utilitzat per dues raons:

1) Com qualsevol programa de tractament lexicomètric, el TACT és capaç de processar un o diversos textos, crear una base de dades textual que pot ser interrogada per l'usuari, i extreure'n diversos resultats en forma de llistats d'unitats (de paraules simples o de concurrències[7]) que poden ser ordenats de diverses maneres: alfabèticament, per ordre de freqüència ascendent o descendent, etc. A més, cada unitat o grup d'unitats seleccionades pot ser analitzada en el seu context lingüístic extraient les concordances pertinents.

2) En segon lloc, i molt important, el programa ens permet veure la distribució de cadascun dels termes escollits, bé individualment o bé agrupats per àrees temàtiques (en aquest cas, dret o medi ambient) mitjançant unes gràfiques, les quals ens indiquen clarament el pes o gradient temàtic que presenten aquestes unitats lèxiques *al llarg del text*. D'aquesta manera, podem analitzar no solament la *densitat temàtica* d'un text sinó la seva distribució en un corpus determinat, i en el nostre treball ens servirà per analitzar la *topografia temàtica* dels textos analitzats.

Finalment, l'atribució d'una marca d'àrea temàtica a les unitats seleccionades s'ha fet contrastant-les amb un corpus de referència, que ha estat el banc de dades terminològiques *Termium*. El procés realitzat en aquest treball ha estat manual; tanmateix, considerem que si els resultats d'aquest treball condueixen a repetir l'experiment amb altres textos, la fase d'atribució temàtica a les unitats és automatitzable amb l'ajuda de corpus de contrast en suport electrònic com *l'Hiperdiccionari*.

En l'elaboració del procés hem seguit les fases següents:[8]

a) Processament del text escollit amb el TACT, completat en totes les fases, fins a generar la base de dades textual que conté totes les unitats del text.

b) Extracció de paraules simples i de concurrències, i selecció de les USE més freqüents a partir de les llistes inicials facilitades pel programa.

---

[6] Podeu trobar referències sobre el programa TACT a les adreces següents: http://www.chass.utoronto.ca:8080/cch/tact.html; http://www.indiana.edu/~letrs/help-services/QuickGuides/about-tact.html; http://tactweb.humanities.mcmaster.ca/tactweb/doc/tact.htm.

[7] És com proposem d'anomenar les *collocations*. Per a una descripció de les concurrències i el seu tractament en la constitució de corpus lingüístics, vegeu, entre d'altres, Sinclair (1981).

[8] Per raons d'espai no ens ha estat possible d'il·lustrar amb detall tot el procés seguit en el treball. Per això hem preparat un text complementari que descriu i il·lustra exhaustivament tota la fase de processament i tractament de les dades. Podeu trobar-lo a la pàgina web http://terminotica.upf.es/topografia/.

c) Assignació d'àrea temàtica a les USE seleccionades; l'atribució es fa contra un corpus de referència extern.[9]

d) Creació de grups d'unitats a la base de dades textual segons l'àrea temàtica atribuïda, a fi de poder analitzar la distribució i el pes de cada grup d'USE en el text.

e) Generació dels resultats: freqüència i distribució de cada unitat en el text, pes de cada àrea temàtica en el text, distribució de cada àrea temàtica al llarg del text.

Anàlisi dels resultats i perspectives de treball

Presentem a continuació una mostra dels resultats obtinguts.

Resultat 1: Llista d'USE simples i complexes extretes amb el TACT.

Freqüencia	**Paraules simples**
189	parts
94	conveni
72	article
70	conferència
45	acord
29	present
28	aplicació
26	paràgraf
25	informació
22	objectiu

Freqüencia	**Paraules simples**
21	secretaria
19	disposicions
19	sessions
19	virtut
18	annex
18	compte
18	dipositari
18	vigor
17	resta
17	zones

Fig. 1: Relació d'USE simples més freqüents per ordre decreixent.

[9] El corpus de referència utilitzat en aquest treball ha estat el banc de dades terminològiques Termium en CD-ROM (versió de 1997). Atès que en aquest experiment només buscàvem àrees temàtiques de primer nivell (la branca o etiqueta temàtica més general), i n'hem trobat dues, considerem que els resultats serien previsiblement semblants amb altres corpus de referència. Aquesta hipòtesi, però, no l'hem confirmat.

Fig. 2: Relació d'USE complexes més freqüents per ordre decreixent.

Denominador	Freqüència i concurrència
#5	46 canvi climàtic
#15	34 efecte de hivernacle
#21	28 països en desenvolupament
#14	24 gasos amb efecte d hivernacle
#7	15 nacions unides
#5	13 període de sessions
#7	12 comissions antropogèniques
#18	12 països desenvolupats
#5	11 protocol de mont-real
#1	10 integració econòmica
#10	9 mesures adoptades
#5	8 òrgans subsidiaris
#5	8 estats membres
#4	8 sistema climàtic
#4	8 medi ambient

Resultat 2: Gràfica de distribució d'una USE complexa al llarg del text:

```
(2)                                                    Total: 46.

0-5%    |    4|XXXX          50-55%  |    2|XX
5-10%   |    4|XXXX          55-60%  |    0|
10-15%  |    4|XXXX          60-65%  |    4|XXXX
15-20%  |    6|XXXXXX        65-70%  |    1|X
20-25%  |    7|XXXXXXX       70-75%  |    0|
25-30%  |    5|XXXXX         75-80%  |    0|
30-35%  |    2|XX            80-85%  |    0|
35-40%  |    1|X             85-90%  |    0|
40-45%  |    2|XX            90-95%  |    0|
45-50%  |    4|XXXX          95-100% |    0|
```

Fig. 3: Distribució de la USE **canvi climàtic** al llarg del text.

Resultat 3: Pes i distribució de les USE de cada àrea temàtica dins el text.

```
(32)                                                  Total: 117.
0-5%    |    9|XXXXXXXXX              50-55%  |    3|XXX
5-10%   |   11|XXXXXXXXXXX             55-60%  |    0|
10-15%  |   13|XXXXXXXXXXXXX           60-65%  |    4|XXXX
15-20%  |   15|XXXXXXXXXXXXXXX         65-70%  |    2|XX
20-25%  |   16|XXXXXXXXXXXXXXXX        70-75%  |    4|XXXX
25-30%  |   13|XXXXXXXXXXXXX           75-80%  |    0|
30-35%  |    7|XXXXXXX                 80-85%  |    0|
35-40%  |    3|XXX                     85-90%  |    0|
40-45%  |   12|XXXXXXXXXXXXX           90-95%  |    0|
45-50%  |    5|XXXXX                   95-100% |    0|
```

Fig. 4: Distribució de l'àrea temàtica **medi ambient** al llarg del text.

```
(24)                                                  Total: 421.
0-5%    |    5|XXXXX
5-10%   |    3|XXX
10-15%  |    5|XXXXX
15-20%  |    9|XXXXXXXXX
20-25%  |    5|XXXXX
25-30%  |    6|XXXXXX
30-35%  |   19|XXXXXXXXXXXXXXXXXX
35-40%  |   15|XXXXXXXXXXXXXXX
40-45%  |   11|XXXXXXXXXXX
45-50%  |   15|XXXXXXXXXXXXXXX
50-55%  |   27|XXXXXXXXXXXXXXXXXXXXXXXXXXX
55-60%  |   34|XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
60-65%  |   28|XXXXXXXXXXXXXXXXXXXXXXXXXXXX
65-70%  |   27|XXXXXXXXXXXXXXXXXXXXXXXXXXX
70-75%  |   21|XXXXXXXXXXXXXXXXXXXX
75-80%  |   40|XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
80-85%  |   27|XXXXXXXXXXXXXXXXXXXXXXXXXXX
85-90%  |   51|XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
90-95%  |   35|XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
95-100% |   38|XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
```

Fig. 5: Distribució de l'àrea temàtica **dret** al llarg del text.

En l'anàlisi de resultats, volem subratllar les conclusions principals a què hem arribat:

a) Després d'analitzar les USE simples i complexes més freqüents en el text, individualment i en grup, observem que hi ha dues àrees temàtiques clarament predominants, i que les unitats de l'àrea de dret tenen un pes temàtic més important que les unitats de medi ambient.

b) Pel que fa al *pes o gradient temàtic* de les unitats del text, considerem que l'anàlisi quantitativa de les dades demostra clarament una situació de *complementarietat temàtica* entre les unitats de medi ambient i les de dret: les unitats de totes dues àrees són clarament predominants en el text.

c) L'anàlisi distribucional de les dades demostra que les USE de l'àrea de dret es reparteixen majoritàriament des del 40% del text fins al final, mentre que les USE de medi ambient són clarament predominants en la primera meitat del text, i per contra desapareixen pràcticament en la segona meitat.

d) Així, pel que fa a la *topografia temàtica* del text, considerem que els resultats descriuen igualment una situació de distribució complementària, atès que les USE d'una àrea i altra tenen un pes diferent al llarg del text.

Des del punt de vista metodològic, considerem que els resultats obtinguts proven que el mètode i el tractament de les dades que hem seguit permeten extreure conclusions rellevants del text analitzat, i validar o refutar les hipòtesis formulades abans de la fase d'anàlisi i processament de les dades. Tanmateix, som conscients que metodològicament som en un punt de partida, no d'arribada, i que en treballs posteriors caldrà afinar decisions, automatitzar algunes fases del procés, i millorar la presentació gràfica de les dades, a fi que la lectura dels resultats obtinguts pugui ser encara més matisada. Amb aquesta intenció, presentem en annex una proposta que hem anomenat *topografia temàtico-terminològica*, concebuda com un mapa topogràfic del text estudiat que cartografia la distribució simultània de diverses USE al llarg del text.[10]

Finalment, a tall de consideració global, ens refermem en el fet que els primers resultats que hem obtingut obren perspectives interessants per la realització de treballs d'interès comú per a terminòlegs, documentalistes i traductors, com apuntàvem en els objectius de la comunicació:

—Per a la terminologia, la nostra aportació pensem que s'insereix en la línia de treballs que actualment analitzen el funcionament de les USE en els textos especialitzats (com els treballs sobre el concepte de densitat terminològica del text), i permetrà caracteritzar més bé un tipus de variació horitzontal poc estudiada, que és la varietat temàtica o diatòpica en l'interior dels textos especialitzats.

—Per a la documentació, pensem que les eines i processos utilitzats en aquest treball poden ser una aportació als procediments d'indexació automàtica de documents en text complet, en la variació anomenada "text lliure" (*free text*); a més podria pro-

[10] La topografia és acompanyada d'una breu llegenda que ajuda a interpretar la gràfica, que ha estat generada amb el full de càlcul Excel. Per a una descripció exhaustiva del procés seguit en la creació de la topografia, vegeu-ne el text a la pàgina web http://terminotica.upf.es/topografia/excel.htm.

porcionar indicacions sobre termes i candidats a descriptors en la indexació amb llenguatges controlats (tesaurus), i, també, podria utilitzar-se en programes de categorització automàtica de documents.

—Per a la traducció, pensem especialment que les noves eines de traducció assistida basades en les memòries de traducció, que s'estan implantant amb força, conduiran a la necessitat de caracteritzar, descriure i etiquetar amb precisió el contingut dels textos que seran utilitzats com a material de referència, i que han de servir per a la pretraducció controlada de nous textos. Igualment, considerem que en empreses de traducció amb un volum important de treball, la caracterització temàtica del text a traduir ajudarà a seleccionar el traductor més adequat per a cada cas.[11]

Bibliografia

Cabré, M.T. (1992). *La terminologia: la teoria, el mètode, les aplicacions*. Barcelona: Empúries.

Cabré, M.T. (1999). *La teoria comunicativa de la terminologia*. Barcelona: IULA, Universitat Pompeu Fabra.

Domènech, M. (1998). *Unitats de Coneixement i textos especialitzats: primera proposta d'anàlisi*. Barcelona: IULA, Universitat Pompeu Fabra. [treball de recerca de doctorat]

Estopà, R. (1999). *Extracció de terminologia:elements per a la construcció d'un SEACUSE (Sistema d'Extracció Automàtica de Candidats a Unitats de Significació Especialitzada)*. Barcelona: IULA, Universitat Pompeu Fabra. [tesi doctoral]

Lancaster, F. W. (1995). *El control del vocabulario en la recuperación de la información*. València: Universitat de València.

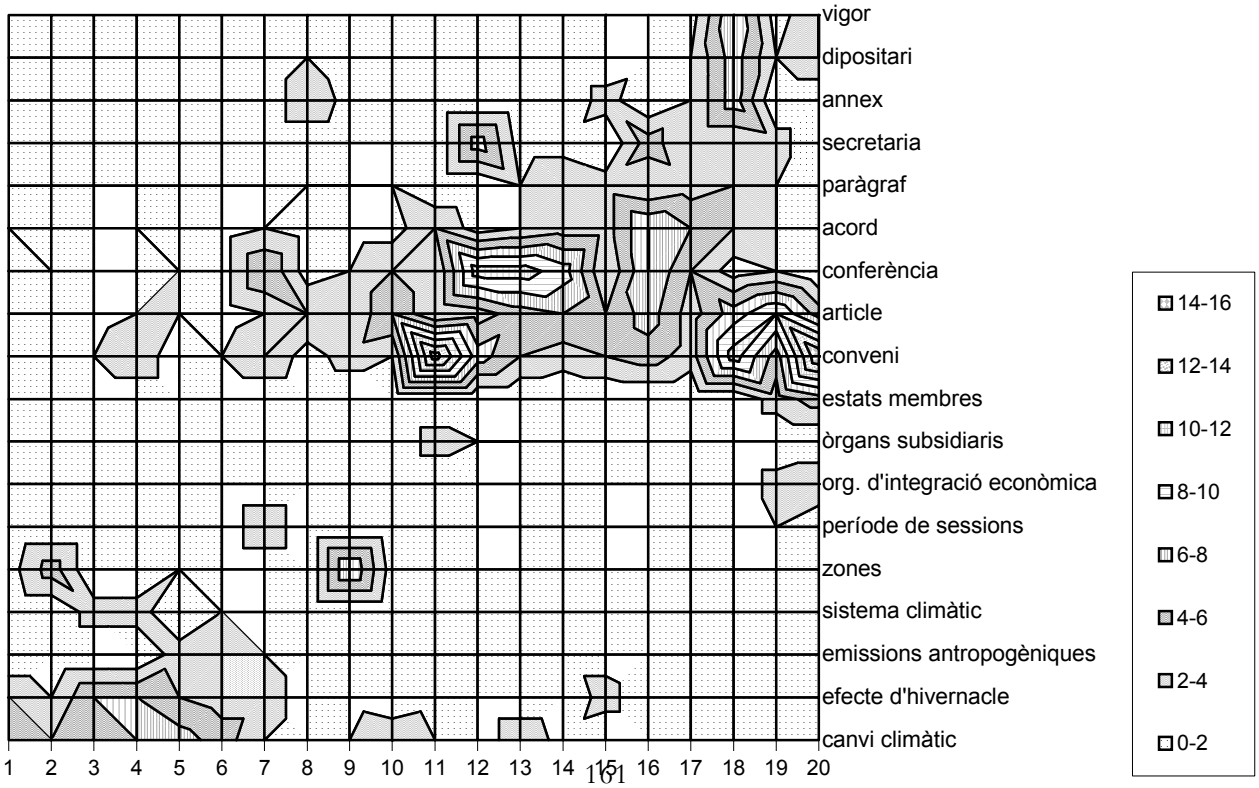
Newmark, P. (1992) *Manual de traducción*. Madrid: Cátedra.

Sinclair, J. (1981) *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Annex: Topografia temàtico-terminològica

El gràfic s'ha d'interpretar d'esquerra a dreta. La coordenada horitzontal indica les 20 parts en les quals s'ha dividit el text. Cadascuna d'aquestes parts representa un 5% del text. En la coordenada vertical apareixen els 18 termes més rellevants i la seva distribució en les 20 parts en què està dividit el text. Cada color representa el nombre d'ocurrències de cada terme en una part determinada del text, per exemple, *conveni* es distribueix des de la tercera part fins al final del text, concentrant la seva representativitat màxima (14-16) en la onzena part i en la vintena.

[11] Els autors d'aquest treball volem agrair al Dr. Lluís Codina els suggeriments i comentaris valuosos que ens ha fet, i que sens dubte han enriquit el text final.


vigor
dipositari
annex
secretaria
paràgraf
acord
conferència
article
conveni
estats membres
òrgans subsidiaris
org. d'integració econòmica
període de sessions
zones
sistema climàtic
emissions antropogèniques
efecte d'hivernacle
canvi climàtic
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
14-16
12-14
10-12
8-10
6-8
4-6
2-4
0-2