Capítol 6. L'anàlisi del corpus compilat

Al capítol 4 (apartat 2.2 "Descripció del corpus", pàgina 211) s'ha descrit el corpus, compilat segons els criteris de cerca i obtenció de recursos textuals indicats al capítol 3 (apartats 2 "La cerca de documents digitals especialitzats", pàgina 166 i 3 "L'extracció de documents digitals de la xarxa", pàgina 175). Per tot l'indicat en aquests apartats, i tal com s'ha exposat al capítol 4, el corpus que hem recollit és un corpus escrit, no etiquetat, especialitzat (en Leònids), sincrònic, textual, gran, documentat, multilingüe (anglès, castellà i català), comparable i *ad hoc*.

Per estudiar-lo farem servir tres eines informàtiques diferents, que es descriuen en el primer apartat d'aquest capítol. A continuació s'analitzarà l'adequació dels corpus per a la traducció de dos textos (un article de premsa i un article de revista acadèmica). Els dos últims apartats d'aquest capítol estan dedicats a il·lustrar les possibilitats de cerca que es poden dur a terme amb aquesta mena de corpus: d'una banda cerques per obtenir dades de manera sistemàtica (apartat 3 "Anàlisi del corpus multilingüe comparable: cerca sistemàtica", pàgina 307) i de l'altra cerques per obtenir dades puntuals (apartat 4 "Anàlisi del corpus multilingüe comparable: cerques puntuals", pàgina 355).

1. Eines de gestió i anàlisi del corpus

L'estudi del corpus recollit es duu a terme amb el suport de tres programes informàtics diferents:

- El Gestor de Recursos Textuals.
- WordSmith Tools.
- L'Extractor d'*n-grams*.

A continuació descriurem cadascuna de les tres eines utilitzades amb detall.

1.1. El Gestor de Recursos Textuals (GeRT)

El GeRT és un programa no comercialitzat desenvolupat en el marc d'aquesta tesi doctoral[1]. La funció fonamental d'aquest programa consisteix a separar un conjunt dels recursos textuals que formen part del corpus per analitzar-los de manera independent. Aquest programa ens facilitarà, per tant, l'anàlisi de *subcorpus* d'arxius que comparteixin alguna de les característiques descrites a l'apartat 2.2 "Descripció del corpus" del capítol 4 (pàgina 211) com ara la llengua, el tipus d'autor, el tipus de recurs textual o altres.

El programari ha estat desenvolupat en llenguatge Delphi, versió 5.0 (Visual Pascal) per a entorn Windows. Utilitza una base de dades que incorpora taules Paradox 7 i Visual Dbase. La base del programari és una taula Visual Dbase que conté les dades relatives als recursos textuals digitals, amb els camps següents:

- Índex
- Autor
- Títol

[1] Programa dissenyat per l'autora d'aquest treball de recerca i desenvolupat per un tècnic. En la versió en suport digital d'aquesta tesi s'inclou una còpia d'aquest programa.

- Responsable secundari
- Títol document
- Llengua
- Data de creació
- Data de actualització
- Data de consulta
- Lloc
- Tipus (a-e)
- Accés (URL)
- Relació dades-text
- Dades autor
- Dades receptor
- Tipus de recurs
- Tipus de lloc
- Contingut

Aquests camps coincideixen amb els paràmetres previstos per a la descripció dels recursos textuals, tal com queda recollit a la taula 4-2. "Esquema de classificació de recursos textuals digitals especialitzats en Leònids" (pàgina 211).

La taula Visual DBase del GeRT conté la informació que es generarà amb els diferents tipus de consultes que aquesta aplicació permet.

El programari es divideix en quatre fitxes: *consultes*, *selecció*, *estadístiques* i *edició*, que es descriuen a continuació.

1.1.1. La fitxa consultes

Aquesta fitxa permet executar consultes dinàmiques sobre qualsevol dels camps de la taula descrita a l'apartat anterior mitjançant el llenguatge de consulta SQL. Els controls situats a la part superior esquerra de la fitxa (vegeu la figura 6-1, pàgina 286) permeten seleccionar un o més camps al mateix temps, alhora que també permet

determinar si el criteri de consulta ha de ser igual (=) que el criteri seleccionat o diferent (<>). Si la consulta fa referència a més d'un camp, es pot determinar si el nou criteri de consulta és acumulatiu respecte dels anteriors (botó I, criteri SQL AND) o alternatiu (botó O criteri SQL OR).

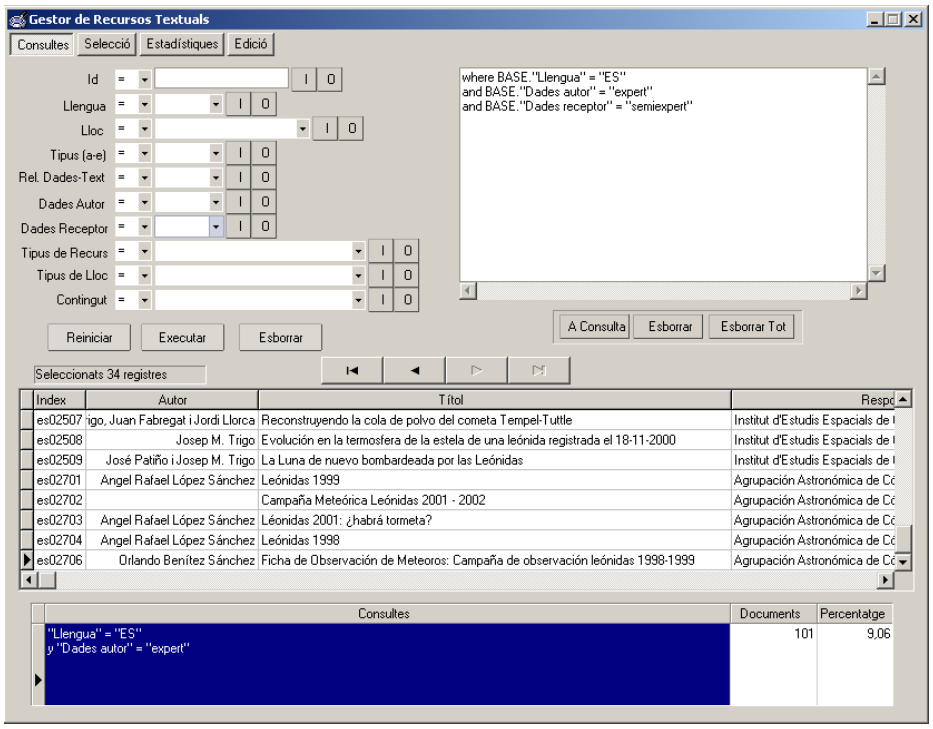


Figura 6-1. Interfície de la fitxa Consultes del GeRT (e. p.)

El valor de la consulta se selecciona de la llista desplegable, en què es presenten tots els valors possibles. Són els botons I i O els que executen la consulta cada vegada que es premen. Això no obstant, també es pot fer la consulta triada prement el botó *Executar*. Si es vol fer una altra consulta, els criteris de consulta anterior s'esborren prement el botó *Esborrar*. En el cas, força improbable, que s'hagués produït un error del sistema en fer la consulta, s'hauria de prémer el botó *Reiniciar.*

A la part superior dreta de la finestra apareix el codi SQL de la consulta: així es pot comprovar quins criteris han estat triats per a la cerca. A la part central de la finestra apareix una graella amb les dades dels recursos textuals filtrats a partir dels criteris de la consulta, i a la part superior d'aquesta graella els controls d'edició de la taula que

ens permetran moure'ns pels diferents registres (com ara, situar-se al principi o al final de la taula).

Per no haver de fer una consulta determinada més d'un cop, encara que és força ràpid, es pot guardar el resultat de la consulta a la graella, situada a la part inferior de la finestra, prement el botó *A consulta*. Així, aquesta segona graella mostrarà la consulta SQL amb els criteris escollits, el nombre de documents que s'han recuperat a partir dels criteris establerts i el percentatge que representen aquests recursos amb relació al total. Per esborrar la consulta activa en aquesta graella s'ha de prémer el botó *Esborrar*, i per buidar la graella completament el botó *Esborrar Tot*.

1.1.2. La fitxa selecció

Aquesta és sense cap mena de dubte la fitxa més complexa de l'aplicació i la més important, ja que permet destriar arxius del corpus a partir de fins a quatre criteris de selecció, els que en cada moment puguin interessar a l'usuari. El criteri fonamental en aquest cas serà el de llengua, ja que d'aquesta manera es poden separar els tres subcorpus monolingües. La resta de paràmetres previstos són:

- Lloc.
- Tipus (a-e).
- Relació dades-text.
- Dades autor.
- Dades receptor.
- Tipus de recurs.
- Tipus de lloc.
- Contingut.

Una vegada triat entre un i quatre criteris de consulta, s'ha de prémer el botó *Executar* i començarà un procés de consultes combinades a l'SQL que presentaran el nombre de documents amb tots els valors del criteri (camp de la taula de documents) triat.

D'aquesta manera, si hem triat com a primer criteri de consulta el camp *Llengua* hi haurà tres consultes SQL amb els tres valors possibles d'aquest camp (anglès, castellà i català). Les dades d'aquest primer criteri de consulta es presenten a la primera de les quatre graelles previstes en aquesta fitxa. En aquesta graella, al camp *Concepte* s'indiquen els valors (registres) del criteri triat, mentre que el camp *Valor* presenta el nombre de documents que coincideixen amb aquest criteri i el camp *Percentatge* indica el percentatge que representa respecte del total.

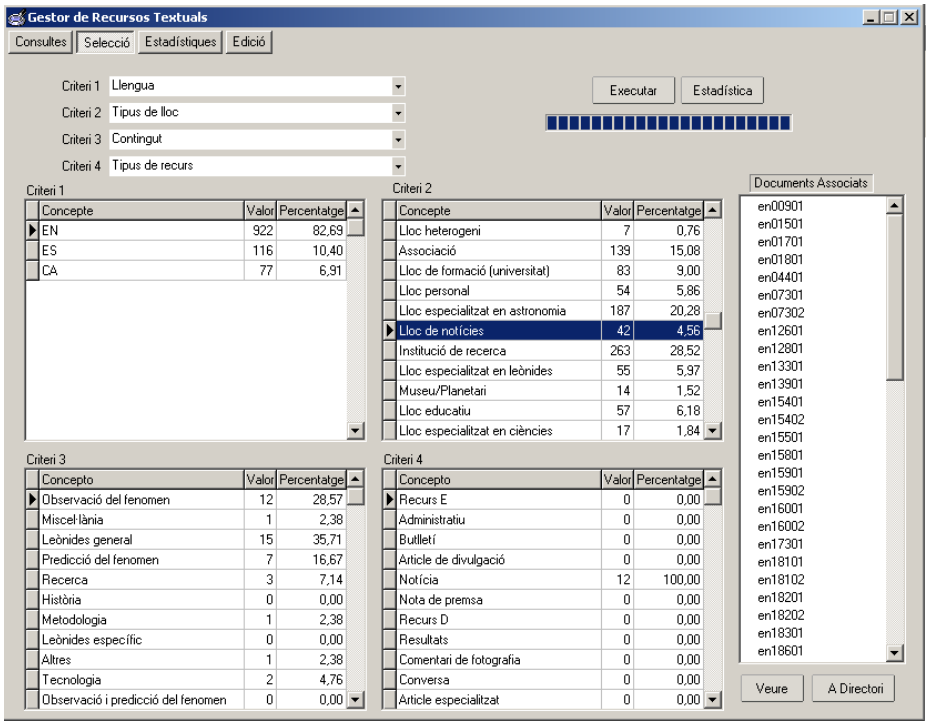


Figura 6-2. Interfície de la fitxa *selecció* del GeRT (e. p.)

Si s'ha triat un segon criteri de consulta, es durà a terme una sèrie més o menys complexa de consultes SQL, combinant els valors del primer criteri amb els valors del segon. Així doncs, continuant amb l'exemple anterior, si el primer criteri de consulta és *Llengua* i el segon *Tipus de Lloc*, es fan 42 (3+39) consultes SQL, ja que hi ha 3 valors possibles pel camp *Llengua* i 13 valors possibles per al camp *Tipus de Lloc*, que cal multiplicar pels 3 valors possibles del primer criteri. Els valors del segon criteri de consulta se situen a la segona graella. Els percentatges recollits en aquesta segona graella no es calculen en funció del total d'arxius del corpus, sinó de la totalitat del

valor escollit al primer criteri. D'aquesta manera, un valor que en aquest nivell signifiqui el 100% no es referirà als 1.115 registres del corpus, sinó a la totalitat dels recursos corresponents al valor seleccionat al criteri anterior.

Si es tria un tercer criteri, la consulta SQL es complica molt, depenent dels valors possibles del nou criteri de selecció. Si a la consulta proposada anteriorment s'afegeix una tercera consulta, com ara el camp *Contingut* (que conté 11 valors possibles), aleshores l'aplicació haurà d'executar 471 (3+39+429) consultes SQL (les combinacions possibles entre *Llengua*, *Tipus de Lloc* i *Contingut*). Les dades del tercer criteri de consulta es troben a la tercera graella, a la part inferior esquerra de la fitxa.

Si es decideix triar un quart criteri de consulta, s'assoleix el grau màxim de complexitat de consultes SQL del programa. D'aquesta manera, en triar com a quart criteri, per exemple *Tipus de Recurs* (que té 21 valors possibles), l'aplicació haurà executat 9.480 (3+39+429+9009) consultes SQL, corresponents a totes les combinacions possibles entre els valors triats. Els valors del quart criteri de consulta se situen a la graella de la part inferior dreta de la fitxa.

A causa de la possible complexitat de les consultes, sota el botó "Executar" se situa una barra de progrés que presenta l'evolució del procés. Per tal de visualitzar les consultes només cal triar els criteris desitjats a les respectives graelles. Hi ha tantes combinacions possibles com valors en els diferents criteris de consulta, de manera que quan es tria un criteri canvien els que en depenen. A l'exemple que hem indicat, quan triem com a *Llengua* el català, els altres criteris, jeràrquicament dependents, mostraran els seus valors en referència als documents en català, i el mateix succeeix quan triem un criteri respecte dels que en depenen (tant en la consulta 1, com en la consulta 2 i en la 3).

A la part dreta de la finestra es troba una llista anomenada *Documents associats*, que mostra els recursos del corpus recollit. Cada vegada que es tria un valor a qualsevol de les quatre graelles de consulta, aquesta llista indica automàticament els arxius que hi coincideixen. A l'exemple, si triem a la primera graella el valor *En* (documents en anglès) la llista presentarà tots els documents que compleixen aquesta condició; si a la segona graella triem *Lloc de Notícies*, a la llista es presenten els documents que compleixen el primer i el segon criteri de consulta. El mateix podem fer a la tercera i

quarta graella, segons els valors que puguin resultar interessants en el tercer i el quart criteris. Els documents triats es poden copiar i guardar en un directori perquè puguin ser consultats posteriorment com a corpus independent polsant el botó *A directori*, situat a la fitxa sota la llista de documents. També es pot visualitzar un document concret que prèviament hagi estat triat de *Documents associats* prement el botó *Veure*.

1.1.3. La fitxa estadístiques

Aquesta finestra presenta un gràfic amb els valors del criteri de consulta de nivell inferior: el quart, si s'han triat quatre criteris; el tercer, si s'han triat tres criteris; el segon, si s'han triat dos criteris; o el primer si només se n'ha triat un. També presenta el percentatge de cada valor recollit a l'estadística, amb relació al nombre total de recursos recollits en aquell criteri de consulta.

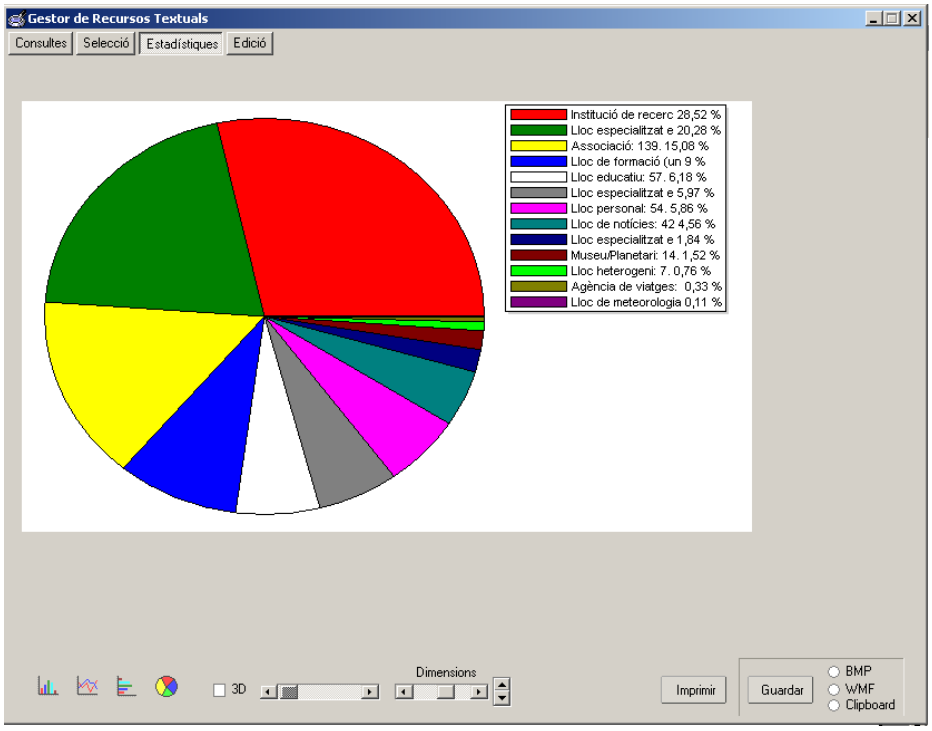


Figura 6-3. Interfície de la fitxa estadístiques del GeRT (e. p.)

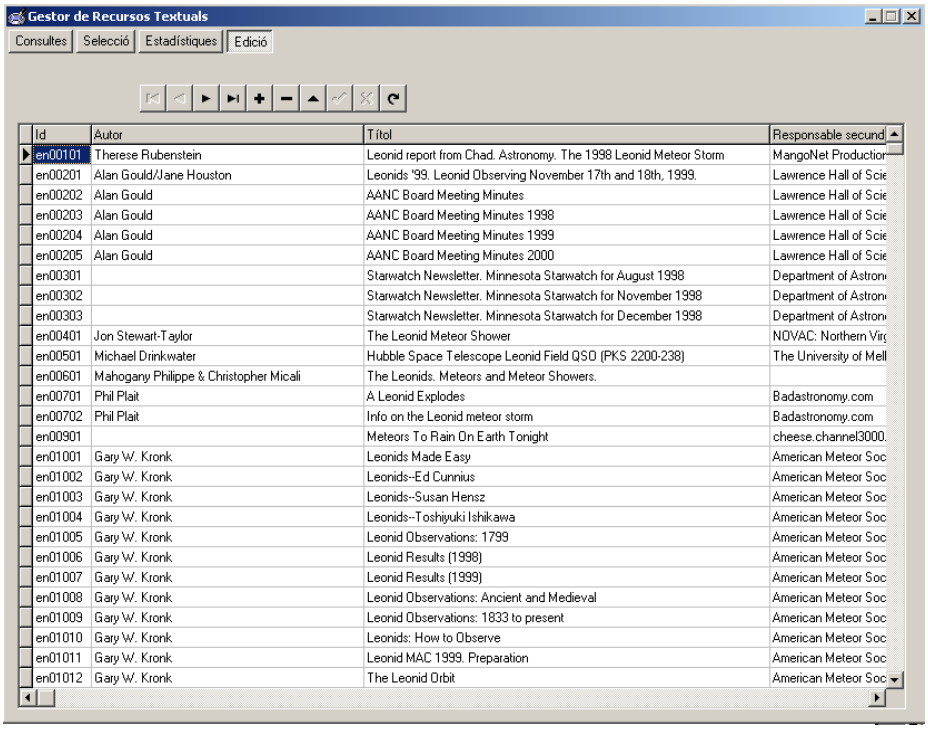
El gràfic estadístic es genera automàticament en activar aquesta finestra. També es pot generar prement el botó *Estadística* a la finestra *Selecció*. La presentació del gràfic estadístic es pot modificar amb els controls situats a la part inferior esquerra de la finestra, de esquerra a dreta són:

- Botons per generar diferents gràfics: en forma de barres verticals, cims, barres horitzontals o pastís.
- Quadre de selecció per a presentar el gràfic en tres dimensions.
- Controls per a definir les dimensions del gràfic.

El gràfic es pot imprimir amb el botó *Imprimir*, es pot guardar en format BMP o WMF, o bé es pot copiar per inserir-lo en una altra aplicació.

1.1.4. Edició

Aquesta finestra permet modificar els valors de la taula en què es descriu cadascun dels recursos textuals o inserir nous registres. Aquesta possibilitat permet utilitzar aquesta aplicació per gestionar qualsevol conjunt de recursos textuals que conformin un corpus, i no només el corpus recollit per a aquest treball de recerca. Per aquest motiu aquesta eina es pot adaptar per respondre a necessitats diferents, de recerca o docents, en el futur.



Id	Autor	Títol	Responsable secund
en00101	Therese Rubenstein	Leonid report from Chad. Astronomy. The 1998 Leonid Meteor Storm	MangoNet Production
en00201	Alan Gould/Jane Houston	Leonids '99. Leonid Observing November 17th and 18th, 1999.	Lawrence Hall of Scie
en00202	Alan Gould	AANC Board Meeting Minutes	Lawrence Hall of Scie
en00203	Alan Gould	AANC Board Meeting Minutes 1998	Lawrence Hall of Scie
en00204	Alan Gould	AANC Board Meeting Minutes 1999	Lawrence Hall of Scie
en00205	Alan Gould	AANC Board Meeting Minutes 2000	Lawrence Hall of Scie
en00301		Starwatch Newsletter. Minnesota Starwatch for August 1998	Department of Astron
en00302		Starwatch Newsletter. Minnesota Starwatch for November 1998	Department of Astron
en00303		Starwatch Newsletter. Minnesota Starwatch for December 1998	Department of Astron
en00401	Jon Stewart-Taylor	The Leonid Meteor Shower	NOVAC: Northern Virg
en00501	Michael Drinkwater	Hubble Space Telescope Leonid Field QSO (PKS 2200-238)	The University of Mell
en00601	Mahogany Philippe & Christopher Micali	The Leonids. Meteors and Meteor Showers.	
en00701	Phil Plait	A Leonid Explodes	Badastronomy.com
en00702	Phil Plait	Info on the Leonid meteor storm	Badastronomy.com
en00901		Meteors To Rain On Earth Tonight	cheese.channel3000.
en01001	Gary W. Kronk	Leonids Made Easy	American Meteor Soc
en01002	Gary W. Kronk	Leonids--Ed Cunnius	American Meteor Soc
en01003	Gary W. Kronk	Leonids--Susan Hensz	American Meteor Soc
en01004	Gary W. Kronk	Leonids--Toshiyuki Ishikawa	American Meteor Soc
en01005	Gary W. Kronk	Leonid Observations: 1799	American Meteor Soc
en01006	Gary W. Kronk	Leonid Results (1998)	American Meteor Soc
en01007	Gary W. Kronk	Leonid Results (1999)	American Meteor Soc
en01008	Gary W. Kronk	Leonid Observations: Ancient and Medieval	American Meteor Soc
en01009	Gary W. Kronk	Leonid Observations: 1833 to present	American Meteor Soc
en01010	Gary W. Kronk	Leonids: How to Observe	American Meteor Soc
en01011	Gary W. Kronk	Leonid MAC 1999. Preparation	American Meteor Soc
en01012	Gary W. Kronk	The Leonid Orbit	American Meteor Soc

Figura 6-4. Interfície de la fitxa Edició del GeRT (e. p.)

1.2. El paquet d'eines WordSmith Tools

WordSmith Tools és un conjunt d'eines de gestió i anàlisi de corpus dissenyat per Mike Scott i comercialitzat per l'Oxford University Press[2]. Aquest programa presenta diverses eines de gran utilitat. Entre aquestes eines, les que utilitzarem en el marc d'aquesta recerca són:

- WORDLIST o extractor de llistes de paraules.
- CONCORD o extractor de concordances.
- KEYWORDS o extractor de llistes de paraules clau.

[2] Trobareu més informació a la pàgina personal de Mike Scott (http://www.lexically.net/wordsmith/index.html) o a la pàgina d'Oxford University Press (http://www.oup.co.uk/isbn/0-19-459286-3).

A continuació descriurem les característiques més rellevants d'aquestes tres eines.

1.2.1. L'extractor de llistes de paraules

L'eina WORDLIST permet extraure llistats amb les paraules incloses en un conjunt de textos. Els llistats es poden generar a partir de tot el corpus, o bé arxiu per arxiu. En el nostre cas, l'opció que ens interessarà serà l'extracció de llistats de paraules del conjunt del corpus.

En cada operació, WORDLIST genera 3 llistats diferents, relatius al mateix conjunt de textos. En primer lloc, un llistat amb totes les paraules (*types*) presents al corpus ordenades en funció de la seva freqüència d'aparició: de la més freqüent a la menys freqüent. Aquesta llista permet detectar les paraules més utilitzades al llarg del corpus, que, d'una banda, seran paraules funcionals o gramaticals (habituals en tota mena de discurs), i de l'altra paraules plenes, amb càrrega semàntica, que seran indicatives de l'àmbit temàtic del corpus.

El segon lloc, WORDLIST genera una llista que recull de nou totes les paraules incloses al corpus, aquest cop ordenant-les de manera alfabètica. Aquesta llista permet cercar paraules per comprovar si es troben al corpus. Tot i ser una llista alfabètica, després de cada paraula també s'indica la seva freqüència d'aparició.

El tercer dels llistats generats recull informació estadística sobre les dades obtingudes, tant pel que fa al conjunt de textos interrogats com a cadascun per separat.

Des d'aquesta eina, un cop identificada una paraula que pugui resultar d'interès, es pot realitzar una cerca de concordances obrint automàticament l'eina Concord.

1.2.2. L'extractor de concordances

CONCORD es pot obrir directament des del controlador de les eines (WordSmith Tools CONTROLLER) o bé des de l'eina WORDLIST. L'extractor de concordances localitza i extrau tots els contextos d'una paraula determinada i els ordena utilitzant aquesta paraula

com a eix vertical. Un cop localitzats els contextos, permet reordenar-los alfabèticament en funció de les paraules a dreta i esquerra de la paraula pivot.

A més de cercar les concordances d'una determinada paraula, CONCORD també pot extraure els contextos d'una paraula en què necessàriament aparegui una altra paraula, determinada per l'usuari, a dreta o esquerra de la paraula pivot.

Opcionalment, es poden extraure concordances a partir de paraules truncades, és a dir, indicant únicament una part de la paraula pivot. D'aquesta manera, CONCORD podrà extraure contextos de paraules que comencin, continguin o finalitzin amb la cadena de caràcters que l'usuari haurà indicat com a paraula pivot. Aquesta opció resulta de gran utilitat quan es treballa amb un corpus no lematitzat, ja que permet, per exemple, recuperar els contextos de totes les formes conjugades d'un verb (si és un verb regular) o els de totes les formes flexionades d'un substantiu.

A partir del llistat de concordances extret, CONCORD pot mostrar els cosituats més habituals de la paraula pivot (les paraules que presenten un major índex d'aparició en el seu context), així com la posició que hi ocupen. Per defecte, el context que es té en compte a l'hora de dur a terme aquesta operació és des de la cinquena paraula a l'esquerra de la paraula pivot (posició –5) fins a la cinquena paraula a la seva dreta (posició +5); tanmateix, aquests valors es poden modificar. El llistat de cosituats pot indicar l'existència d'unitats de significat ampliades al voltant de la paraula pivot, o patrons estructurals habituals.

Finalment, també pot resultar de gran interès l'opció d'identificació d'agrupacions de paraules (*clusters*) en el context de la paraula pivot. Tanmateix, aquesta opció presenta dues limitacions: d'una banda, només extrau aquelles agrupacions que es troben entre les posicions –5 i +5 respecte a la paraula pivot (paràmetre per defecte que es pot ampliar o reduir, però que sempre es té en compte), per la qual cosa no pot cercar agrupacions a tot el corpus; d'altra banda, una agrupació és considerada un conjunt d'un nombre de paraules (determinat per l'usuari) que es repeteix amb una freqüència mínima (que per defecte és 3 cops). Per aquest motiu, només pot recuperar agrupacions d'un nombre determinat de paraules (dos o tres, o quatre, etc.), i no agrupacions formades per un nombre indeterminat de paraules (dues o més) que es

repeteixen amb una freqüència mínima; és a dir, agrupacions de dues paraules i/o de tres paraules, i/o de quatre paraules, etc., que es repeteixin amb una freqüència mínima. És per això que l'operació d'extracció d'agrupacions de paraules es durà a terme amb l'Extractor d'*n-grams* (vegeu l'apartat 1.3 "L'Extractor d'ngrams" d'aquest capítol, pàgina 295).

1.2.3. L'extractor de llistes de paraules clau

La tercera de les eines utilitzades en l'anàlisi del corpus compilat és KEYWORD. Aquesta eina compara un corpus especialitzat amb un de general per tal d'identificar les paraules clau del primer. Tècnicament, aquesta eina compara la llista de paraules del corpus especialitzat amb la del corpus general, tenint en compte que aquest darrer corpus ha de ser necessàriament més gran que el primer. La llista resultant recull les paraules que, pel seu ús, són representatives del corpus especialitzat.

Aquesta llista de paraules contindrà paraules que no apareixen en el corpus general així com paraules amb una freqüència d'aparició notablement més alta en el corpus especialitzat que en el corpus general, cosa que denota un ús específic d'aquestes paraules. A diferència de la llista de freqüències produïda per l'eina WordList, a la llista de paraules clau no hi apareixen paraules funcionals o gramaticals, ja que aquestes paraules generalment s'utilitzen amb una freqüència equivalent en textos especialitzats i generals.

1.3. L'Extractor d'*n-grams*

Aquest programa ha estat dissenyat i desenvolupat per Antoni Oliver Pérez[3] en l'àmbit del màster en Tradumàtica (Traducció i Tecnologies de la Informació i la Comunicació),

[3] Doctorand de la Universitat de Barcelona i professor del Màster Tradumàtica (Traducció i Tecnologies de la Informació i la Comunicació) (http://www.fti.uab.es/pg.tradumatica).

i no es comercialitza. L'Extractor d'*n-grams* està dissenyat en Perl, i calcula els n-grams[4] d'ordre 1 fins a 5 a partir d'un arxiu o conjunt d'arxius. El resultat, el llistat d'agrupacions de paraules o n-grams, el presenta ordenant-los de més a menys freqüència, i el desa en un arxiu de sortida en format text. Té la capacitat de tractar recursivament tot un directori, cosa que és molt útil per al tractament de corpus voluminosos.

Figura 6-5. Controlador de l'Extractor d'*N-grams* i finestra de configuració del programa (e. p.)

L'Extractor d'*N-grams* permet fer un filtratge de les agrupacions de paraules detectades mitjançant una llista de paraules buides (*stop-words list*), per tal de no recuperar agrupacions formades per paraules buides (agrupacions com ara *per la qual cosa* o *per tal de* en català). En concret, el filtratge elimina els n-grams que continguin una paraula buida en posició extrema, és a dir que la primera o l'última paraula de l'n-gram sigui una paraula buida de la llista.

L'extracció dels n-grams s'inicia seleccionant l'arxiu d'entrada amb el botó *Seleccionar origen* del controlador. Si el que es desitja és tractar recursivament tot un directori cal prèviament marcar l'opció *Tot el directori*. A continuació caldrà indicar l'arxiu de destí. En aquest arxiu es guardaran els n-grams en ordre descendent a la freqüència d'aparició, de més freqüent a menys freqüent.

Abans de començar el procés cal seleccionar la configuració desitjada amb el botó *Configuració*. Aquest botó ens obre la pantalla de configuració, la qual ens dóna les possibilitats següents:

[4] Vegeu l'apartat 3.1 "Elements d'anàlisi de la lingüística de corpus" del capítol 5, pàgina 270.

- Seleccionar l'ordre dels n-grams. Per defecte està seleccionat l'n 2 a 5.
- Si volem filtrar el resultat amb paraules buides, cal marcar l'opció *Filtratge amb stop-words* i indicar el fitxer de paraules buides amb el botó *Seleccionar fitxer stop-words*.

Per acceptar les opcions pitgem el botó *Acceptar*. Per començar el procés de càlcul d'n-grams cal pitjar el botó *Començar*. El procés acaba quan aquest botó torna a la seva posició original.

El següent llistat és un exemple d'una extracció d'n-grams realitzada amb aquest programa:

563	meteor shower
386	Leonid meteor
269	meteor storm
208	solar system
199	Leonid meteor shower
190	Leonid Meteor
186	meteor showers
186	meteors per hour
171	Leonid meteors
168	Astronomical Society
158	Meteor Shower
153	Air Force
127	Meteor Society
115	Leonid shower
113	shooting stars
110	bright meteors
106	Leonid storm
104	meteoroid stream
98	parent comet
95	population index

Taula 6-1. Primers 20 n-grams del corpus anglès (entre dues i cinc paraules per agrupació), precedits per la seva freqüència d'aparició (e. p.)

2. Comprovació de l'adequació del corpus multilingüe

La comprovació de l'adequació del corpus multilingüe com a corpus *ad hoc* del qual podem extraure informació durant la traducció d'un text especialitzat es basa en la comparació de les paraules més representatives del corpus amb les més representatives del text que cal traduir. Tot i que confiem en la utilitat d'aquesta metodologia de treball, sigui quina sigui la combinació lingüística de la traducció, en aquest cas seguirem l'esquema de la traducció directa (d'anglès a castellà o català).

L'adequació del corpus, doncs, en el cas del corpus anglès, consistirà a comparar la llista de paraules clau de l'original a traduir amb la llista de paraules clau del corpus anglès. Abans de dur a terme aquesta operació, descriurem dos textos en llengua anglesa que prendrem com a originals.

El primer dels textos escollits és un article de diari, titulat "Leonids: Meteor Shower Power", escrit per P. Friedlander i publicat al *Washington Post* l'11 de novembre de 2002 (pàgina B8). Aquest article de premsa està escrit presumiblement per un semiexpert, ja que enlloc no s'indica que sigui un expert. Inclou citacions textuals d'experts, cites bibliogràfiques i propostes de fonts per obtenir més informació sobre el tema. És un text curt, de 640 paraules, adreçat a un lector llec, amb coneixements de cultura general mitjans/alts (el perfil del lector habitual d'una publicació diària de premsa seriosa, no sensacionalista) i, per tant, un text poc especialitzat.

El segon dels textos escollits és un article de revista acadèmica, titulat "The Leonid Meteor Shower: Historical Visual Observations", escrit per P. Brown, del Department of Physics and Astronomy de la University of Western Ontario (London, Ontario, Canada), i publicat a la revista *Icarus* el 1999 (número 138, pàgines 287-308). Aquest article acadèmic està escrit per un expert. Està dividit en tres parts clarament diferenciables: un resum inicial, el cos de l'article i una bibliografia final. Al resum inicial s'indiquen les paraules clau que representen el contingut de l'article (METEORS; LEONIDS; VISUAL OBSERVATIONS). El cos de l'article està dividit en subapartats que ordenen el contingut

del text, que alhora també està il·lustrat amb gràfics. És un text relativament llarg, de 12.951 paraules, adreçat a un lector expert, el lector habitual de la revista.

Tant en el cas de l'article de premsa com en el de l'article acadèmic, la paraula LEONID apareix directament al títol. Aquest fet, juntament amb la llista de les seves paraules clau, justifiquen la cerca de textos paral·lels a partir del terme i l'àmbit temàtic a què pertany (LEONID + ASTRONOMY, vegeu l'apartat 2.1 "La cerca dels recursos textuals digitals especialitzats sobre els Leònids per a la posterior creació de corpus monolingües comparables" del capítol 3, pàgina 167).

2.1. Comprovació de l'adequació del corpus anglès

La comprovació de l'adequació del corpus, en aquest cas l'anglès, es realitzarà comparant les paraules clau de cadascun dels articles amb les del corpus. Per tal d'obtenir les paraules clau dels articles, i utilitzant l'eina KEYWORD de WordSmith Tools (vegeu l'apartat 1.2.3 "L'extractor de llistes de paraules clau" d'aquest capítol, pàgina 295), compararem, d'una banda, la llista de paraules generada a partir de cadascun dels articles amb la llista de paraules d'un corpus de caràcter genèric de la llengua anglesa, com és el British National Corpus (BNC)[5]. D'aquesta manera obtindrem una llista amb les paraules més representatives del contingut dels dos articles, d'una banda les de l'article de premsa i d'altra banda les de l'article acadèmic.

De la mateixa manera, també extraurem la llista de les paraules clau del subconjunt del corpus comparable format pels recursos textuals en llengua anglesa, que denominem corpus anglès, ja que, de manera independent a la resta de textos del corpus (els

[5] El BNC és una col·lecció de mostres de textos, escrits i transcrits de la llengua oral, representatius de la llengua anglesa moderna que conté 100 milions de paraules. Fou compilat per un consorci d'editorials de diccionaris (Oxford University Press, Longman, Chambers-Larousse) i de centres de recerca (Oxford University, Lancaster University i la British Library). La consulta d'aquest corpus es realitza amb una aplicació informàtica pròpia, el SARA, tot i que en el nostre cas hem extret una llista de paraules amb WordSmith Tools per poder obtenir les paraules clau posteriorment.

recursos en castellà i català), constitueix per ell mateix un corpus monolingüe especialitzat en aquesta llengua.

	PARAULES CLAU DE L'ARTICLE DE PREMSA
1	ASTRONOMER
2	WWW
3	NOV
4	M
5	METEORS
6	NASM
7	METEOR
8	EDU
9	LEONID
10	P
11	OBSERVATORY
12	PLANETARIUM
13	SKY
14	ASTRONOMERS
15	SPACE
16	APOLLO
17	MUSEUM
18	LECTURE
19	STORM

	PARAULES CLAU DE L'ARTICLE ACADÈMIC
1	LEONID
2	LEONIDS
3	ZHR
4	OBSERVATIONS
5	ASTRON
6	SHOWER
7	METEOR
8	METEORS
9	PEAK
10	ACTIVITY
11	ZHRS
12	STORM
13	J
14	TEMPEL
15	OLIVIER
16	MAXIMUM
17	PROFILE
18	FIG
19	OLMSTED
20	NODAL
21	NOVEMBER
22	OBSERVERS
23	TUTTLE
24	GAUSSIAN
25	YEOMANS
26	DATA
27	LONGITUDE
28	HR
29	MOON
30	POP

	PARAULES CLAU DEL CORPUS ANGLÈS
1	METEOR
2	LEONID
3	METEORS
4	LEONIDS
5	SHOWER
6	COMET
7	STORM
8	UT
9	EARTH
10	SKY
11	NASA
12	NOVEMBER
13	NOV
14	SPACE
15	PEAK
16	ORBIT
17	HTTP
18	TUTTLE
19	ASTRONOMY
20	SOLAR
21	TEMPEL
22	RADIANT
23	ZHR
24	OBSERVING
25	TELESCOPE
26	OBSERVERS
27	DUST
28	MOON
29	OBSERVATIONS
30	OBSERVATORY

Taula 6-2. Paraules clau dels articles i del corpus anglès respecte al BNC (e. p.)

Mentre que de l'article de premsa només hem aconseguit una llista de dinou paraules clau, les llistes produïdes a partir de l'article acadèmic i del corpus anglès han estat més extenses. Tanmateix, a la taula 6-2 es mostren les 30 primeres paraules clau de cada llista. Entre les coincidències més rellevants entre totes tres llistes cal destacar el fet

que tant la paraula LEONID com la paraula METEOR es trobin entre les 10 primeres paraules.

En la llista de paraules clau queda palès que l'eina WordSmith Tools pren com a paraula qualsevol cadena de caràcters entre espais en blanc, per la qual cosa recull com a paraules expressions com ara WWW o EDU. Aquestes expressions, pròpies de la URL d'una pàgina web, no consten al BNC, per la qual cosa queden recollides a la llista de paraules clau de l'article de premsa. D'altra banda, paraules com STORM, que formen part del vocabulari anglès habitual i, per tant, estan incloses en el BNC, apareixen en aquest article amb una freqüència relativa molt més alta (el 0,42‰) que en el corpus general (inferior al 0,01‰), per la qual cosa també formen part de la llista de paraules clau. Aquesta diferència en la freqüència d'aparició indica un possible ús específic en aquest àmbit, hipòtesi que queda corroborada en comprovar que en la llista de paraules clau de l'article acadèmic també es dóna la mateixa situació.

Entre les coincidències més rellevants que garanteixen l'adequació del corpus anglès compilat *ad hoc* per a la traducció de l'article de premsa cal destacar, a més de LEONID, METEOR i STORM, coincidències com SPACE i OBSERVATORY.

Pel que fa a l'article acadèmic, crida especialment l'atenció la inclusió de la paraula ACTIVITY entre les seves primeres trenta paraules clau, ja que es tracta d'una paraula d'ús freqüent en llengua anglesa. De fet, ACTIVITY apareix al BNC amb una freqüència del 0,01‰ (en total, 11.642 vegades). Tanmateix, l'aparició d'aquesta paraula a l'article acadèmic resulta relativament més alta que en el corpus general, ja que hi apareix amb una freqüència del 0,67‰ (en total, 113 vegades). De nou, aquesta diferència en la freqüència d'aparició denota un ús específic d'ACTIVITY en l'article acadèmic. A més, aquesta paraula ocupa el lloc quaranta-sisè en la llista de paraules clau del corpus, amb una freqüència del 0,13‰.

Una altra freqüència que hem de recalcar és la de la paraula NOVEMBER, recollida a la llista de paraules clau de l'article acadèmic, com també a la del corpus. Aquesta paraula apareix amb una freqüència relativa del 0,35‰, mentre que en el BNC la seva

freqüència és inapreciable (inferior al 0,01‰)[6]. Per la seva banda, en el corpus anglès NOVEMBER presenta una freqüència d'ús del 0,25%. Aquest fet indica un ús d'aquesta paraula molt més freqüent en aquest àmbit temàtic que no pas en el la llengua general, cosa que no atribuïm a una utilització especial de NOVEMBER, sinó a una vinculació entre aquest mes i el tema que tracten tant els articles[7] com el corpus.

A més de paraules pròpies del vocabulari general, la freqüència de les quals pot indicar un ús especial en l'àmbit i que, en coincidir tant en la llista de paraules clau dels articles com en la llista del corpus, en garanteix l'adequació, en la llista de l'article especialitzat hi apareixen sigles, com ara ZHR[8], i noms propis, com són TEMPLE i TUTTLE, que també apareixen a la llista de paraules clau del corpus.

Finalment, l'aparició de certes unitats entre les paraules clau del corpus, com ara HTTP i NASA indiquen la procedència dels recursos textuals recollits. La primera indica l'origen dels recursos textuals digitals que configuren el corpus, Internet; la segona fa referència a un dels organismes més rellevants en l'àmbit de l'astronomia, per la qual cosa no només ha estat emissora de part dels recursos del corpus, sinó que també s'hi fa referència des de recursos elaborats per altres institucions o per autors independents.

2.2. Paraules clau del corpus castellà

Tant en el cas del català, com es veurà més endavant, com en el del castellà, l'adequació del corpus *ad hoc* no es pot comprovar comparant-lo amb textos originals que s'han de traduir, ja que són les llengües a les quals traduirem i, per tant, no hi

[6] El programari utilitzat no és capaç de calcular freqüències inferiors al 0,01‰, per la qual cosa les freqüències d'aparició més baixes resulten inapreciables.

[7] En el cas de l'article de premsa, no apareix NOVEMBER, sinó NOV, una forma abreviada.

[8] ZHR, tal com és podrà desprendre de les concordances que s'utilitzaran durant l'anàlisi del corpus anglès, són les sigles de Zenith Hourly Rate.

tenim cap text de partida. Això no obstant, i de manera sil·logística, si hem utilitzat les mateixes paraules clau per documentar-nos tant en anglès com en castellà i en català, i si el corpus anglès s'adequa a l'àmbit temàtic, coincidint en els usos específics del lèxic dels originals que cal traduir, podríem concloure que les parts en castellà i català del corpus comparable també s'adequaran al tema.

Tanmateix, en analitzar les paraules clau del corpus castellà s'hi poden identificar certes paraules que el faran coincidir amb el corpus anglès i, ateses les seves característiques, ens permetran deduir-ne l'adequació.

La llista de paraules clau del corpus castellà l'hem aconseguida comparant-lo amb el corpus general del castellà Lexesp[9]. El resultat de comparar la llista de paraules del corpus castellà compilat (corpus especialitzat) amb la llista de paraules del Lexesp (corpus genèric) ha donat com a resultat la llista de paraules clau recollida a la taula següent:

1	LEÓNIDAS	16	HTTP
2	METEOROS	17	NASA
3	LEO	18	TORMENTA
4	NOV	19	PER
5	OBSERVACIÓN	20	MAGNITUD
6	COMETA	21	ASTRORED
7	ACTIVIDAD	22	ESP
8	LLUVIA	23	RADIANTE
9	INFO	24	LUNA
10	NOVIEMBRE	25	MÁXIMO
11	HORA	26	OBSERVAR
12	ESTRELLAS	27	WWW
13	ASTRO	28	ASTRONÓMICA
14	ASTRONOMÍA	29	OBSERVADORES
15	R	30	FUGACES

Taula 6-3. 30 primeres paraules clau del corpus castellà en relació amb el corpus LEXESP (e. p.)

[9] *Lexesp: Léxico informatizado del español*, compilat pel Centre de Llenguatge i Computació de la Universitat de Barcelona, amb un total de 6 milions de paraules.

Les dues paraules que clarament en garanteixen l'adequació són LEÓNIDAS i METEOROS, que el traductor, partint de la seva cultura general, podrà identificar com a equivalents de LEONIDS i METEORS, paraules aquestes últimes que figuren tant en el corpus anglès com en els articles. D'altra banda, la presència de NOVIEMBRE entre les principals paraules clau del corpus, amb una freqüència relativa d'aparició del 0,18‰ (mentre que en el Lexesp és inferior al 0,01‰), també indica aquesta correlació temàtica entre el corpus castellà i l'anglès.

Altrament, el fet que NASA aparegui en la setzena posició de la llista de paraules clau també ens permet entreveure la correlació temàtica entre el corpus castellà i l'anglès, i, per tant, l'adequació del corpus castellà a l'àmbit temàtic dels originals. Aquest fet també queda corroborat amb la presència d'altres paraules, com ara NOVIEMBRE, OBSERVAR i ACTIVIDAD.

2.3. Paraules clau del corpus català

El corpus català, tal com s'ha descrit a l'apartat 3.1 "El procés d'identificació dels recursos textuals dels corpus" del capítol 3 (pàgina 177), està format per recursos textuals obtinguts a partir de dos cerques diferents: d'una banda, la cerca utilitzant les paraules clau LEÒNIDS i ASTRONOMIA i les seves variants ortogràfiques (correctes i incorrectes); d'altra banda, la cerca basada en les paraules clau METEOR i ASTRONOMIA. Per aquest motiu, analitzarem en primer lloc l'adequació del subcorpus que resulta de la primera cerca, i a continuació l'adequació del conjunt de recursos que integren el corpus català.

El subcorpus format pels recursos textuals obtinguts a partir de la primera cerca, igual que succeïa en el cas del corpus castellà, es podria considerar equivalent al corpus anglès pel que fa a la seva adequació, ja que ha estat compilat mitjançant la mateixa estratègia de cerca. Tanmateix, de nou, en observar les paraules clau d'aquest conjunt de recursos textuals, també trobem certs elements que ens garanteixen la seva

adequació. Per tal d'extreure les paraules clau d'aquest subcorpus, el compararem amb una part del corpus de l'Institut d'Estudis Catalans[10].

1	METEORS	16	TAMBÈ
2	ÈS	17	PLUJA
3	COMETA	18	PÀGINA
4	WEB	19	METEORITS
5	VILAWEB	20	LEÒNIDS
6	LLUNA	21	INFORMACIÓ
7	NOV	22	FUJI
8	NOVEMBRE	23	D'ESTELS
9	MÈS	24	NEBULOSA
10	CCD	25	HALE
11	TM	26	LEÒNIDES
12	P	27	PAÏS
13	INTERNET	28	MM
14	ECLIPSI	29	BOPP
15	NGC	30	CATALÀ

Taula 6-4. 30 primeres paraules clau del subcorpus català aconseguit a partir de la cerca de LEÒNIDS i ASTRONOMIA en relació al corpus de l'IEC (e. p.)

De nou, la presència de NOVEMBRE i NOV representa un punt de coincidència amb les paraules clau del corpus anglès i castellà, com també METEORS i METEORIT. En aquest cas, és de recalcar que LEÒNIDS i LEÒNIDES no ocupen una posició tant privilegiada com als altres casos. Tanmateix, si sumem la freqüència d'aparició de totes dues denominacions, el resultat seria equiparable als obtinguts en els llistats de paraules clau dels corpus anglès i castellà.

D'altra banda, en analitzar l'adequació de tot el corpus català i extreure les seves paraules clau, els resultats no mostren una adequació tan evident com en el cas anterior.

[10] Aquest subcorpus està format per un total de 7.272.114 paraules i correspon als documents més actuals recollits al corpus de l'IEC.

1	TERRA	16	OF
2	SOBRE	17	DURANT
3	SOL	18	METEORIT
4	SER	19	BARCELONA
5	DES	20	NOVEMBRE
6	VAN	21	COMETA
7	THE	22	SOLAR
8	VILAWEB	23	COMETES
9	LLUNA	24	MART
10	METEORITS	25	KM
11	DIA	26	INFORMACIÓ
12	WEB	27	NIT
13	AVUI	28	P
14	MILIONS	29	LLUM
15	METEORS	30	PLUJA

Taula 6-5. 30 primeres paraules clau del corpus català en relació al corpus de l'IEC (e. p.)

En aquest cas, la presència de paraules derivades de METEOR és recalcable, i també la de COMETA, tant en singular com en plural. No obstant això, no hi trobem cap referència temporal que ens permeti garantir l'adequació del corpus, com en canvi podíem fer a partir de les paraules clau del corpus anglès i castellà. També es troba a faltar la paraula clau LEÒNID, que no hi consta.