Capítol 4. La classificació dels corpus de recursos textuals

Un cop recollits tots els recursos textuals que s'analitzaran com a textos paral·lels al capítol 6, i per tal que aquests recursos poguessin configurar un corpus, s'han tingut en compte tota una sèrie de paràmetres que, d'una banda, permet classificar-los i conèixer les seves característiques extralingüístiques més rellevants i, de l'altra, que ens permetrà dividir els corpus finals en subcorpus en funció d'aquestes mateixes característiques.

Abans de presentar aquesta classificació (a la qual es dedica l'apartat 2 "La classificació dels recursos textuals digitals especialitzats dedicats als Leònids", pàgina 197) descriurem breument en el proper apartat els conceptes principals provinents de la lingüística de corpus així com els seus postulats teòrics i metodològics.

1. La lingüística de corpus

La lingüística de corpus es pot considerar una metodologia que analitza textos en tant que realització d'un sistema lingüístic (Biber, 1988; Biber, Conrad i Reppen, 1996 i 1998; Biber i Conrad, 1999; Leech, 1991; Sánchez i Cantos, 1997), o bé el fonament teòric d'un nou paradigma d'anàlisi de la llengua basat en la indissolubilitat de sentit i forma (ex. Stubbs, 2001 i 2002; Tognini-Bonelli, 1996 i 2001). El seu plantejament empíric, centrat en l'estudi de l'ús de la llengua i, per tant, prenent com a objecte d'estudi textos escrits o orals, entra en clara contraposició amb el tipus d'estudi realitzat tradicionalment, que es basa principalment en l'observació del sistema lingüístic (Leech, 1991: 8). En paraules de Biber, Conrad i Reppen: "[r]ather than looking at what is theoretically possible in a language, we study the actual language used in naturally occurring texts" (1998: 1).

Aquest nou apropament permet descobrir els nexes d'unió entre el que virtualment pot ser i el que finalment és en llengua, i així es pot tornar a connectar la llengua real (en tant que sistema) amb els textos reals (en tant que realització d'aquest sistema). De fet, teòrics com de Beaugrande apostaren per aquesta metodologia en considerar que "corpus linguistics may well provide the great missing links between virtual and actual and reconnect language with text, or between 'language with parole', or between 'competence with performance', and so on" (de Beaugrande, 1999: 256), en clara al·lusió als postulats chomskyans i saussurians.

La lingüística de corpus, doncs, es presenta com a alternativa a la metodologia tradicional, la qual, basada principalment en la introspecció i la intuïció lingüística, parteix del coneixement que els usuaris tenen sobre la llengua per il·lustrar-lo mes tard amb exemples. Contràriament a aquesta metodologia, la lingüística de corpus proposa un apropament empíric que permeti descriure l'ús que una comunitat fa de la llengua tot partint de dades autèntiques. Es tracta, doncs, d'una metodologia fonamentalment inductiva, segons la qual és la descripció de dades textuals concretes la que permet formular hipòtesis que tinguin en compte els fets observats (Tognini-Bonelli, 2001: 2).

Tot i que aquesta metodologia s'ha posat en pràctica recentment recaptant resultats notables, val a dir que no es tracta d'una proposta nova. Leech (1991) troba arrels de la lingüística de corpus en el corrent estructuralista nord-americà tal com l'entenien Leonard Bloomfield i els seus seguidors. Aquest corrent proposava una visió de la ciència altament influïda per les doctrines positivista i conductista, per a les quals el conjunt dels textos és el primer dels elements que la lingüística ha de poder explicar. Aprofundint en aquest origen estructuralista, altres autors, com ara Sinclair, Stubbs, Teubert o Tognini-Bonelli contemplen la lingüística de corpus com la continuació dels postulats de John R. Firth, el qual insistí que la noció 'llengua' com a objecte de recerca només es pot observar en els textos escrits o orals. Segons els principis de Firth, desenvolupats per Halliday i aplicats per Sinclair, la lingüística és una ciència de caire social i aplicat que s'hauria d'estudiar a partir de dades autèntiques, refusant la recerca intuïtiva que recorre a oracions inventades. És més, advoquen per l'observació d'unitats textuals senceres, i no d'oracions aïllades, i la realització d'estudis comparatius d'aquestes unitats a partir de corpus de textos. Firth i els seus seguidors consideren que els estudis lingüístics s'han de basar essencialment en el significat, i que **forma** (i en conseqüència estructura) i **significat** són inseparables; tal com assenyala Tognini-Bonelli, quan en aquest cas parla del treball de John Sinclair:

> "In this respect, Sinclair, like Firth and Halliday, is adamant in claiming that the distinction between form and meaning is only a methodological convenience and this leads him to posit formal observations as criteria for analysing." (Tognini-Bonelli, 2001: 99)

Aquests plantejaments estructuralistes teòrics no es van poder dur a la pràctica sobre conjunts de textos prou voluminosos per ser representatius d'una llengua fins a l'arribada de la informàtica als estudis lingüístics. Leech (1991) considera que els avenços informàtics han significat el ressorgiment dels estudis sobre corpus, que d'alguna manera havien quedat de banda a causa, també, de la preponderància de les teories chomskianes. Biber, Conrad i Reppen resumeixen els avantatges de l'ús d'eines informàtiques en els estudis basats en corpus en els termes següents:

> "Computers make possible to identify and analyze complex patterns of language use, allowing the storage and analysis of a larger database of natural language than could be dealt with by hand. Furthermore, computers provide consistent, reliable analyses – they don't change their mind or become tired during an analysis." (Biber, Conrad i Reppent, 1998: 4).

Així doncs, els avantatges de la utilització d'ordinadors en la recerca sobre la llengua són fonamentalment els dos següents: 1) la capacitat d'emmagatzemar i analitzar un gran volum d'informació, que no està a l'abast de la capacitat humana; i 2) la garantia d'uns resultats finals. Tanmateix, i tal com aquests autors també indiquen, els estudis basats en corpus no consisteixen únicament en el recompte de característiques o estructures lingüístiques, sinó que la seva rellevància rau en les "qualitative functional interpretations of quantitative patterns" (Biber, Conrad i Reppent, 1998: 5). D'aquesta manera, la recerca es basarà en dades autèntiques obtingudes amb un objectiu determinat que els investigadors interpreten, i no en un exercici introspectiu, tot i que no es renuncia a la intuïció de l'usuari de la llengua; en paraules de Leech, es tracta de "a question of corpus plus intuition, rather than of or intuition" (1991: 74). De fet, l'eina bàsica amb què s'analitzen els corpus, els cercadors de concordances o KWIC (*Key Words in Context*), que cerquen una paraula i ofereixen tots els seus contextos ordenats fent servir la mateixa paraula com a eix central vertical dels segments trobats, no pot arribar a conclusions per ella mateixa, sinó que és l'usuari qui interpreta els resultats que li proporciona el KWIC i considera si són dades rellevants o no.

Com a metodologia, la lingüística de corpus ha estat molt ben rebuda en disciplines relacionades amb l'estudi de la llengua, sobretot les de caràcter més aplicat, i entre elles especialment en la lexicologia (i la seva vessant més aplicada, la lexicografia). En aquest àmbit, la implantació d'aquesta metodologia ha permès l'elaboració de diccionaris basats en dades extretes de grans corpus. Els avantatges més subratllables de la utilització de grans corpus com a font d'informació per a la creació de recursos lexicogràfics són els següents:

- Proporcionen material autèntic, que pot resultar d'utilitat a l'hora d'indicar exemples d'ús en l'entrada de cada paraula.

- Permeten que els lexicògrafs prenguin decisions relatives a canvis de sentit a partir de dades fefaents.
- Proporcionen informació sobre, per exemple, estructures gramaticals, categoritzacions o registres.
- Ofereixen informació sobre freqüències d'ús.
- Proporcionen dades sobre neologismes, noves combinacions de paraules i col·locacions. (Meijs, 1996: 102)

Si bé l'aplicació més destacable de la lingüística de corpus ha estat en l'àmbit de la lexicografia, fins al punt de donar lloc a una àrea comuna denominada lexicografia basada en corpus (*corpus-based lexicography*), aquesta metodologia empírica també ha estat molt ben acollida en àmbits propers, com la traducció automàtica, l'ensenyament de segones llengües o la lingüística textual. En terminologia, l'extracció automàtica o semiautomàtica d'informació a partir de corpus és una activitat incipient (Vivaldi i Rodríguez, 2001); tanmateix, en la part aplicada d'aquesta disciplina també es parla de terminologia basada en corpus (vegeu l'apartat 2.3 "La metodologia de treball de la terminologia", del capítol 5, pàgina 266).

Per tal que les dades obtingudes resultin fiables, s'han d'establir certs criteris d'admissió i classificació dels textos que formen part del corpus. A continuació analitzarem els paràmetres que es tenen en compte a l'hora de dissenyar un corpus, així com la tipologia de corpus resultant i la seva aplicació més habitual.

1.1. Els tipus de corpus

Amb el ressorgiment d'aquesta metodologia empírica, la mateixa definició del concepte de corpus s'ha convertit en un dels aspectes al qual els teòrics han prestat més atenció. Com a resum de les característiques indicades pels autors més representatius (Abaitua, 2000; Biber, Conrad i Reppen, 1998; Corpas, 2001; Leech 1991; Marcos Marín, 1994;

McEnery i Wilson, 1996; Sánchez i Cantos, 1997; Sinclair 1991 i 1996[1]; Stubbs, 1996 i 2002; Teubert 1996 i 2001; Tognini-Bonelli, 1996 i 2001), direm que un corpus es caracteritza principalment per ser un conjunt suficientment gran de **dades reals** de la llengua (escrita o parlada) que es vol investigar i estar en un **format processable** per ordinadors. El conjunt de textos del corpus ha de ser **representatiu** de la llengua o varietat de llengua que es vol analitzar, ja sigui un dialecte o una varietat lingüística professional, per tal que els resultats de l'estudi puguin ser extrapolables a aquella varietat de llengua. Per aquest motiu, els textos que formen part del corpus han de ser un **exemple** de la varietat lingüística que s'hi vol representar. A més, el conjunt de textos del corpus ha d'estar **ordenat** en funció de l'objectiu de la recerca que es vulgui dur a terme.

Aquests cinc aspectes, l'autenticitat de les dades, el seu format digital, la representativitat del conjunt de les dades i el caràcter il·lustratiu de cadascuna de les mostres, així com els criteris amb què s'ordenen per formar el corpus, resulten fonamentals a l'hora de dissenyar un corpus, i es concretaran d'una manera o altra en funció del tipus de recerca que es vulgui dur a terme.

L'autenticitat de les dades és una premissa que pràcticament es dóna per sobreentesa en lingüística de corpus. Tot i així, s'acostuma a realitzar un preprocessament de les dades, generalment amb l'objectiu de donar-hi el format digital necessari per tal d'analitzar-les. Si el que es vol dur a terme és un estudi fonològic, els textos seran orals i s'obtindrà un **corpus oral**, mentre que si es vol realitzar un altre tipus d'estudi, els textos seran escrits o transcripcions de textos orals i formaran un **corpus escrit**. Sovint, aquesta fase de preprocessament inclou l'etiquetatge dels textos. L'etiquetatge consisteix a marcar als textos els elements que després facilitaran un tipus determinat d'anàlisi; generalment es marquen elements morfo-sintàctics o lexicosemàntics. D'altra banda, també es poden marcar elements macroestructurals o microestructurals, com són els paràgrafs, les oracions i les frases del text. Així, doncs, en funció del tipus

[1] Aquesta obra està recollida a la bibliografia sota la referència EAGLES 1996. Sinclair fou l'autor de les directrius sobre disseny de corpus de l'EXPERT ADVISORY GROUP ON LANGUAGE ENGINEERING STANDARDS (EAGLES), una de les iniciatives de la Comissió Europea en el marc del programa de Recerca i Enginyeria Lingüística de la Direcció General XIII.

d'estudi que es vulgui dur a terme es dissenyarà un **corpus etiquetat** o un **corpus no etiquetat**.

La representativitat del corpus és un paràmetre ineludiblement vinculat a l'objectiu de la recerca per a la qual es vol fer servir. Es pot recórrer a la metodologia de la lingüística de corpus per dur a terme estudis sobre l'estat actual d'una llengua, de la seva evolució o del seu estat en un moment històric concret, partint d'un **corpus sincrònic**, d'un **corpus diacrònic** o d'un **corpus periòdic o cronològic** respectivament; d'altra banda, si es vol analitzar un tipus de llengua particular es dissenyarà un **corpus especialitzat**, mentre que si l'objecte d'estudi és la llengua sense cap limitació temàtica o d'altre tipus es partirà d'un **corpus general**.[2]

Els corpus poden estar formats per textos sencers (**corpus textual**) o bé per fragments de textos que actuen com a mostra representativa de l'estat de la llengua amb totes les seves varietats (**corpus de referència**). La quantitat de text que s'inclou pot afectar la representativitat del corpus i, en conseqüència, els resultats finals de l'estudi; si en el disseny del corpus es té en compte l'existència de diverses varietats de llengua i es volen incloure amb un determinat pes, calculant el percentatge de textos de cada varietat que hi formarà part, l'estudi estarà basat en un **corpus equilibrat**. Per oposició, els que no tinguin en compte la representativitat i l'equilibri entre variants seran **corpus grans** o **extensos**. Hi ha un altre tipus de corpus que, pel tipus de recerca que permeten desenvolupar, han de mantenir un volum de dades constant, de manera que a mesura que incorporen nous textos o fragments de textos n'eliminen quantitats equivalents d'antics. Aquests corpus, denominats **corpus monitor**, permeten "observar cambios recientes en el uso de la lengua, convirtiéndose en una referencia viva de la propia evolución ligüística" (Corpas, 2001: 158).

A més de considerar el text com a exemple d'una varietat de llengua, tal com s'ha indicat en el paràgraf anterior, introduint-lo completament al corpus o bé extraient-ne un fragment, també pot resultar interessant, en funció del tipus d'estudi que es vol

[2] Corpas inclou en la seva classificació un cinquè tipus de corpus en funció de la llengua que s'estudia, el **corpus canònic**, format per les obres d'un determinat autor.

realitzar, conèixer aspectes extralingüístics que pugin resultar rellevants. Aquesta informació es pot recollir en arxius independents (denominats DTD o *document type definition*), en un arxiu registre al qual els usuaris tenen accés o directament com a capçalera de cadascun dels textos del corpus. El corpus que proporciona aquesta informació es denomina **corpus documentat**, mentre que els corpus que no la proporcionen es consideren **corpus no documentats**.

La tipologia exposada fins al moment pot aplicar-se tant a corpus que continguin textos en una sola llengua, els **corpus monolingües**, com els que en continguin més d'una, que pot ser un **corpus bilingüe** (amb textos en dues llengües) o un **corpus multilingüe** (amb textos en tres o més llengües). Aquest segon tipus de corpus es pot subdividir en diferents categories en funció de la relació que es pugui establir entre els textos en diferents llengües i el tipus de recerca per a la qual es compilen.

El corpus format per textos originals en dues o més llengües que comparteixen una situació sociopragmàtica equivalent es denominen **corpus comparables** (Bowker, 2000: 20; Peters and Picchi, 1997: 254; Teubert, 1996: 245). Aquests textos comparteixen característiques com ara el tema, el període de publicació, el tipus de text, etc.

Un altre tipus de corpus amb textos amb més d'una llengua és l'anomenat **corpus paral·lel**, que conté textos que guarden entre ells alguna de les relacions següents:

- Textos escrits en una llengua A i la seva traducció a una llengua B (i C ...).
- Textos escrits originalment en una llengua A i B, amb les seves respectives traduccions.[3]
- Textos traduïts a les llengües A, B i C, escrits originalment en la llengua Z. (Teubert i Kervio-Berthou, 2000: 145).

[3] Aquest tipus de col·lecció de textos, en la tipologia de Baker (1995: 234) es considera un **corpus comparable**, categoria que defineix com un conjunt de textos originals i traduïts a les llengües A i B, i no com conjunt de textos originals en més d'una llengua.

Formalment, aquests corpus poden mantenir els textos en cadascuna de les llengües per separat, o bé alinear-los de manera que cada oració en una llengua vagi acompanyada per l'oració equivalent en les altres llengües. Aquest tipus de corpus paral·lel rep el nom de **corpus alineat**. Tanmateix, els corpus paral·lels que no són susceptibles de ser alineats reben la denominació de **corpus de traducció lliure** (*free translation corpus*, Tognini-Bonelli, 2001: 6).

Els corpus bilingües o multilingües permeten dur a terme estudis interlingüístics. En traductologia, aquests corpus permeten dur a terme els següents tipus de recerques:

- Estudis comparatius entre més d'una llengua l'objectiu dels quals és identificar equivalències, ja sigui en l'àmbit textual o de gènere, morfosintàctic (gramàtiques comparades) o lèxic.
- Estudis comparatius entre textos originals i traduïts en una mateixa llengua. Aquest tipus de recerca permet reflexionar sobre les diferències que hi pugui haver entre textos originals i traduïts, determinar-ne les causes i elaborar material didàctic i postulats teòrics per pal·liar aquestes diferències.

Quant a la recerca que es vol dur a terme no està limitada per una varietat lingüística o la comparació entre dues llengües, sinó per la informació final que es vol obtenir es considera que es recull un **corpus amb una finalitat especial** (*special purpose corpus*, Pearson, 1998: 48) o un **corpus *ad hoc*** (Corpas, 2001: 173). La composició d'aquests corpus està condicionada per la finalitat de l'estudi que es realitzarà; es pot extreure, per exemple, d'un corpus de referència, però no necessàriament ha de respectar els seus paràmetres de representativitat o d'equilibri. Es recullen per tal d'extreure'n definicions o observar candidats a unitats de sentit especialitzat, i s'utilitzen com a recurs per a la didàctica i la pràctica de la traducció.

Així doncs podem concloure que els conjunts de textos recollits i classificats, tal com veurem en el proper apartat, conformen un corpus amb les característiques següents:

- **Escrit**: conté únicament dades autèntiques produïdes per escrit.
- **No etiquetat**: conté text pur (*raw text*), sense cap mena d'etiquetatge.

- **Especialitzat**: inclou textos dedicats a un àmbit temàtic determinat, en concret al fenomen astronòmic dels Leònids.
- **Sincrònic**: tots els textos han estat publicats entre l'any 1997 i el 2001.
- **Textual**: conté unitats textuals senceres que són pàgines web. Tanmateix, la pàgina web és una unitat virtual, però no coincideix necessàriament amb la unitat de sentit, que és el document hipertextual. Si bé, llavors, aquest corpus no coincideix plenament amb la definició de corpus textual, tampoc no pot considerar-se un corpus de referència, ja que, encara que el text d'una pàgina web és un fragment d'un document hipertextual, aquest fragment no està determinat per cap criteri de representativitat o de manteniment de l'equilibri entre varietats lingüístiques incloses al corpus. Per aquest motiu hem decidit considerar-lo un corpus textual.
- **No equilibrat**: si bé el conjunt de textos en llengua anglesa sí que es pot considerar un corpus monolingüe gran o extens, el volum d'informació en les altres dues llengües es tan baix que no pot rebre aquesta denominació. Per aquest motiu hem decidit caracteritzar-lo tot posant en relleu l'absència d'equilibri entre el volum d'informació en cadascuna de les llengües.
- **Documentat**: el corpus va acompanyat d'un arxiu en format text amb la descripció dels textos (vegeu l'apartat 2.1.6 "Esquema final de classificació dels recursos digitals" d'aquest mateix capítol, pàgina 209).
- **Multilingüe**: conté textos en castellà, català i anglès.
- **Comparable**: els textos són originals i coincideixen en tema, període de publicació i situació comunicativa.
- ***Ad hoc***: s'ha recollit per fer-lo servir com a font de recursos per a la traducció especialitzada.

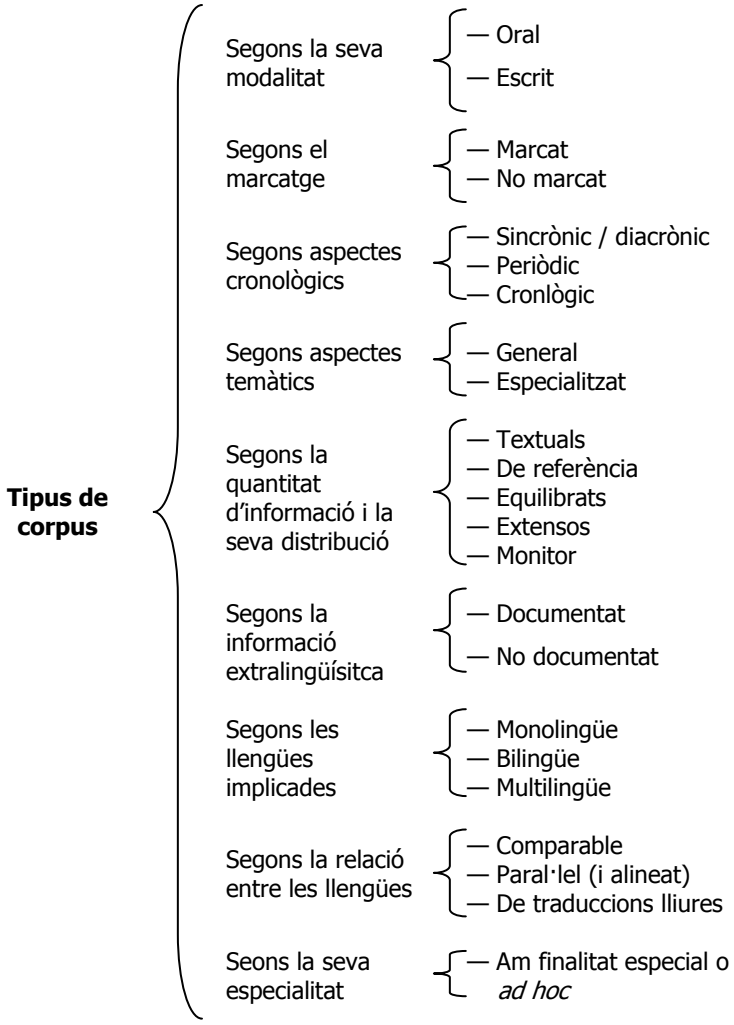

Figura 4-1. Esquema resum dels tipus de corpus (e. p.)

1.2. La composició d'un corpus *ad hoc*

En decidir crear un corpus *ad hoc*, el seu disseny ha de respondre a l'objectiu de la recerca que volem dur a terme. En el nostre cas, l'objectiu de la recerca és doble:

- Establir el potencial de la informació textual digital publicada al web pública com a recurs per a la traducció especialitzada.
- Proposar un protocol metodològic d'extracció d'informació de corpus similars útil en la pràctica professional de la traducció.

Obeint, doncs, a aquest doble objectiu, el corpus resultant presenta unes característiques molt específiques que descriurem a continuació basant-nos en els quatre criteris descriptius fonamentals proposats per l'Expert Advisory Group on Language Engineering Standards (EAGLES, 1996), que són la quantitat o dimensió, la qualitat, la simplicitat i la documentació.

Tal com assenyala Sinclair en les directrius d'EAGLES (1996), la dimensió d'un corpus, entesa com el volum de paraules, depèn en la pràctica de la dificultat amb què es puguin recollir els seus components (els textos que en formen part). En el nostre cas, i tal com s'ha assenyalat al capítol anterior, els components del corpus són tots els recursos textuals a l'abast de qualsevol usuari d'Internet recuperats a partir d'una cerca basada en paraules clau que en delimiten l'àmbit temàtic (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)[4]. Si finalment el subcorpus anglès resulta molt més voluminós que el castellà i el català, haurà estat una conseqüència directa del desequilibri entre llengües que hi ha a Internet.

Pel que fa a la qualitat dels components del corpus, el nivell mínim exigible segons EAGLES és que siguin autèntics, com ho són en aquest cas. Atès que el primer dels objectius d'aquesta recerca consisteix a comprovar fins a quin punt la informació publicada al web públic pot resultar útil com a font de recursos per a la traducció especialitzada, no s'ha cregut convenient partir de criteris apriorístics d'avaluació de la qualitat, com proposen, per exemple, Corpas (2002), Pérez Hernández (2002a i 2002b) o Pearson (1998).

[4] D'aquesta manera, a més, coincidim amb un dels criteris de creació de corpus *ad hoc* indicats per Pérez Hernández (2002b: 142), en concret el de pertinença a un domini d'especialitat.

El tercer dels criteris proposats per EAGLES defensa la simplicitat dels components inclosos en el corpus. La simplicitat dels components implica l'absència d'elements que, en estar barrejats amb el text, puguin distorsionar l'anàlisi del corpus i els resultats finals. Per aquest motiu, si es vol incloure informació addicional, lingüística o extralingüística, s'ha de separar clarament entre els components textuals que s'analitzaran i aquests elements addicionals. En aquest sentit, EAGLES proposa que els components del corpus es trobin en format text pla (ASCII), i que qualsevol marcatge addicional es faci mitjançant etiquetes SGML. En el nostre cas, tots els recursos textuals, que originalment es trobaven en format HTML, han estat convertits al format ASCII am la finalitat, d'una banda, de seguir la proposta d'EAGLES i, de l'altra, d'evitar el possible soroll que les etiquetes HTML podrien provocar en l'anàlisi del corpus i en els seus resultats.

EAGLES proposa com a últim criteri la documentació dels components del corpus, i així l'hem aplicat. En aquest cas, els components del corpus han estat classificats en funció de tota una sèrie de paràmetres extralingüístics, que s'abordaran en el proper apartat. Aquesta informació ha estat recollida en un arxiu de registre i descripció independent que es pot consultar mitjançant una eina informàtica dissenyada amb aquest propòsit per l'autora que, a més, facilitarà la subdivisió del corpus a partir dels paràmetres descriptius seguits.

D'altra banda, Pérez Hernández (2002a i 2002b), a més d'adoptar els criteris d'EAGLES, també proposa escollir els components del corpus *ad hoc* tenint en compte criteris de qualitat, temàtics i de caràcter pragmàtic (fonamentalment la data de publicació i les condicions comunicatives dels textos). Com hem indicat anteriorment, en el nostre cas l'estudi de la qualitat dels recursos textuals, en tant que fonts d'informació per a la traducció especialitzada, és un dels objectius de la nostra recerca, i per això no hem cregut convenient considerar-lo com a criteri *a priori*. Altrament, tant el criteri temàtic com el mitjà comunicatiu (un dels criteris pragmàtics proposats per Pérez Hernández), formen part de la definició explícita del nostre corpus especialitzat *ad hoc*: està dedicat a un àmbit concret de l'astronomia i conté únicament recursos extrets del WWW públic.

Finalment, Pérez Hernández també proposa incloure en el corpus únicament aquells components que responguin al tipus textual, el nivell de tecnicitat i el tipus de receptor corresponents a la varietat de llengua que es vol observar. En el nostre cas, i tal com s'analitzarà amb detall a continuació, aquests criteris han estat tinguts en compte, entre d'altres; tanmateix, no s'han fet servir com a paràmetre de filtratge dels recursos que s'han inclòs al corpus, sinó únicament com a paràmetres descriptors mitjançant els quals podrem estudiar amb més profunditat el potencial informatiu dels recursos textuals digitals.

2. La classificació dels recursos textuals digitals especialitzats dedicats als Leònids

Davant de la incapacitat del traductor de jutjar per ell mateix la qualitat de cadascun dels recursos textuals que consulta, es veu obligat a recolzar-se en paràmetres ja establerts que li permetin decidir si es troba davant d'un recurs de qualitat o no. En alguns casos es recorre conscientment o inconscientment a paràmetres establerts per centres d'informació com els descrits a l'apartat 1.1.2 "Tipus de recursos web i qualitat de la informació" del capítol 3 (pàgina 127). Es tracta de criteris que permeten jutjar el valor d'un recurs de manera objectiva i absoluta en funció d'una sèrie de premisses establertes.

Aquests criteris d'avaluació tenen en compte diferents aspectes, com són ara l'autor, l'actualitat i la precisió de la informació, el tractament i l'originalitat del contingut, el propòsit del recurs, l'existència d'enllaços externs, la seva ergonomia, l'entorn informàtic i la presència de citacions d'altres recursos, com també el receptor ideal del recurs.[5] Així doncs, es considera un recurs de qualitat aquell que té com a autor un expert en la matèria, és actual, original, etc. Altres recursos, en els quals per exemple no figura l'autor o no són actuals, no es consideren de qualitat i, per tant, en un centre d'informació, com ara una biblioteca, serien descartats.

El traductor, però, no sempre cerca documents de qualitat en termes absoluts, sinó que sovint necessita documents que siguin de qualitat amb relació al seu encàrrec de traducció (que podrien ser no actuals o esbiaixats, com l'original que ha de traduir). D'altra banda, a Internet es troba una alta quantitat de recursos textuals que no acompleixen els criteris de qualitat indicats anteriorment i que potser individualment no ofereixen un mínim nivell de fiabilitat, tot i que això no vulgui dir que necessàriament continguin informació errònia; tanmateix, el fet d'analitzar-los en conjunt i de manera quantitativa, com es farà al capítol 6, sí que pot dotar-los de la fiabilitat que individualment no tenen.

[5] S'ha fet una descripció més detallada d'aquests aspectes a l'apartat 1.1.2 "Tipus de recursos web i qualitat de la informació" d'aquest capítol, en concret, a la pàgina 133.

Per aquest motiu, a l'hora de classificar els recursos textuals dels tres corpus monolingües tindrem en compte tant els criteris proposats des de les ciències de la informació, que valoren la qualitat de manera absoluta, com altres que poden ajudar al traductor a estimar el seu valor relatiu, en relació amb l'encàrrec de traducció i l'original que ha de traduir. Inicialment no es descartarà cap recurs per manca de qualitat, tot i que la descripció de cadascun d'ells en funció dels criteris que a continuació es presenten sí que permetrà dividir els corpus en subcorpus a l'hora de fer l'anàlisi lingüística i conceptual.

2.1. Criteris de classificació dels recursos textuals

A continuació descriurem tots els criteris inclosos com a paràmetres de classificació dels corpus, que es poden agrupar de la manera següent:

1. Dades administratives.
2. Dades sobre l'autor i el receptor.
3. Dades sobre l'estructura del document hipertextual.
4. Tipus de lloc.
5. Tipus de recurs textual.

Finalment es presentarà la fitxa de classificació dels recursos textuals i s'analitzaran els recursos textuals que presenten les combinacions de paràmetres més freqüents.

2.1.1. Dades administratives

Per dades administratives entenem totes aquelles dades que permeten identificar el recurs textual a Internet. La majoria també són necessàries a l'hora de fer una correcta citació bibliogràfica del recurs (Estivill i Urbano, 1997), tot i que sovint no hi consten explícitament.

En concret, les dades administratives recollides són les següents:

- **Autor o responsable principal**: persona que ha elaborat el recurs.
- **Títol del recurs digital**: el títol que rep el recurs. En aquest cas s'ha tingut en compte tant el títol que apareix dins del recurs textual (l'equivalent al títol d'un article en format paper) com la informació inclosa al metacamp "títol" del document en format HTML, que apareix a la barra superior del navegador.
- **Responsable secundari**: institució a la qual pertany l'autor i/o responsable de la plana que inclou el recurs.
- **Títol del document digital**: el títol que rep el lloc web que inclou el recurs digital. En aquest cas també s'ha tingut en compte la informació recollida al metacamp "títol". Sovint el nom del responsable secundari també és el títol del document digital.
- **Lloc de publicació**: generalment coincideix amb el lloc de la seu central del responsable secundari. Només s'ha indicat si al recurs hi constava explícitament o si es podia desprendre del domini de la seva URL.
- **Llengua**: la llengua de redacció del recurs. Tots els recursos es prenen com a originals, ja que en cap recurs no s'indica que sigui una traducció (tot i que tampoc no s'indica el contrari).
- **Data de creació del recurs digital**: la data en què el recurs es va publicar per primer cop, si hi consta.
- **Data d'actualització del recurs digital**: l'últim cop que el recurs va ser actualitzat, si hi consta.
- **Data de consulta del recurs digital**: la data en què es va descarregar el recurs d'Internet.
- **Adreça URL del recurs digital**: el localitzador del recurs a Internet.
- **Adreça URL de la pàgina inicial del document digital**: el localitzador del document hipertextual a Internet, en cas que aquell recurs en concret deixi d'existir a la xarxa, almenys amb aquella adreça.

2.1.2. L'autor i el receptor

Si en molts casos l'autor del recurs no hi consta explícitament, el seu receptor ideal no hi consta gairebé mai. Si bé els recursos textuals es redacten pensant en una situació comunicativa determinada que inclou un receptor ideal, la comunicació mitjançant llocs públics d'Internet és absolutament oberta i no presenta cap restricció d'accés. Per aquest motiu es podria parlar del receptor ideal d'un recurs ateses les seves característiques, però en cap cas de receptor únic.

Per tot això, i basant-nos en la tipologia d'interlocutors de la comunicació especialitzada de Ciapuscio i Kuguel (2002), hem adoptat la categorització següent:

- **Tipus d'autor**:
  - *Expert*: capaç d'expressar coneixements teòrics i aplicats.
  - *Semiexpert*: capaç d'expressar coneixements aplicats o de reelaborar el discurs d'un expert per tal que el pugui llegir un llec.
- **Tipus de receptor**:
  - *Expert*: capaç d'assimilar coneixements teòrics complexos i aplicats.
  - *Semiexpert*: capaç d'assimilar coneixements teòrics simples i aplicats.
  - *Llec*: capaç d'assimilar coneixements introductoris.

Com ja s'ha assenyalat, els recursos sovint no indiquen el seu autor. Per aquest motiu, i davant la impossibilitat de caracteritzar l'autor real de cada recurs, s'ha optat per caracteritzar el perfil mínim que l'autor de cada recurs ha de posseir per tal de poder-lo elaborar. En el cas del receptor s'ha actuat de la mateixa manera.

2.1.3. Dades sobre l'estructura del document hipertextual

Tot i partir de la hipòtesi que l'estructura del document hipetextual que es publica a Internet no acostuma a reflectir l'estructura conceptual de l'àmbit temàtic que tracta, no hem volgut perdre l'ocasió de comprovar-ho amb els recursos recollits. A més de recollir informació sobre l'entorn del recurs digital (sempre dins del document

hipertextual al qual pertany), també hem volgut posar en relleu la naturalesa multimèdia dels recursos tot indicant els elements que el componen.

Així doncs, les dades sobre l'estructura del document recollides són les següents:

- **Components del recurs digital**: arxius d'imatge, so, vídeo o que permeten interacció que complementen el recurs textual.
- **Recursos que envolten el recurs digital**: les adreces URL d'altres recursos d'aquell mateix document hipertextual des dels quals es pot accedir al recurs que s'està descrivint mitjançant un enllaç intern. D'aquesta manera es pot reproduir l'estructura del document prenent com a centre el recurs que s'està analitzant.

2.1.4. Tipus de lloc web

Per a la caracterització dels llocs web s'han adoptat dues tipologies:

a) La tipologia d'Alexander i Tate (2001), que prové de les ciències de la informació.

b) El catàleg de llocs web observats durant la classificació dels recursos textuals.

Com es veurà finalment, el catàleg de llocs web descrits es pot prendre com una subdivisió de la tipologia d'Alexander i Tate, ja que es tenen en compte altres criteris que aquests autors obvien.

a) Tipologia proposada per Alexander i Tate

La tipologia d'Alexander i Tate (2001), descrita a l'apartat 1.1.2 "Tipus de recursos web i qualitat de la informació" del capítol 3 (pàgina 127), pren com a criteri principal

l'objectiu de l'emissor. Així doncs, els llocs web es classifiquen en funció del propòsit de l'autor o, subsidiàriament, el responsable secundari.

La classificació adoptada és la següent:

- **Llocs A**: llocs de suport.
- **Llocs B**: llocs publicitaris i de negocis.
- **Llocs C**: llocs de notícies.
- **Llocs D**: llocs informatius.
- **Llocs E**: llocs personals.

De tota manera, i com que sovint els llocs inclouen pàgines amb les quals es vol aconseguir propòsits diferents, de vegades aquesta classificació no s'ha aplicat només sobre llocs, sinó també directament sobre els recursos textuals classificats.

b) Tipologia de llocs amb recursos textuals d'astronomia observats

Durant l'anàlisi dels recursos textuals per a la seva classificació hem anat trobant trets comuns entre diferents llocs que ens han ajudat a caracteritzar-los d'una manera més detallada, basant-nos no només en l'objectiu de l'autor sinó també en el tipus de responsable secundari. Alguns dels llocs web descrits a continuació es poden trobar en qualsevol àmbit temàtic, mentre que altres són exclusius d'àmbits científics o d'astronomia.

- **Associació**: lloc web d'una associació, de caràcter amateur, amb dades sobre les seves activitats i informacions addicionals.
- **Institució de recerca**: lloc web d'un institut de recerca en què es publica informació relativa a la seva recerca.
- **Museu/planetari**: lloc web a cavall entre les institucions de recerca i els llocs educatius o de formació. Generalment informen sobre la recerca que emparen i ofereixen, a més, informació addicional.
- **Lloc de formació (universitat)**: lloc web d'una institució dedicada a la formació d'experts.

- **Lloc de notícies**: llocs dedicats a la difusió d'informació en general, sovint associats a empreses dedicades a la comunicació.
- **Lloc de meteorologia**: llocs dedicats a la difusió d'informació meteorològica, sovint associats a empreses dedicades a la comunicació o a institucions públiques o privades que generen informació d'aquesta mena.
- **Lloc educatiu**: lloc web d'una institució dedicada a l'educació primària o secundària. Poden ser llocs dels mateixos centres d'educació o bé d'editorials o similars que proporcionen informació addicional sobre matèries educatives.
- **Lloc especialitzat en ciències**: lloc web equivalent a una revista especialitzada dedicada a les ciències.
- **Lloc especialitzat en astronomia**: lloc web equivalent a una revista especialitzada dedicada a l'astronomia.
- **Lloc especialitzat en Leònids**: lloc web equivalent a una revista especialitzada dedicada als Leònids.
- **Lloc heterogeni**: lloc dedicat a qualsevol altre tema que conté un recurs textual dedicat als Leònids.
- **Agència de viatges**: lloc d'una agència de viatges que conté un recurs textual dedicat als Leònids (generalment informació addicional a una oferta de viatge per veure aquest fenomen astronòmic).
- **Lloc personal**: lloc dedicat a pàgines personals o pàgina personal dins d'un lloc institucional.

(Al capítol 7 "Proposta metodològica sobre l'obtenció i l'explotació de recursos textuals digitals com a textos paral·lels" (pàgina 431) es durà a terme una extrapolació d'aquesta classificació independentment de l'àmbit temàtic.)

2.1.5. Tipus de recurs textual

A l'hora de catalogar els recursos textuals hem tingut en compte diferents factors:

- La relació entre dades i text, que descriu el grau d'abstracció en la redacció del recurs.

- Els trets comuns observats durant la classificació dels recursos textuals.
- Els trets comuns del contingut del recursos textuals.

a) La relació dades/text

Tal com Hoffmann assenyalava (1998: 64), el nivell d'abstracció es pot veure reflectit en la dimensió lingüística d'un text, de manera que com més abstracte és la manera de tractar el tema, més fàcil resulta trobar expressions alienes al llenguatge natural entre els enunciats del text. Tanmateix, hem pogut comprovar que aquestes expressions artificials, com ara fórmules numèriques o alfanumèriques, en aquest cas no són un indicador directe del nivell d'abstracció teòrica del recurs textual a què es referia Hoffmann, sinó que també es fan servir en recursos de caire més aplicat i metodològic. Per aquest motiu no podem fer una associació directa entre nivell d'abstracció i interlocutors, recollint la hipòtesi de Hoffmann segons la qual la comunicació entre experts és la més abstracta.

La presència d'expressions artificials condicionarà obligatòriament la fase d'anàlisi lingüisticoconceptual dels corpus, per la qual cosa hem cregut oportú indicar si el recurs textual està format únicament per text o no i, en aquest cas, en quina relació. Per aquest motiu hem previst els paràmetres següents:

- **Nivell 4**: recurs textual que conté únicament text redactat.
- **Nivell 3**: recurs textual que conté text redactat que es completa amb dades[6].
- **Nivell 2**: recurs textual que conté dades que es completen amb text[7].
- **Nivell 1**: recurs textual que conté pràcticament només dades[8].

[6] Per exemple, el recurs EN06901.

[7] Per exemple, el recurs EN05646.

b) Tipologia observada de recursos textuals

Durant l'anàlisi dels recursos textuals es van anar identificant recursos que compartien tota una sèrie de característiques, com ara la funció, la seva forma, la perspectiva sobre el tema o altres, que ens permetia agrupar-los sota etiquetes descriptives com són ara les següents:

- **Administratiu**: recurs textual creat en el marc d'una associació i generalment intern en què es repassen les activitats realitzades, com també altres temes administratius.
- **Article**: text que tracta un tema de manera asincrònica.
  - **Article acadèmic**: coincideix amb el prototipus d'article acadèmic en format paper, fins al punt que tampoc no inclou cap element que no pertanyi directament al text, com ara *banners* publicitaris o menús que l'incloguin dins d'un document hipertextual. Semblen textos pensats per ser publicats en paper i que finalment també han estat publicats en format digital.
  - **Article de divulgació**: recurs textual publicat per una institució rellevant en l'àmbit amb la intenció de proporcionar informació sobre un tema. Formalment, està inclòs al document hipertextual mitjançant un menú i també acostuma a oferir una llista d'enllaços cap a recursos més especialitzats que permeten completar la informació proporcionada al final del cos del recurs.
  - **Article especialitzat**: recurs textual elaborat per un expert o publicat en un lloc especialitzat que té com a objectiu la comunicació entre experts o semiexperts. Formalment pot presentar esquemes molt diferents, amb presència o absència de menús i llista d'enllaços addicional.
- **Butlletí**: publicació regular d'una associació o institució. Generalment inclou informació sobre més d'un aspecte. La informació del butlletí pot aparèixer

[8] Per exemple, el recurs EN04602.

inclosa en un sol arxiu o distribuint cada unitat informativa en un arxiu diferent.

- **Comentari de fotografia**: recursos en què els elements centrals són fotografies que van acompanyades per petits comentaris textuals.
- **Compendi d'informacions**: recurs format per breus notes informatives generalment sobre diferents temes. Aquests recursos són de caràcter sincrònic i divulgatiu.
- **Conversa**: recurs textual de caràcter sincrònic amb informació especialitzada i dades sobre circumstàncies personals de l'autor, sense separar un tipus d'informació de l'altra. El to és habitualment col·loquial i és l'equivalent a una conversa entre experts o semiexperts.
- **Entrevista**: entrevista a un expert. Recurs amb forma de pregunta–resposta.
- **FAQ (*frequently asked questions* o preguntes més freqüents)**: recurs textual divulgatiu asincrònic que presenta la informació en forma de preguntes i respostes. Les preguntes es converteixen en subapartats del recurs, i la seva funció és la d'introduir un aspecte concret del tema, per la qual cosa són els elements més destacats de la seva macroestructura.
- **Glossari**: recurs monolingüe amb forma de llista alfabètica de termes acompanyats per una definició.
- **Llistat/índex**: recurs que consisteix en un llistat de dades, generalment enllaços hipertextuals amb comentaris.
- **Material de formació**: recursos textuals asincrònics desenvolupats en l'àmbit d'una institució o un lloc web dedicat a la formació primària o secundària. Acostumen a ser lectures addicionals o materials de suport d'un curs.
- **Nota de premsa**: recurs textual sincrònic publicat per un expert o una institució experta per proporcionar informació sobre un tema a periodistes, que després la processaran i la difondran entre llecs.
- **Notícia**: recurs textual sincrònic que generalment presenta unes característiques formals molt concretes: data al començament del recurs, menú o llista d'enllaços que el vinculen a la resta de recursos del document i

complementen la notícia (com ara notícies anteriors), llista d'enllaços externs, espai per opinar sobre la notícia i funció per enviar-la a algú mitjançant el correu electrònic.

- **Recurs D**: recurs textual que tracta un tema de manera asincrònica sense cap més marca específica. Pertany a una unitat textual superior subdividida en arxius, per la qual cosa es pot considerar un subapartat. No acostuma a presentar cap format en concret, però sovint inclou enllaços que el vinculen al subapartat immediatament superior (l'índex o pàgina inicial del recurs textual), així com als recursos immediatament anteriors i posteriors.
- **Recurs E**: recurs textual personal, que no presenta cap tret habitual clar.
- **Recurs interactiu**: recurs format per elements en qualsevol de les morfologies de la informació amb el qual el lector o receptor interactua per tal d'obtenir una informació final.
- **Resultats d'observació**: recurs en forma de taula en què es descriu una sessió d'observació del fenomen astronòmic, en el nostre cas els Leònids. La informació s'expressa mitjançant abreviatures i valors alfanumèrics.

En aquesta relació de tipus de recursos textuals es troben clarament tant els equivalents de gèneres convencionals com els tipus de recursos propis de l'àmbit de l'astronomia:

Distribució dels tipus de recursos textuals			
En l'àmbit general[9]		**En l'àmbit de l'astronomia**	
AMB EQUIVALENTS EN FORMAT PAPER	SENSE EQUIVALENTS EN FORMAT PAPER	EXCLUSIUS D'ASTRONOMIA	NO EXCLUSIUS D'ASTRONOMIA
Administratiu	Article	Resultats d'observació	Administratiu
Article acadèmic	Compendi d'informacions		Article
Article de divulgació			Article acadèmic
Article especialitzat	Conversa (experts i/o semiexperts)		Article de divulgació
Butlletí			Article especialitzat
Comentari de fotografia	FAQ		Butlletí
	Llistat/Índex		Comentari de fotografia
Entrevista	Recurs D		Compendi d'informacions
Glossari	Recurs E		Conversa (experts i/o semiexperts)
Material de formació	Recurs interactiu		Entrevista
Nota de premsa			FAQ
Notícia			Glossari
Resultats d'observació			Llistat/Índex
			Material de formació
			Nota de premsa
			Notícia
			Recurs D
			Recurs E
			Recurs interactiu

Taula 4-1. Classificació de la tipologia de recursos textuals digitals (e. p.)

c) Tipologia de continguts observada

Pel que fa al contingut dels diferents recursos hem observat les regularitats temàtiques següents:

- **Història**: informació de caire històric sobre els Leònids.
- **Leònids específic**: informació sobre un aspecte en concret de els Leònids.
- **Leònids general**: informació sobre els Leònids des d'una perspectiva generalista, adoptant diverses perspectives.

[9] Aquesta comparació parteix de les característiques dels recursos en format digital observats, de manera que es considera que tenen un equivalent en format paper quan en aquests últims presenten les mateixes característiques que els digitals.

- **Metodologia**: informació sobre la metodologia que cal seguir per observar els Leònids.
- **Miscel·lània**: recurs textual en què es tracten temes de diversa índole, entre els que hi ha els Leònids.
- **Observació del fenomen**: informació sobre l'observació dels Leònids.
- **Observació i predicció del fenomen**: informació sobre la predicció i l'observació dels Leònids.
- **Predicció del fenomen**: informació sobre la predicció dels Leònids.
- **Recerca**: informació sobre investigacions que tenen a veure amb els Leònids duta a terme per institucions superiors o de recerca.
- **Tecnologia**: informació en relació amb l'efecte dels Leònids sobre elements tecnològics, com ara els satèl·lits.
- **Altres**: informació sobre altres temes en què es fa referència als Leònids.

2.1.6. Esquema final de classificació dels recursos digitals

Així doncs, i com a resum dels apartats anteriors, la informació recollida sobre cadascun dels recursos textuals que formen part dels corpus monolingües és la que es reflecteix en aquesta fitxa:

Fitxa de classificació dels recursos textuals digitals		
Dades administratives	Autor o responsable principal	
	Títol del recurs digital	
	Responsable secundari	
	Títol del document digital	
	Lloc de publicació	
	Llengua	
	Data de creació del recurs digital	
	Data d'actualització del recurs digital	
	Data de consulta del recurs digital	
	Adreça URL del recurs digital	
	Adreça de la pàgina inicial del document digital	
Autor/Receptor	Tipus d'autor	Expert
		Semiexpert
	Tipus de receptor	Expert
		Semiexpert
		Llec
Dades sobre l'estructura	Components del recurs digital	
	Recursos que envolten el recurs digital	
Tipus de lloc web	Segons la classificació d'Alexander i Tate	a: Llocs de suport i defensa
		b: Llocs publicitàries i de negocis
		c: Llocs de notícies
		d: Llocs informatives
		e: Llocs personals
	Tipus de llocs identificats	Agència de viatges
		Associació
		Institució de recerca
		Lloc de formació (universitat)
		Lloc de meteorologia
		Lloc de notícies
		Lloc educatiu
		Lloc especialitzat en ciències
		Lloc especialitzat en astronomia
		Lloc especialitzat en els Leònids
		Lloc heterogeni
		Lloc personal
		Museu/Planetari
Tipus de recurs	Relació dades/text	4
		3
		2
		1
	Tipus de recurs observats	Administratiu
		Article
		Article acadèmic
		Article de divulgació
		Article especialitzat
		Butlletí

		Comentari de fotografia
		Compendi d'informacions
		Conversa
		Entrevista
		FAQ
		Glossari
		Llistat/índex
		Material de formació
		Nota de premsa
		Notícia
		Recurs D
		Recurs E
		Recurs interactiu
		Resultats d'observació
	Continguts	Història
		Leònids específic
		Leònids general
		Metodologia
		Miscel·lània
		Observació del fenomen
		Observació i predicció del fenomen
		Predicció del fenomen
		Recerca
		Tecnologia
		Altres

Taula 4-2. Esquema de classificació dels recursos textuals digitals especialitzats en Leònids (e. p.)

Aquesta descripció tan detallada ens permetrà subdividir cadascun dels tres corpus i realitzar anàlisis de subcorpus resultants en funció de trets comuns entre els recursos, com poden ser el tipus de lloc, el tipus de recurs o el tipus de contingut. D'aquesta manera es podrà deduir quins són els recursos que ha de cercar el traductor per a dur a terme la seva documentació.

2.2. Descripció del corpus

En aquest apartat es presenta una descripció exhaustiva dels corpus recollits a partir de diferents combinacions de paràmetres. Aquestes combinacions que abordarem ens permetran conèixer més profundament les característiques de caire fonamentalment

pragmàtic dels recursos textuals que formen part dels corpus, fet que facilitarà l'anàlisi que posteriorment es durà a terme (vegeu el capítol 6 "L'anàlisi del corpus compilat", pàgina 283).

Per a la caracterització dels corpus que a continuació presentem recorrerem de manera il·lustrativa a les dades relatives al corpus anglès, el qual, atès el seu volum, és el que presenta una major complexitat a tots els nivells. Si s'escau, però, també s'indicaran les possibles diferències que es puguin donar entre les característiques del corpus anglès i les dels corpus català i castellà. Tanmateix, a l'annex B s'inclouen les taules més representatives de del corpus català i castellà.

2.2.1. Descripció dels corpus monolingües

El corpus monolingüe especialitzat anglès està format per 922 arxius que contenen un total de 1.446.436 paraules (*tokens*) corresponents a 36.351 formes (*types*). El corpus castellà està format per 115 arxius que sumen un total de 149.501 paraules i 12.268 formes. El corpus català el formen 77 arxius, 21 dels quals han estat obtinguts amb la cerca a partir de la paraula clau "Leònids" i les seves variacions, i 56 amb la cerca del seu hiperònim; en total, aquest corpus conté 89.406 paraules corresponents a 12.698 formes diferents (22.072 paraules i 4.683 formes del subconjunt de recursos obtinguts amb la paraula clau "Leònids" i 67.334 paraules i 10.479 formes del segon subconjunt).

En no estar etiquetat de cap manera, entre les formes computades s'inclouen, per exemple, formes plurals i singulars d'un mateix substantiu o diferents formes conjugades d'un verb, per la qual cosa no es pot parlar de formes canòniques.

La informació administrativa recollida sobre la totalitat dels recursos resulta relativament baixa, puix que són dades que no s'acostumen a indicar explícitament. Sovint la informació sobre el lloc de publicació, o fins i tot el nom de l'autor o responsable principal del recurs, s'obvia o no hi consta. Aquest fet ja seria motiu d'exclusió dels índexs de recursos de qualitat.

Tots els recursos recollits compten amb un títol de presentació. Malauradament, però, la resta de dades administratives només hi consten en el volum de recursos següent:

Dades administratives	Recursos en anglès		Recursos en castellà		Recursos en català	
Autor o responsable principal	512	(55,5%)	95	(82,6%)	13	(16,9%)
Responsable secundari	901	(97,7%)	104	(90,4%)	69	(89,6%)
Títol del document hipertextual	905	(98,2%)	108	(93,9%)	73	(94,8%)
Data de creació del recurs	412	(44,7%)	61	(53,1%)	20	(25,9%)
Data d'actualització	221	(23,9%)	5	(4,3%)	10	(12,9%)
Lloc de publicació	238	(25,8%)	114	(99,1%)	77	(100%)

Taula 4-3. Informació administrativa que consta als recursos (e. p.)

Els corpus en anglès, castellà i català, la situació és similar excepte en el cas del lloc de publicació, que apareix en un percentatge molt més alt, ja sigui directament al recurs textual inclòs al corpus o a la plana inicial del document hipertextual a què pertany. També crida l'atenció la presència gairebé a tots els recursos del títol del document hipertextual, cosa que probablement està relacionada amb el fet que sovint el responsable secundari dóna nom al seu lloc web.

Pel que fa als objectius amb els quals s'analitzarà el corpus, la dada que més es troba a faltar en el corpus anglès és el lloc de publicació. Com que no és possible saber si, en aquest cas, l'anglès és la llengua materna del responsable principal, aquesta discriminació es podria haver dut a terme subsidiàriament a partir del lloc de publicació. De tota manera, i tot i que les dades següents només fan referència al 25,81% dels recursos en anglès, la distribució per països de procedència és aquesta:

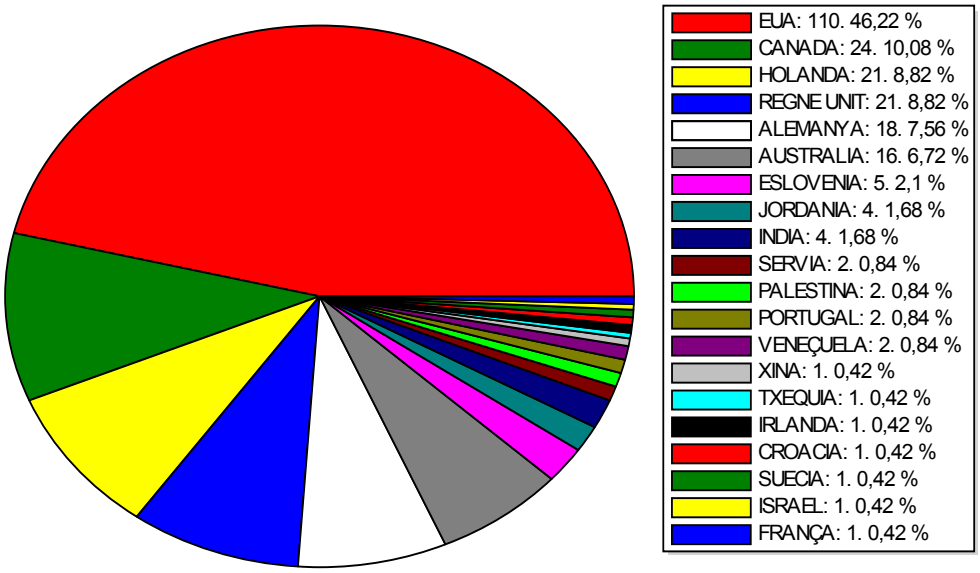

Figura 4-2. Procedència dels recursos digitals en llengua anglesa (e. p.)

En conjunt, el 72,26% dels recursos provenen de països de parla anglesa (Estats Units, Canadà, Regne Unit, Austràlia i Irlanda), cosa que estableix una relació pràcticament de 3:1 entre els recursos produïts en països de llengua anglesa i els que no.

Aquesta subdivisió del corpus en funció del lloc de publicació també pot resultar rellevant en el cas del corpus castellà, no tant per identificar textos elaborats per autors que no són de llengua materna castellana com pel fet de poder tenir en compte les diferents varietats lingüístiques d'aquesta llengua. De tota manera, les dades recollides a la taula 4-3 (pàgina 213) sobre el lloc de publicació dels recursos fan palès el fet que als recursos en llengua catalana i castellana és més habitual trobar aquesta dada. Tanmateix, tot i que de vegades no s'indica explícitament el lloc de publicació del recurs, aquesta dada sí que es fa explícita mitjançant l'adreça URL del recurs,

a) Descripció en funció de l'autor i el receptor

La majoria dels recursos textuals recollits en el corpus anglès han estat produïts per autors experts. [10]

Tipus d'autor		
Expert	711	(77,11%)
Semiexpert	204	(22,13%)
Llec	7	(0,76%)

Taula 4-4. Tipus d'autors dels recursos en llengua anglesa (e. p.)

Crida l'atenció l'existència de recursos elaborats per autors llecs, que inicialment no havia estat prevista. Es tracta de petites descripcions fetes per testimonis del fenomen dels Leònids en diferents moments històrics. Aquest fet únicament es dóna en el corpus en llengua anglesa.

Considerem que un recurs ha estat elaborat per un expert quan es produeixen qualsevol dels següents condicionants:

- L'autor s'identifica com a expert (astrònom, llicenciat en ciències físiques, doctor en astronomia o membre d'una institució de recerca).
- El recurs forma part d'un lloc web en el qual només hi poden publicar experts.
- El recurs inclou continguts teòrics. En el cas dels Leònids, entenem per contingut teòric tot allò que no es pot desprendre de l'observació directa del fenomen o de la lectura d'obres assequibles a algú no format específicament en astronomia.

Així doncs, els recursos catalogats com a produïts per semiexperts han estat tots els que no acomplien cap dels condicionants anteriors. Generalment en aquest recursos hi consta explícitament el caràcter aficionat de l'autor, o bé estan inclosos en llocs web

[10] En el cas del català, però, la majoria de recursos textuals han estat realitzats per autors semiexperts, fet que crida l'atenció.

propis d'aficionats. També acostumen a incloure continguts de caràcter metodològic, que es pot adquirir amb l'experiència tot i no haver estat format en astronomia.

Els recursos produïts per experts tenen com a receptor típic tant experts com semiexperts i llecs, en la relació següent:

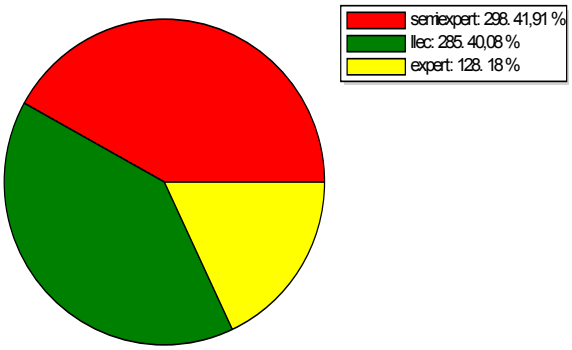

Figura 4-3. Receptors prototípics del textos produïts per experts en llengua anglesa (e. p.)

Aquesta situació es veu reflectida pràcticament igual en el corpus castellà, mentre que en català l'expert no sembla que sigui el receptor típic de cap d'aquests recursos.

Pel que fa a la distribució dels tipus de recurs elaborats per un autor expert en llengua anglesa, la relació és la següent:

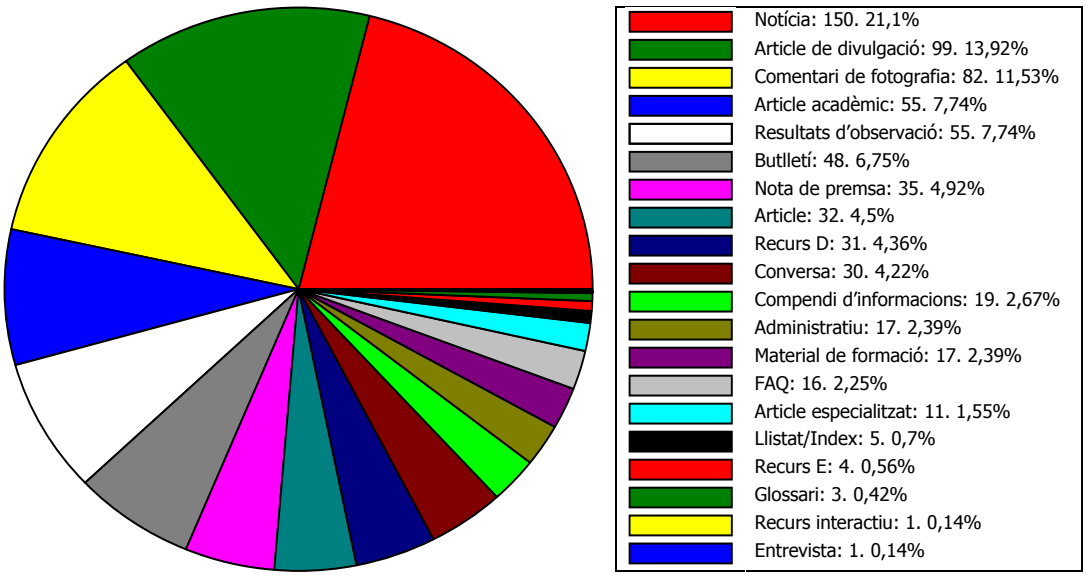

Figura 4-4. Tipus de recursos redactats per un autor expert en llengua anglesa (e. p.)

Els quatre tipus de recursos principals, que sumen un 54,29% del total, són notícies, articles de divulgació, comentaris de fotografies i articles acadèmics. En el cas del castellà i el català resulta encara més evident la utilització d'aquest mitjà per a la divulgació de coneixement, ja que els tipus de recursos més emprats són la notícia, el butlletí i l'article de divulgació.

Els receptors dels tipus de recurs més freqüents en llengua anglesa quan han estat elaborats per experts són els següents:

Figura A. Notícia

Figura B. Article de divulgació

Figura C. Comentari de fotografia

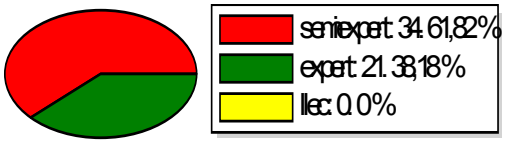

Figura D. Article acadèmic

Figura 4-5. Receptors dels tipus de recursos més habituals produïts per experts en llengua anglesa (e. p.)

Els resultats d'aquests gràfics, en què predominen clarament els receptors típics no experts, es veuen reflectits a la taula següent. Aquest fet també es pot observar en els corpus català i castellà.

Tipus de receptors dels recursos més habituals produïts per experts		
Expert	57	(14,77%)
Semiexpert	131	(33,94%)
Llec	198	(51,29%)
TOTAL	386	

Taula 4-5. Receptors dels recursos més habituals produïts per experts en llengua anglesa (e. p.)

Observant-ho des d'un punt de vista global, la major part dels recursos que habitualment redacten autors experts estan dirigits a receptors llecs (el 51,29% en el cas de l'anglès, i percentatges similars en castellà i català). El segon lloc l'ocupen els recursos adreçats a receptors semiexperts prototípics.

D'altra banda, els recursos produïts per autors semiexperts tenen com a receptor ideal en llengua anglesa semiexperts i llecs seguint la distribució següent:

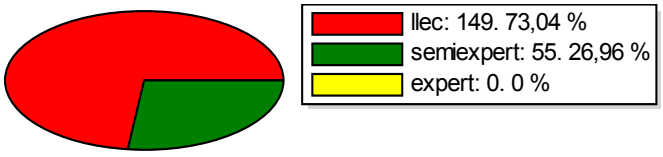

Figura 4-6. Receptors dels recursos produïts per autors semiexperts en llengua anglesa (e. p.)

Tant en el corpus castellà com el català es pot observar la mateixa tendència.

Resulta obvi, doncs, que entre el receptor de recursos redactats per autors semiexperts no es troben els experts, ja que en classificar els recursos s'identifica el receptor en funció dels coneixements mínims que ha de tenir per tal de poder-lo comprendre. En cap cas no es vol donar per suposat que un recurs produït per un autor semiexpert no pugui resultar d'interès per a un expert.

Els tipus de recursos que elaboren els autors semiexperts són els següents:

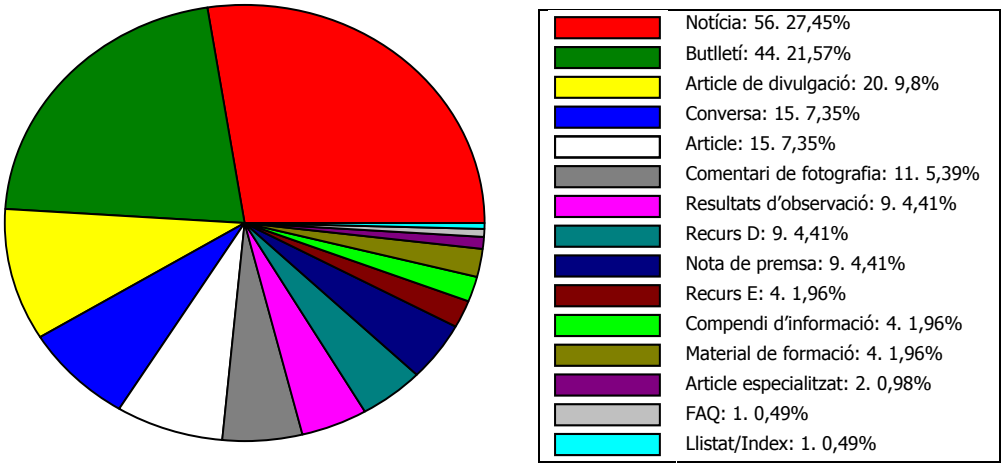

Figura 4-7. Tipus de recursos produïts per autors semiexperts en llengua anglesa (e. p.)

Com també era de preveure, entre els recursos realitzats per autors semiexperts no s'inclouen els articles acadèmics, que és un tipus de recurs reservats a autors experts. Els tres tipus de recursos més habituals són la notícia, el butlletí i l'article de divulgació, i entre tots tres representen el 58,82% del total. Aquesta tendència es manté, amb petites diferències, tant en el corpus castellà com en el català.

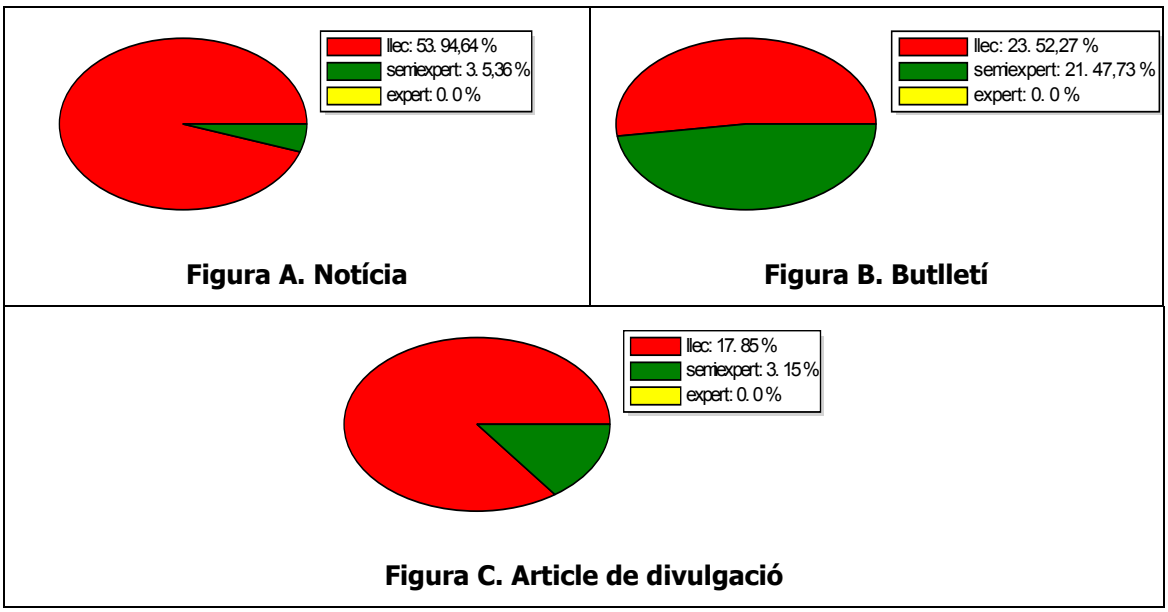

Figura 4-8. Receptors dels tipus de recursos més habituals produïts per semiexperts en llengua anglesa (e. p.)

Dels gràfics de la figura 4-8 es desprèn que la major part d'aquests recursos estan adreçats a receptors llecs. De tota manera, destaca el fet que en el cas dels butlletins la relació entre receptors llecs i semiexperts és gairebé la mateixa.

D'aquest conjunt de dades, i com a conclusió d'aquest apartat, es desprenen tres fets que resulten rellevants per a la nostra recerca:

- Els recursos elaborats per experts i dirigits a experts suposen aproximadament una cinquena dels recollits, de la qual cosa es pot deduir que els experts utilitzen aquest mitjà de comunicació, la publicació a través de pàgines web públiques, per comunicar-se principalment amb no experts.
- Els tipus de recursos que majoritàriament elaboren els experts, siguin notícies, articles de divulgació, comentaris de fotografies o butlletins (segons la llengua), fet que corrobora la hipòtesi del punt anterior tot indicant que la intenció dels experts és divulgar coneixement.

b) Descripció en funció del tipus de lloc web

En classificar els recursos textuals obtinguts hem utilitzat dues categoritzacions de llocs web: la primera els classifica en funció de l'objectiu de l'autor i la segona en funció del tipus de responsable secundari.

Llocs A	0	(0%)
Llocs B	7	(0,76%)
Llocs C	71	(7,70%)
Llocs D	745	(80,80%)
Llocs E	99	(10,74%)

Taula 4-6. Recursos en llengua anglesa segons el tipus de lloc a què pertanyen (classificació d'Alexander i Tate, 2001) (e. p.)

Com calia esperar, cap dels recursos recollits no pertany a un lloc web de tipus A, és a dir, un lloc web l'objectiu del qual sigui la defensa d'una idea, com podria ser el lloc d'un partit polític. D'altra banda, sí que hem trobat recursos que provenen de llocs web de tipus B, publicitaris o de negocis, tot i que únicament en llengua anglesa. Es tracta de recursos que contenen ressenyes de llibres publicats recentment, l'objectiu dels quals és vendre aquests llibres, ja que també inclouen un enllaç hipertextual cap a l'editorial que els publica.

Tanmateix, la major part dels recursos pertanyen a llocs de notícies (tipus C), llocs informatius (tipus D) i llocs personals (tipus E). El tipus de lloc web que més recursos aporta al nostre corpus és el D, el lloc informatiu. El seu objectiu, sobretot en l'àmbit de l'astronomia, és oferir informació objectiva sobre fets concrets; poden ser llocs molt diversos, dirigits a diferents tipus de receptors, que estructuren els arxius que els conformen en funció dels temes que es tracten o de dates en què van succeir i que, per les seves característiques, podrien considerar-se l'equivalent virtual de les revistes científiques no acadèmiques.

Alguns d'aquests llocs de tipus D estan dedicats exclusivament a l'astronomia, d'altres a la ciència en general. Gairebé a tots hi ha publicitat, i és per aquest motiu que formen part del web visible, ja que si s'haguessin de consultar per subscripció pertanyerien al web invisible i no es podria recuperar mitjançant els cercadors habituals.

Pel que fa a la classificació dels lloc web en funció del tipus de responsable secundari, la distribució dels recursos recollits és la següent:

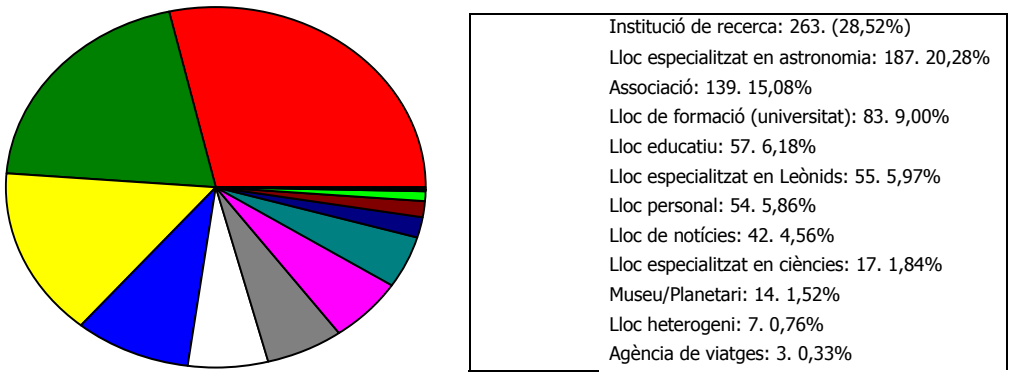

Figura 4-9. Recursos en llengua anglesa segons el tipus de lloc a què pertanyen (classificació en funció del responsable secundari) (e. p.)

Com es pot comprovar, la majoria de recursos provenen de llocs web d'institucions de recerca, llocs especialitzats en astronomia i associacions dedicades a l'astronomia, així com d'universitats en la seva vessant d'institució de formació superior. De nou, aquesta situació es dóna amb variacions poc significatives en el corpus castellà, mentre que en el català resulta molt més freqüent els recursos que formen part de llocs educatius, que sumen un total del 54,5%.

c) Descripció del corpus en funció del tipus de recurs textual

En anglès es troba la diversitat més gran de recursos textuals. La distribució en xifres d'aquests recursos, que tant en el corpus castellà com en el català resulten molt similars, és la següent:

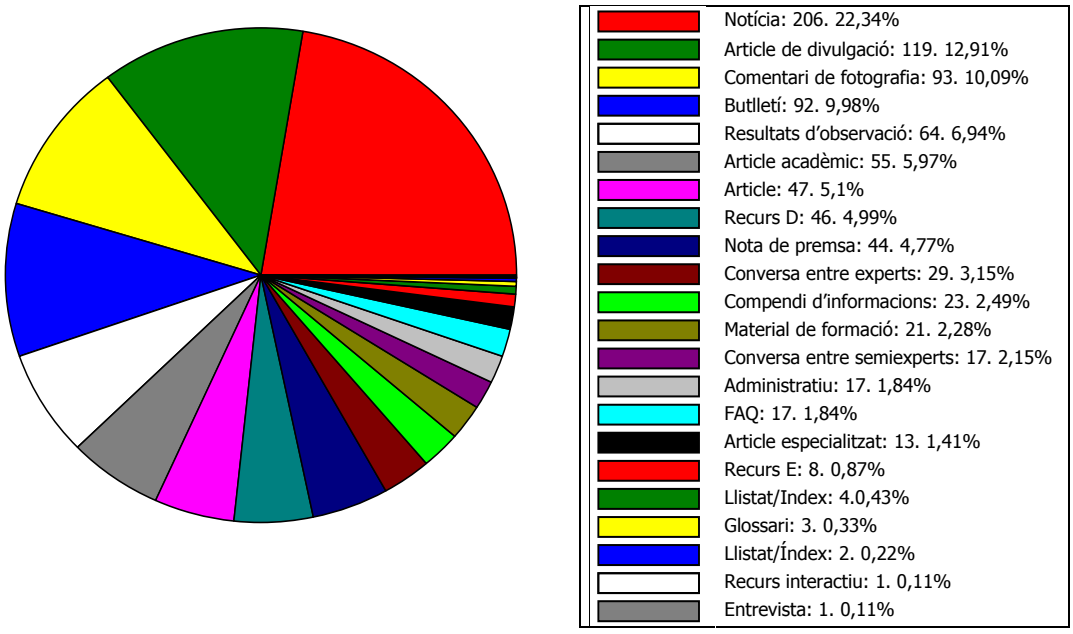

Figura 4-10. Tipus de recursos en llengua anglesa (e. p.)

Per tal d'analitzar aquestes dades en context, intentarem reproduir la situació comunicativa ideal d'on formen part tot analitzant el tipus de lloc al qual pertanyen, així com el seu autor i receptor típics, puix que, en cercar textos paral·lels, el traductor buscarà reproduir fonamentalment la situació comunicativa de la seva traducció basant-se en aquests paràmetres.

El tipus de recurs més habitual és la notícia, que hem definit com a recurs textual sincrònic als fets que presenta o a la informació que ofereix. Es tracta, doncs, d'un recurs d'informació immediata.

La major part dels recursos textuals de notícies, el 72,28% en el cas de l'anglès, han estat produïdes per experts. La distribució d'aquests recursos entre els diferents tipus de lloc web és la següent:

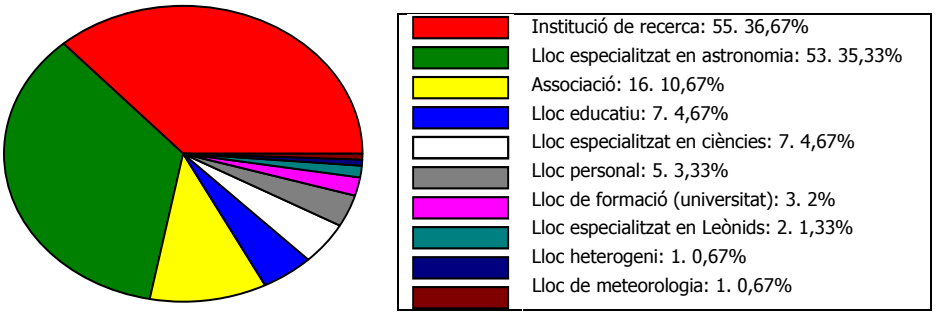

Figura 4-11. Llocs web en què apareixen notícies elaborades per experts en llengua anglesa (e. p.)

Tal com mostra el gràfic, els principals llocs web que proporcionen informació sobre fets, en aquest cas de l'àmbit de l'astronomia, de manera sincrònica són les institucions de recerca i els llocs i les associacions especialitzades en astronomia.

D'altra banda, els recursos textuals de notícies elaborats per semiexperts, el 27,18% del total, es distribueix entre els diferents tipus de llocs web de la manera següent:

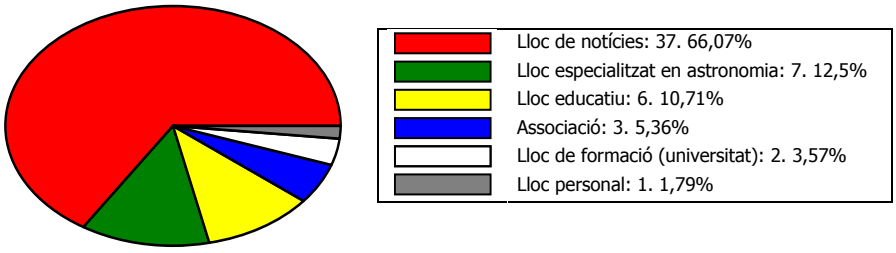

Figura 4-12. Llocs web en què apareixen notícies elaborades per experts en llengua anglesa (e. p.)

Pel que fa als altres tipus de recursos textuals més habituals, la seva situació comunicativa resulta força similar.

Tots aquests tipus de recursos textuals han estat elaborats majoritàriament per experts. Destaca especialment el fet que cap d'aquests recursos no formen part de llocs de tipus genèric, com podria ser un lloc de notícies. Tots ells pertanyen a llocs específics, ja sigui per raó de l'àmbit temàtic (ciències, astronomia o Leònids) o al tipus d'audiència a què es dirigeixen (llocs educatius o de formació superior).

Segons aquestes dades, els llocs web de notícies, doncs, resulten molt més homogenis pel que fa al tipus de recursos textuals que contenen que no pas la resta de tipus de llocs web. Dit en altres paraules, si bé és cert que en els diferents tipus de llocs es poden trobar notícies en major o menor mesura, els llocs de notícies només contenen aquest tipus de recurs. Sense haver pogut comprovar-ho de manera fefaent, suposem que l'heterogeneïtat de la resta de llocs web es deu fonamentalment als motius següents:

- No publiquen informació amb un únic objectiu, sinó que utilitzen un mateix lloc web per facilitar la comunicació entre experts (per exemple, informació sobre recerca), expert → semiexpert (notes de premsa), expert → llec (notícies), semiexpert ↔ semiexpert (conversa), semiexpert → llec (material de formació).
- Tot i tenir un únic objectiu, en no seguir el procés editorial habitual en la publicació en format paper, el resultat és molt més heterogeni del que es desitja.

De nou, el conjunt d'aquestes dades, així com les que es desprenen de l'anàlisi dels corpus català i castellà, que no presenten variacions significatives, corrobora la hipòtesi assenyalada anteriorment, segons la qual el web públic es un mitjà de comunicació utilitzat principalment per a la difusió d'informació. A més, també es pot afirmar que la difusió es fa fonamentalment de manera sincrònica al fet del qual s'informa, ja que el recurs més habitual és la notícia.

d) Descripció dels corpus en funció del contingut dels recursos textuals

A més del tipus de lloc web i del tipus de recurs textual, l'altre criteri utilitzat per classificar els recursos que formen part del corpus és el seu contingut. En analitzar cadascun dels recursos per tal de classificar-los es van detectar certes regularitats temàtiques que van ser agrupades en els següents onze blocs:

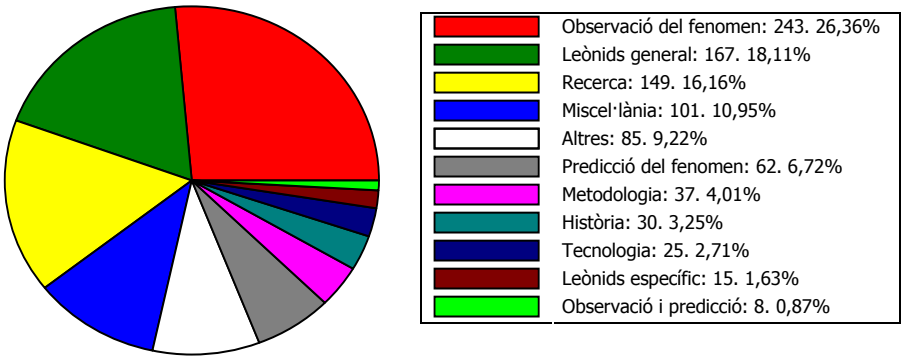

Figura 4-13. Els recursos en llengua anglesa segons el seu contingut (e. p.)

Aquest paràmetre de classificació s'ha tingut en compte únicament per comprovar si es dóna una correlació entre el contingut i el tipus de recurs o lloc web i completar així l'anàlisi dels corpus recollits.

CONTINGUT / TIPUS DE RECURS	Història	Leònids específic	Leònids general	Metodologia	Miscel·lània	Observació	Observació i predicció	Predicció	Recerca	Tecnologia	Altres	
Administratiu	...	...	...	2	**15**	...	...	...	...	...	...	**17**
Article	2	3	7	3	1	10	2	6	1	...	**12**	**47**
Art. acadèmic	2	...	...	...	...	3	...	2	**44**	...	4	**55**
Art. divulgació	*10*	2	***42***	5	7	9	2	9	5	8	*20*	**119**
Art. especialitzat	1	2	...	1	1	2	...	3	...	...	3	**13**
Butlletí	...	...	5	...	***59***	17	...	3	2	...	6	**92**
Comentari fotografia	6	...	37	5	...	38	...	5	2	...	...	**93**
Compendi d'informacions	...	...	4	...	**10**	3	...	1	3	1	1	**23**
Conversa	...	...	1	...	...	**43**	...	...	2	...	...	**46**
Entrevista	...	...	...	...	...	**1**	...	...	...	...	...	**1**
FAQ	...	...	**12**	1	...	...	...	...	1	1	2	**17**
Glossari	...	...	...	...	...	...	...	...	...	...	**3**	**3**
Llistat/Índex	1	...	**3**	...	1	...	...	...	...	...	1	**6**
Material formació	...	...	5	...	...	6	...	1	...	...	**9**	**21**
Nota de premsa	...	...	8	...	3	6	1	8	**13**	2	3	**44**
Notícia	2	*6*	38	8	2	42	*3*	*22*	***56***	*13*	14	**206**
Recurs D	6	2	4	*10*	1	6	...	1	**11**	...	5	**46**
Recurs E	...	...	1	2	1	2	...	1	...	...	1	**8**
Recurs interactiu	...	...	...	...	...	...	...	...	**1**	...	...	**1**
Resultats d'observació	...	...	...	...	...	***55***	...	...	8	...	1	**64**
	30	*15*	*167*	*37*	*101*	*243*	*8*	*62*	*149*	*25*	*85*	***922***

Taula 4-7. Relació tipus de recurs / contingut en llengua anglesa (e. p.)[11]

Lataula 4-7 presenta certes regularitats pel que fa a la relació que es dóna entre tipus de recurs i contingut, de les quals comentarem les més destacables:

- Hi ha una correlació elevada entre els següents tipus de recurs i continguts:
  - **Butlletí**, contingut **miscel·lània**: el butlletí en format paper és una publicació en la qual es tracta més d'un tema, i aquests sovint pertanyen al mateix àmbit temàtic o comparteixen alguna característica (tenen lloc a una ciutat concreta, estan vinculats a una mateixa institució, o qualsevol altre nexe comú). El butlletí en format digital també presenta aquesta característica, fins al punt que, en els casos que hem recollit, no es divideix formalment en arxius independents, sinó que manté el disseny del butlletí tradicional.

[11] En aquesta taula s'assenyala amb negreta el contingut amb major presència en un tipus de recurs, i amb cursiva es marca el tipus de recurs textual que més tracta un contingut determinat.

- **Resultat d'observació**, contingut **observació del fenomen**: si bé pot resultar obvi que el tipus de recurs que resulta de l'observació recull informació sobre l'observació del fenomen, no ho és tant el fet invers. Sobre l'observació dels Leònids es parla en recursos de diferent tipus, com ara en notícies o comentaris de fotografies, però el recurs en què més s'aborda aquest tema de manera específica és el resultat d'observació.
- **Article de divulgació**, contingut **Leònids general**: la funció principal de l'article de divulgació pràcticament l'obliga a tractar el tema d'una manera genèrica, pràcticament introductòria, que no permet que pugui estar dedicat a una perspectiva concreta.
- **Notícia**, contingut **recerca**: tot i que la majoria d'articles acadèmics estan dedicats a temes relacionats amb la recerca, la majoria dels recursos amb continguts vinculats a aquest tema són notícies, moltes de les quals formen part de llocs web especialitzats o pertanyents a institucions de recerca. Aquest fet posa en relleu el caràcter de publicació immediata de la notícia.

✸ La vinculació d'un tipus de recurs a un contingut determinat es dóna amb claredat únicament en tres casos:

- **Article acadèmic**, principalment dedicat a **recerca**: l'article acadèmic, que podríem caracteritzar com a producte d'una major reflexió que la notícia, deixa de ser funcional en l'àmbit del web públic en benefici de la notícia; per aquest motiu arribem a la conclusió que els articles acadèmics publicats a Internet no van ser creats pensant a publicar-los en aquest mitjà, sinó que un cop utilitzats en el seu context natural (com a ponència en un congrés o com a article en una publicació dirigida exclusivament a experts, en format paper o digital) són reciclats per part dels seus autors o s'ofereixen a un públic més ampli.
- **Conversa**, principalment dedicada a l'**observació del fenomen**: aquest tipus de recurs, que entenem com la reproducció escrita i amb una sola veu de la conversa espontània, està dedicat majoritàriament a comentar l'activitat dels experts i semiexperts en aquest àmbit, que és

l'observació directa del fenomen astronòmic. Per aquest motiu, aquesta relació, que en analitzar-la pot no resultar gaire sorprenent, sí que ens permet detectar certa correlació entre l'activitat dels experts i el tipus de recurs més espontani i allunyat de la publicació en format paper, seguint el procés editorial tradicional.

- **Compendi d'informació**, principalment dedicat a **miscel·lània**: aquest recurs temàtic comparteix moltes característiques amb el butlletí, tret que el nexe d'unió entre temes que inclou pot no ser tan explícit com en el cas d'aquest segon recurs. Tot i així, la naturalesa del recurs compendi d'informació fa que principalment no estigui dedicat a un tema en concret, sinó que el conformin una miscel·lània de temes.

La informació que aporta la taula 4-7 (pàgina 226) es completa amb el contingut de la taula 4-9 inclosa al final d'aquest capítol (pàgina 233), que aborda la mateixa relació indicant valors relatius que resulten més fàcils de comparar.

La segona comparativa realitzada observa els recursos a partir del seu contingut i del lloc web a què pertanyen. Els resultats estan recollits a la taula 4-8:

CONTINGUT / TIPUS DE LLOC WEB	Història	Leònids específic	Leònids general	Metodologia	Miscel·lània	Observació	Observació i predicció	Predicció	Recerca	Tecnologia	Altres	
Agència de viatges	...	...	...	...	1	1	...	...	...	...	1	**3**
Associació	5	2	13	8	***45***	40	1	5	6	...	14	**139**
Institució de recerca	3	*8*	*52*	4	7	**69**	*3*	*17*	*67*	10	*23*	**263**
Lloc de formació (universitat)	5	...	**20**	...	18	15	1	9	4	1	10	**83**
Lloc de notícies	...	...	**15**	1	1	12	...	7	3	2	1	**42**
Lloc de meteorologia	...	...	1	...	...	...	...	...	...	...	...	**1**
Lloc educatiu	1	...	**25**	3	3	6	1	7	...	1	10	**57**
Lloc especialitzat en astronomia	*10*	4	17	*14*	18	***73***	1	11	15	8	16	**187**
Lloc especialitzat en ciències	2	...	**5**	1	...	2	...	2	2	2	1	**17**
Lloc especialitzat en els Leònids	1	...	1	3	2	1	...	...	**46**	...	1	**55**
Lloc heterogeni	1	...	1	...	...	2	...	1	2	...	...	**7**
Lloc personal	2	1	13	2	1	**20**	1	3	4	1	6	**54**
Museu/Planetari	...	...	4	1	**5**	2	...	...	...	...	2	**14**
	30	*15*	*167*	*37*	*101*	*243*	*8*	*62*	*149*	*25*	*85*	***922***

Taula 4-8. Relació tipus de lloc / contingut en llengua anglesa (e. p.)[12]

De les dades d'aquesta taula es desprenen les conclusions següents:

- Hi ha una correlació elevada entre els següents tipus de lloc web i continguts:
  - **Associació**, contingut **miscel·lània**: si la correlació entre butlletí i miscel·lània podia resultar òbvia, el mateix succeix amb aqueta correlació. Les associacions d'amateurs es dediquen a activitats molt diverses, per la qual cosa els seus llocs també ho són.
  - **Lloc especialtizat en astronomia**, contingut **observació del fenòmen**: aquesta correlació, tot i que no resulta innovadora, és difícil de garantir atès que no podem comprovar la correlació inversa.
- Vinculació d'un tipus de lloc web a un contingut:

[12] En aquesta taula s'assenyala amb negreta el contingut amb major presència en un tipus de lloc web, i amb cursiva es marca el tipus de lloc en què es tracta principalment un contingut determinat.

- **Institució de recerca** i **lloc personal**, principalment dedicats a l'**observació del fenòmen**: com a activitat pròpia d'experts i semiexperts, l'observació dels Leònids és el contingut més habitual de pàgines personals i dels llocs web en què publiquen els experts, com són els de les institucions de recerca. Aquest és el contingut que en major mesura es dóna en aquests tipus de lloc web; tanmateix, però, la seva heterogeneïtat temàtica és molt alta, per la qual cosa la vinculació no resulta gaire estable

✸ Vinculació d'un contingut a un tipus de lloc web:

- **Recerca**, es troba principalment en llocs d'**institucions de recerca**: aquesta és una vinculació entre contingut i lloc web que calia esperar. Aquesta mena de relacions entre lloc web i contingut es poden preveure de manera intuïtiva; per aquest motiu acostumen a ser la primera opció de documentació dels traductors.

De nou, les dades absolutes recollides a la taula 4-8 es completen amb les dades relatives de la taula 4-10. "Relació tipus de lloc / contingut en llengua anglesa", que s'inclou al final del aquest capítol. Altrament, les dades relatives als corpus castellà i català es poden trobar a l'annex B, pàgina 618.

L'anàlisi descriptiva dels corpus realitzat en aquest apartat ens ha permès conèixer més profundament algunes de les característiques externes dels recursos textuals recollits. Més endavant, sobretot al capítol 6 (pàgina 283) dedicat a l'anàlisi dels corpus, utilitzarem aquests paràmetres per subidividir cadascun dels corpus en conjunts de recursos textuals homogenis pel que fa a les seves característiques externes i que, per tant, poden resultar més apropiats com a textos paral·lels d'una traducció determinada.

Finalment, tal com s'ha indicat a l'apartat 1.1.2 del capítol 3 "Tipus de recursos web i qualitat de la informació" (pàgina 127), un dels aspectes més rellevants de la publicació d'informació a Internet és l'absència d'indicadors que garanteixin qualitat dels recursos, com ho són per exemple el responsable principal i la data de producció i/o revisió de

cada recurs. Aquest fet també es posa de manifest a la taula 4-3. "Informació administrativa que consta al recurs" (pàgina 213), en la qual s'indica, per exemple, que únicament al 51,7% de tots els recursos s'assenyala explícitament el nom del seu responsable principal.

Tanmateix, tal com intentarem demostrar en el capítol 6, pensem que, si bé els recursos textuals que no presenten els indicadors necessaris per garantir la seva qualitat no poden ser prou fiables de forma individual per al traductor, en conjunt sí que poden ser-ho. En altres paraules, pensem que el criteri de quantitat inherent a la metodologia d'obtenció d'informació que proposarem ens garantirà la qualitat del conjunt, és a dir, de cadascun dels corpus, tot i que els recursos que en formen part no presentin els indicadors necessaris que la garanteixen.

Abans d'encetar l'anàlisi dels recursos recollits, i de revisar els elements d'anàlisi més apropiats que provenen de la lingüística de corpus, en aquest capítol hem observat els paràmetres que donen lloc als diferents tipus de corpus. Després d'analitzar les diferents tipologies de corpus, hem presentat els paràmetres de classificació dels recursos textuals digitals especialitzats que integren el nostre corpus. En especial, hem prestat especial atenció a criteris de classificació relatius a l'autor, al lloc web i directament al recurs textual. Els criteris de classificació emprats, a partir dels quals s'han descrit els recursos recollits, ens permetran dividir el corpus compilat en subcorpus de recursos amb característiques extralingüístiques comunes, i analitzar-los per separat.

La descripció dels recursos també ens ha permès conèixer amb més profunditat els corpus monolingües que conjuntament integren aquest corpus multilingüe comparable especialitzat *ad hoc*. En analitzar cada corpus monolingüe per separat, l'aspecte més destacable és el gran desequilibri quantitatiu que hi ha entre el corpus anglès i els altres dos corpus, el català i el castellà. En ser molt més voluminós, l'anglès també presenta un major nivell d'heterogeneïtat, tant pel que fa als tipus de recursos que l'integren com a les circumstàncies comunicatives que els envolten.

	Història		Leònids específic		Leònids general		Metodologia		Miscel·lània		Observació		Observació i predicció		Predicció		Recerca		Tecnologia		Altres	
Administratiu	**...**		**...**		**...**		**11,8%**		**88,2%**		**...**		**...**		**...**		**...**		**...**		**...**	
		...		...		...		5,4%		15%		...		...		...		...		...		...
Article	**4,2%**		**6,4%**		**14,9%**		**6,4%**		**2,1%**		**21,2%**		**4,3%**		**12,8%**		**2,1%**				**25,5%**	
		6,7%		20%		4,2%		8,1%		0,9%		4,1%		25%		9,7%		0,7%				14,1%
Art. acadèmic	**3,6%**		**...**		**...**		**...**		**...**		**5,4%**		**...**		**3,6%**		**80%**		**...**		**7,3%**	
		6,7%		...		...		...		...		1,2%		...		3,2%		29,5%		...		4,7%
Art. divulgatiu	**8,4%**		**1,7%**		**35,3%**		**4,2%**		**5,9%**		**7,6%**		**1,7%**		**7,6%**		**4,2%**		**6,7%**		**16,8%**	
		33,3%		13,3%		25,1%		13,5%		6,9%		3,7%		25%		14,5%		3,4%		32%		23,5%
Art. especialitzat	**7,7%**		**15,4%**		**...**		**7,7%**		**7,7%**		**15,4%**		**...**		**23,1%**		**...**		**...**		**23,1%**	
		3,3%		13,3%		...		2,7%		0,9%		0,8%		...		4,8%		...		...		3,5%
Butlletí	**...**		**...**		**5,4%**		**...**		**64,1%**		**18,5%**		**...**		**3,3%**		**2,2%**		**...**		**6,5%**	
		...		...		2,9%		...		58,4%		7%		...		4,8%		1,3%		...		7,1%
Comentari de fotografia	**6,4%**		**...**		**39,8%**		**5,4%**		**...**		**40,9%**		**...**		**5,4%**		**2,1%**		**...**		**...**	
		20%		...		22,7%		22,7%		...		15,6%		...		8,1%		1,4%		...		...
Compendi d'informacions	**...**		**...**		**17,4%**		**...**		**43,5%**		**13%**		**...**		**4,3%**		**13%**		**4,3%**		**4,3%**	
		...		...		2,4%		...		9,9%		1,2%		...		1,6%		2%		4%		1,2%
Conversa	**...**		**...**		**2,2%**		**...**		**...**		**93,5%**		**...**		**...**		**4,3%**		**...**		**...**	
		...		...		0,6%		...		...		17,7%		...		...		1,3%		...		...
Entrevista	**...**		**...**		**...**		**...**		**...**		**100%**		**...**		**...**		**...**		**...**		**...**	
		...		...		...		...		...		0,4%		...		...		...		...		...
FAQ	**...**		**...**		**70,6%**		**5,9%**		**...**		**...**		**...**		**...**		**5,9%**		**5,9%**		**11,8%**	
		...		...		7,2%		7,2%		...		...		...		...		0,7%		4%		2,35%
Glossari	**...**		**...**		**...**		**...**		**...**		**...**		**...**		**...**		**...**		**...**		**100%**	
		...		...		...		...		...		...		...		...		...		...		3,5%
Llistat/Índex	**16,7%**		**...**		**50%**		**...**		**16,7%**		**...**		**...**		**...**		**...**		**...**		**16,7%**	
		3,3%		...		1,8%		...		0,4%		...		...		...		...		...		1,2%
Material de formació	**...**		**...**		**23,8%**		**...**		**...**		**28,6%**		**...**		**4,8%**		**...**		**...**		**42,9%**	
		...		...		2,9%		...		...		2,5%		...		1,6%		...		...		10,6%
Nota de premsa	**...**		**...**		**18,2%**		**...**		**6,8%**		**13,6%**		**...**		**18,2%**		**29,5%**		**4,5%**		**6,8%**	
		...		...		4,8%		...		2,5%		2,5%		...		12,9%		8,7%		8%		3,5%
Notícia	**0,9%**		**2,9%**		**18,4%**		**3,9%**		**0,9%**		**20,4%**		**...**		**10,7%**		**27,2%**		**6,3%**		**6,8%**	
		6,7%		40%		22,75%		21,6%		17,3%		17,3%		...		35,5%		37,6%		54%		16,5%
Recurs D	**13%**		**13%**		**8,7%**		**21´7%**		**2,2%**		**13%**		**2,3%**		**2,2%**		**23,9%**		**...**		**10,9%**	
		20%		13,3%		2,4%		27,1%		2,5%		2,5%		12,5%		1,6%		7,4%		...		5,9%
Recurs E	**...**		**...**		**12,5%**		**25%**		**12,5%**		**25%**		**1,5%**		**12,5%**		**...**		**...**		**12,5%**	
		...		...		0,6%		5,4%		0,9%		0,8%		37,5%		1,6%		...		...		1,2%
Recurs interactiu	**...**		**...**		**...**		**...**		**...**		**...**		**...**		**...**		**100%**		**...**		**...**	
		...		...		...		...		...		...		...		...		0,7%		...		...
Resultats d'observació	**...**		**...**		**...**		**...**		**...**		**85,9%**		**...**		**...**		**12,5%**		**...**		**1,6%**	
		...		...		...		...		...		22,6%		...		...		5,37%		...		1,2%

Taula 4-9. Relació tipus de recurs / contingut en llengua anglesa (valors relatius) (e. p.)

	Història		Leònids específic		Leònids general		Metodologia		Miscel·lània		Observació		Observació i predicció		Predicció		Recerca		Tecnologia		Altres	
Agència de viatges	**...**	...	**...**	...	**...**	...	**...**	...	**33,3%**	1%	**33,3%**	0,4%	**...**	...	**...**	...	**...**	...	**...**	...	**33,3%**	1,2%
Associació	**3,6%**	16,7%	**1,4%**	13,3%	**9,3%**	7,8%	**5,8%**	21,6%	**32,4%**	44,5%	**28,8%**	16,5%	**0,7%**	12,5%	**3,6%**	8,1%	**4,3%**	4,1%	**...**	...	**10,1%**	16,5%
Institució de recerca	**1,1%**	10%	**3,1%**	53,3%	**19,8%**	31,2%	**1,5%**	10,8%	**2,7%**	6,9%	**26,3%**	28,4%	**1,1%**	37,5%	**6,5%**	27,4%	**25,5%**	44,9%	**3,8%**	40%	**8,7%**	27,1%
Lloc de formació (universitat)	**6,1%**	16,7%	**...**	...	**24,1%**	11,9%	**...**	...	**21,7%**	17,8%	**18,1%**	6,2%	**1,2%**	12,5%	**10,8%**	14,5%	**4,8%**	2,7%	**1,2%**	4%	**12,1%**	11,7%
Lloc de meteorologia	**...**	...	**...**	...	**100%**	0,6%	**...**	...	**...**	...	**...**	...	**...**	...	**...**	...	**...**	...	**...**	...	**...**	...
Lloc de notícies	**...**	...	**...**	...	**35,7%**	9%	**2,4%**	2,7%	**2,4%**	1%	**28,6%**	4,9%	**...**	...	**16,7%**	11,3%	**7,1%**	2%	**4,8%**	8%	**2,4%**	1,2%
Lloc educatiu	**1,7%**	3,3%	**...**	...	**43,9%**	14,9%	**5,3%**	8,1%	**5,3%**	2,9%	**10,5%**	2,5%	**1,7%**	12,5%	**12,3%**	11,3%	**...**	...	**1,7%**	4%	**17,5%**	11,8%
Lloc especialitzat en astronomia	**5,3%**	33,3%	**2,1%**	26,7%	**9,1%**	10,2%	**7,5%**	37,8%	**9,6%**	17,8%	**39,1%**	30%	**0,5%**	12,5%	**5,9%**	17,7%	**8,1%**	10,1%	**4,3%**	32%	**8,6%**	18,8%
Lloc especialitzat en ciències	**11,8%**	6,7%	**...**	...	**29,4%**	2,9%	**5,9%**	2,7%	**...**	...	**11,8%**	0,8%	**...**	...	**11,8%**	3,2%	**11,8%**	1,3%	**11,8%**	8%	**5,9%**	1,2%
Lloc especialitzat en Leònids	**1,8%**	3,3%	**...**	...	**1,8%**	0,6%	**5,5%**	8,1%	**3,6%**	1,9%	**1,8%**	0,4%	**...**	...	**...**	...	**83,6%**	30,9%	**...**	...	**1,8%**	1,2%
Lloc heterogeni	**14,3%**	3,3%	**...**	...	**14,3%**	0,6%	**...**	...	**...**	...	**28,6%**	0,8%	**...**	...	**14,3%**	1,6%	**28,6%**	1,3%	**...**	...	**...**	...
Lloc personal	**3,7%**	6,7%	**1,8%**	6,7%	**24,1%**	7,8%	**3,7%**	5,4%	**1,8%**	1%	**37,1%**	8,2%	**1,8%**	12,5%	**5,6%**	4,8%	**7,4%**	2,7%	**1,8%**	4%	**11,1%**	7,1%
Museu/Planetari	**...**	...	**...**	...	**28,6%**	2,4%	**7,1%**	2,7%	**35,7%**	4,9%	**14,3%**	0,8%	**...**	...	**...**	...	**...**	...	**...**	...	**14,3%**	2,3%

Taula 4-10. Relació tipus de lloc / contingut en llengua anglesa (e. p.)