Jaume Martí i Marina Salse (coord.)
La terminologia i la documentació: relacions i sinergies
Barcelona: Institut d'Estudis Catalans, 2010, p. 47-73 DOI: 10.2436/15.2503.02.2

Invitació a l'estudi estadístic del llenguatge

Rogelio Nazar
Institut Universitari de Lingüística Aplicada
Universitat Pompeu Fabra
Barcelona

Resum

El tema d'aquesta presentació és la cruïlla interdisciplinària entre la lingüística i l'estadística. Està adreçada a lingüistes, per als quals pot tenir un interès teòric, o a professionals que treballen amb la llengua, per als quals pot tenir un interès pràctic. Enfoca el concepte de probabilitat de combinatòria de paraules des de tres perspectives diferents: *a*) els estudis d'associació entre les unitats que es combinen, *b*) la distribució en el corpus d'aquesta combinació d'unitats, i, finalment, *c*) les maneres de mesurar la similitud entre unitats d'acord amb les seves possibilitats de combinació. Tots aquests temes hi són tractats d'una manera estrictament teòrica i van acompanyats d'exemples d'aplicació pràctica en terminologia i en documentació. L'objectiu és demostrar que la utilització d'eines estadístiques en aquests camps és un complement necessari per a la intuïció dels investigadors.

Paraules clau: corpus textuals, estadística, lingüística quantitativa, llenguatge, probabilitat combinatòria.

Abstract: *Invitation to the statistical study of language*

The topic of this presentation is the interdisciplinary nexus between linguistics and statistics. It targets linguists, for whom it may have a theoretical interest, or professionals that work with language, for whom it may have a practical interest. It focuses on the concept of the combinatory probability of words from three different perspectives: *a*) the studies of association between the units that are combined, *b*) the distribution of this combination of units in the corpus, and finally *c*) the ways of measuring similarity between units according to the combination possibilities. All these topics are addressed in a strictly theoretical fashion and are illustrated by examples of practical application in terminology and in documentation. The objective is to demonstrate that the use of statistical tools in these fields is a necessary complement to the researcher's intuition.

KEY WORDS: text corpora, statistics, quantitative linguistics, language, combinatorial probability.

1. INTRODUCCIÓ

Aquesta comunicació està dirigida a persones que no tenen coneixements previs sobre l'encreuament interdisciplinari entre l'estadística i el llenguatge. Té el doble propòsit de ser una aportació des del punt de vista de la lingüística teòrica i, a la vegada, ser útil per a la documentació i per a la terminologia. Consegüentment, inclou exemples de com aquest coneixement teòric es pot aplicar a la solució de problemes pràctics.

La intenció és introduir a la temàtica, però també conscienciar i aclarir. Conscienciar, perquè la lingüística quantitativa no només és una àrea marginal en lingüística, sinó que, a més, moltes vegades tant lingüistes com estadístics n'ignoren l'existència. Aclarir, perquè la relació entre estadística i llengua no és cap novetat ni pertany al món de les «noves tecnologies». Estem parlant d'una tradició que fa més de cinquanta anys que difon conceptes i mètodes que no tenen una relació inherent amb la informàtica. La utilització d'ordinadors és evidentment necessària per a dur a terme estudis en lingüística quantitativa, però parlar d'aquests temes no significa parlar d'un programa informàtic, perquè això equival a confondre el fenomen observat amb l'instrument d'observació. Certament, els mitjans són determinants, ja que, com deia Saussure, el punt de vista defineix l'objecte. Tanmateix, això no ha de dur a l'error de reificar les idees en la forma d'un programari. En definitiva, l'important és conèixer quins estudis s'han fet o es poden fer i prendre consciència que aquesta disciplina no es limita al recompte de vegades que dues paraules apareixen juntes en un corpus.

Pel que fa a la meva legitimitat com a orador, sóc aquí per la meva funció a l'IULA,[1] consistent a assimilar el coneixement que ja existeix sobre lingüística quantitativa, aplicar aquest coneixement a la solució de problemes pràctics i, a la vegada, intentar proposar algun coneixement nou en els fòrums científics.

No presento res de nou en aquesta comunicació. Faré, en canvi, un recorregut per algunes idees que he tractat ja en altres treballs. És important advertir que no represento necessàriament l'opinió dels meus companys de feina. Em refereixo particularment a un protocol que inclou un compromís amb la independència de llengua, que consisteix a esbrinar primer, i sempre que sigui possible, fins a quin punt es pot arribar a treure conclusions útils sense introduir coneixement explícit sobre una llengua en particular.

1. Institut Universitari de Lingüística Aplicada (*http://www.iula.upf.edu*).

Aquesta comunicació està organitzada de la manera següent: en la pròxima secció 2, analitzarem la confrontació existent entre dues formes molt diferents d'apropar-se a l'estudi de la llengua, davant de les quals la lingüística es troba en una posició ambigua: el món humanístic, per anomenar-lo d'alguna manera encara que sembli lleugerament imprecís, i el món científic, en particular el món de les «ciències dures» per oposició a les ciències socials, on el pensament quantitatiu és, a vegades, encara sospitós. A continuació, en la secció 3, entrarem en la matèria de l'anàlisi lingüística enfocada des de la perspectiva estadística. Analitzarem concretament el concepte de combinatòria de paraules. Veurem el concepte de probabilitat de combinatòria de paraules des de tres perspectives diferents: en la subsecció 3.1, els estudis d'associació entre les unitats que es combinen; en la subsecció 3.2, la manera en què aquesta combinació d'unitats es distribueix en un corpus i les conclusions que en podem derivar; i, finalment, en la subsecció 3.3, les formes de mesurar la similitud entre unitats d'acord amb les seves possibilitats de combinació. Com a exemple, analitzarem el bigrama i establirem el significat d'aquesta unitat més enllà de la seva definició formal, per saber en profunditat quin tipus d'informació codifica. Veurem que, encara que sembli sorprenent, la nostra identitat individual i col·lectiva està continguda en el bigrama. Com a exemple de les aplicacions pràctiques, en la secció 4 veurem la classificació de documents en diverses variants —subsecció 4.1 i 4.2—, així com elements per a la caracterització del significat i la desambiguació de terminologia i, en la subsecció 4.3, el descobriment de neologia. Existeixen altres possibilitats d'aplicació, entre les quals trobem línies de recerca en curs, com ara l'extracció automàtica de terminologia especialitzada o l'extracció de terminologia bilingüe de corpus no paral·lels, però aquestes línies, malgrat el seu interès, no es tractaran aquí per les limitacions d'espai.

2. El xoc entre dues cultures

Wilhelm Dilthey (1883) va advertir ja les diferències epistemològiques entre les ciències naturals, d'una banda, i les ciències socials i humanitats (o ciències de l'esperit), de l'altra, continuant una línia de pensament que va iniciar Kant. Mentre que en les ciències naturals preval un pensament mecanicista, amb el qual es pot predir la conseqüència de determinats esdeveniments, en les ciències de l'esperit, en canvi, aquest determinisme no és possible. La resposta d'un ésser humà davant d'un determinat esdeveniment és en última instància imprevisible. Fins i tot en aquestes circumstàncies, les ciències de l'esperit ens permeten almenys comprendre (*verstehen*) les circumstàncies històriques i individuals que envolten el que és humà.

La fita, però, en la història de la presa de consciència de la divisió de la cultura en el saber científic i el saber humanístic —divisió que encara estructura els cur-

rículums de l'educació secundària— és una conferència donada per C. P. Snow (1959), en la qual descriu la sospita mútua i la incomprensió existent entre científics i intel·lectuals. Tot i que pertanyin a les capes més educades de la població, els dos col·lectius són ignorants l'un de l'altre. Si bé després va moderar el seu discurs, en aquella ocasió Snow va plantejar que la gent que té un pensament de tipus tècnic és en general inculta, i els intel·lectuals, per la seva part, hostils a aquest pensament, són generalment incapaços de comprendre els conceptes científics més elementals.

Aquesta separació és particularment interessant en el si de les ciències socials, considerades «ciències toves» per oposició al rigor de les ciències naturals, les «ciències dures». La inclinació dels científics socials per una o per una altra branca de pensament dependrà de l'orientació ideològica personal o de la de cada facultat o departament, però entre els intel·lectuals de les ciències socials és comú advertir una reticència *a priori* cap a tot pensament de tipus tècnic en l'estudi del que és humà. Aquesta reticència està representada en la idea de Cornelius Castoriadis (1975) sobre el fet que amb un llenguatge reduït a allò que és instrumental es pot operar i calcular, però no es pot pensar, una idea amb ressonàncies a la polèmica constatació feta per Heidegger sobre la idea que «la ciència no pensa».

En sociologia, aquesta diferència va estar clarament representada per l'oposició entre el pensament crític i la reflexió filosòfica i històrica de l'Escola de Frankfurt davant l'hàbit dels sociòlegs nord-americans de la Mass Communication Research de promoure l'aplicació de mètodes quantitatius per sobre de la reflexió teòrica, enfrontament que va continuar tot i la col·laboració entre alguns dels màxims exponents d'ambdós bàndols, com Theodor Adorno i Paul Lazarsfeld.

El cas és particularment interessant en la lingüística, si es vol, «la més dura de les ciències toves». Fins i tot lingüistes experimentats expressen sorpresa en prendre consciència que existeix una lingüística quantitativa. Els que són «de lletres» no saben «de nombres». Mandelbrot (1961) encara estava en el moment oportú per a revitalitzar la pregunta sobre què és la lingüística i establir una diferència entre gramàtics i lingüistes. En el cas dels primers, preval el coneixement d'una llengua en particular i del que pot ser i el que no pot ser gramaticalment correcte; mentre que, segons aquest autor, la lingüística pertany al món de les ciències dures, i, en aquest sentit, l'important no són tant les característiques particulars, que són d'una infinita diversitat, sinó les propietats estructurals del llenguatge (actitud contra la qual Saussure segurament no tindria res a dir). L'estudi d'aquestes propietats possibilita enunciats científics amb una validesa que transcendeix el coneixement que es tingui d'una llengua en particular, la qual cosa està d'acord amb l'esperit científic que és procliu a la generalització, ja que no hi ha o no hi hauria d'haver ciència del que és particular.

L'encreuament interdisciplinari, però, és difícil. Les persones que venim d'àmbits més propers a la lingüística en general estem poc informats sobre els conceptes matemàtics més elementals i resulta laboriós començar de zero en el camp, sobretot per a qui no té els hàbits de pensament de les ciències dures. Tanmateix, aquest és, sens dubte, un camp d'estudi que justifica el desafiament; per això, per mitjà d'aquesta presentació, pretenc contagiar l'interès i aportar arguments a la confusió de les barreres entre ciències dures i toves, o entre coneixement científic i coneixement humanístic en general.

Aquestes barreres ja es confonen i la lingüística no n'és l'únic exemple. La teoria literària, món humanístic per antonomàsia, comença a patir també el setge de l'estadística. Un exemple n'és l'aportació que l'estadística està fent en les disputes sobre l'autoria d'obres literàries, en casos que inclouen figures prominents com la de Shakespeare (Vickers, 2002).

3. La informació com a probabilitat

En la línia de Shannon (1948) podem estimar la *quantitat d'informació* com la probabilitat d'ocurrència d'un signe en un missatge, una mesura de la quantitat de sorpresa que ens pot provocar un determinat esdeveniment. Per explicar-ho amb paraules senzilles, en determinats contextos sabem que hi ha esdeveniments que són més o menys normals i d'altres, inesperats. En el llenguatge hi ha certes concatenacions que són més predictibles que d'altres. Si cada dia, en sortir de la feina, el cap diu «fins demà» al treballador, després d'una sèrie d'esdeveniments d'aquest tipus l'enunciat resulta poc informatiu. Però si un determinat dia el text canvia per «aquesta empresa ja no seguirà comptant amb els seus serveis», direm que aquest segon enunciat és comparativament més informatiu, és a dir, causa major sorpresa. Aquesta sorpresa està directament relacionada amb la probabilitat d'aparició d'aquest missatge (la sorpresa no serà tan gran si l'empleat està acostumat a ser acomiadat de diferents feines).

El criteri de la freqüència com a estimació de probabilitat és el mateix que apliquem quan ens trobem en la situació de treure boles d'una urna. Si suposem que cada bola té la mateixa probabilitat de ser escollida, si en treure les boles d'una en una observem que les boles de vegades són negres i altres vegades són blanques, i després de treure cent boles ens adonem que hem obtingut norantacinc boles negres, aquesta circumstància, encara que sigui de manera intuïtiva, ens farà sospitar que la propera bola, la 101, tindrà un 95 % de probabilitats de ser negra.

Podem aplicar aquesta intuïció a l'estudi del llenguatge i adjudicar així un valor d'informació als signes, d'acord amb la seva probabilitat d'aparició en un missatge. En la fórmula [1], la probabilitat d'aparició del signe *i* és expressada com a

$p(i)$, $f(i)$ seria la freqüència d'una determinada paraula en un determinat corpus i N, la quantitat total de paraules d'aquest corpus.

$$p(i) = f(i) / N \quad [1]$$

En el lèxic tenim paraules que són més o menys informatives. L'aparició de paraules com *el*, *de* o *que* en un text ens sorprèn poc, i per això diem que són poc informatives. Si ordenem totes les paraules d'un corpus per freqüència decreixent, observarem que la freqüència d'una unitat està en funció de la seva posició en el rang (r); per tant, es compleix —aproximadament— la fórmula [2]:

$$f(x) = 1/r \quad [2]$$

Si multipliquem la freqüència d'una unitat pel seu rang (equació [3]) obtenim un valor constant c.

$$c = f \cdot r \quad [3]$$

La corba de la funció [2] representa també la distribució de la renda en les societats capitalistes —la llei de Pareto— per a Vilfredo Pareto, que la va descriure el 1906. Ordenats de major a menor renda, s'adverteix com són uns pocs els individus que posseixen la major part de la riquesa, mentre que la gran majoria en percep una mínima part. Entre els lingüistes, el descobriment s'atribueix a J. Estoup, l'any 1916, tot i que va ser divulgada per G. Zipf l'any 1949. L'interès per la llei de Zipf va decaure, però, a partir de l'estudi de Mandelbrot (1961), que la va reformular (fórmula [4]) per tal que s'adaptés millor a les dades observades, particularment en els rangs més alts i més baixos de la corba.

$$f(x) = P \cdot (r + p)^{-B} \quad [4]$$

En la fórmula de Mandelbrot, f és la freqüència i r el rang, mentre que P, p i B són paràmetres constants. Herdan (1964), però, objecta que aquests paràmetres no són constants sinó que depenen de la mida del corpus. La conseqüència d'això és que la fórmula no podria ser aplicada per a la comparació de mostres de mida diferent amb la finalitat, per exemple, de comparar la riquesa lèxica de les mostres.

La riquesa del vocabulari està directament relacionada amb la quantitat d'informació dels signes, la qual cosa determina el grau de dificultat de lectura o densitat d'un text. Això és el que Mandelbrot anomena la *temperatura del discurs*. En el seu cas, plantejava la relació entre l'extensió i el vocabulari d'un text, és a dir,

la quantitat de paraules diferents dividida per la quantitat total de paraules. Però podem establir diferents mesures de riquesa del vocabulari per a un autor o un text no solament segons això, sinó també posant en relació un text analitzat amb un coneixement previ que puguem tenir de la llengua en què està escrit. Aquest coneixement previ pot tenir la forma d'un model de llengua elaborat sobre la base d'un corpus d'una extensió de *n* milions de paraules, un corpus que podríem anomenar *corpus de referència* d'una llengua, conformat per textos de premsa o d'altres gèneres, que pertanyen a una determinada llengua o varietat dialectal. Mal anomenat «corpus de referència», perquè aquest corpus, per més gran que sigui, sempre tindrà un determinat biaix i no arribarà a ser veritablement una referència de la llengua. Aquest model, però, ens permet saber la raresa de les paraules que utilitza un text (o un autor), ja que per a nosaltres representaria un estàndard de llengua «normal».

3.1. *Associació*

Malgrat l'interès que pugui tenir l'assignació individual d'informació per als signes, és molt més interessant estimar les seves probabilitats de combinatòria. Si els signes es combinessin en el llenguatge de manera aleatòria, les seves probabilitats de combinació serien iguals a la multiplicació de les seves probabilitats individuals. La probabilitat de combinació aleatòria de les paraules *i* i *j* (fórmula [5]) defineix que la probabilitat d'aparició conjunta de *i* i *j* (expressada aquí com a intersecció) és igual a la de *i* multiplicada per la de *j*.

$$p(i \cap j) = p(i) \cdot p(j) \qquad [5]$$

Hi ha una aclaparadora quantitat i diversitat de mesures per a calcular les probabilitats de combinació de les paraules —o esdeveniments, en general— (Muller, 1973; Manning i Schütze, 1999; Evert, 2004; entre altres). En lingüística podem veure aquestes mesures aplicades a l'extracció de terminologia especialitzada polilexemàtica o a l'estudi de les col·locacions, tot un capítol en l'estudi del llenguatge. Les combinacions de paraules no són donades solament per la gramàtica, i això té indubtablement el seu correlat en les freqüències de coocurrència. En anglès, es diu *strong coffee*, però no *powerful coffee*. No obstant això, diem una *powerful computer*, però no una *strong computer*.[2] En cada llengua, i fins i tot en cada

2. Aquest últim exemple és interessant, perquè actualment ambdues seqüències de paraules tenen pràcticament la mateixa freqüència a Google; cosa que pot enganyar l'usuari desprevingut, perquè la segona forma, *strong computer*, apareix sempre formant part d'estructures més grans com *strong computer password*. És a dir, el nucli del que depèn *strong* no és en aquest cas *computer* sinó *password*, o *skills*, o *science background*, etcètera.

domini d'especialitat, existeixen certes preferències en les combinacions de paraules de diverses categories (verb-nom; adjectiu-nom; nom-nom, etc.). Per una raó pragmàtica, les coses acostumen a dir-se d'una determinada manera, i si bé la gramàtica ens permetria formular el text d'una altra, fent-ho així correríem el risc de confondre el receptor si ja existeix, en aquesta llengua, domini o registre, una manera típica o idiosincràtica de dir el que volem dir.

Les estadístiques d'associació ens poden informar sobre la manera típica en la qual es combinen les paraules d'una llengua perquè responen a la pregunta sobre quina és la probabilitat que dos esdeveniments ocorrin junts en una mateixa situació o, més precisament, si la freqüència d'aparició de dos esdeveniments en una mateixa situació es pot adjudicar a l'atzar. Un esdeveniment pot ser l'aparició d'una paraula i la situació pot ser un text, un paràgraf, una oració, una «finestra» de *n* paraules, etc. També es pot tractar de l'aparició de les paraules de forma concantenada, o no. Si es tracta d'una seqüència de dues paraules podem parlar d'un *bigrama*, d'un *trigrama* en el cas de tres unitats o d'un *n-grama* per a *n* unitats. Però cal tenir en compte que un *n*-grama podria ser definit d'una altra manera, com una seqüència de lletres o de categories morfològiques. A més a més, la coocurrència pot ser definida d'una manera diferent de la seqüencial. Podem definir *coocurrència* com l'aparició de les dues paraules en una finestra de context sense importarnos l'ordre en què apareixen. Les figures 1 i 2 mostren, per exemple, un criteri de coocurrència que consisteix a comprovar quantes vegades apareixen les paraules —a diferents distàncies i en diferent ordre— en una finestra de context de vint paraules.[3] En ambdós casos, estem analitzant les paraules que coocorren amb la forma

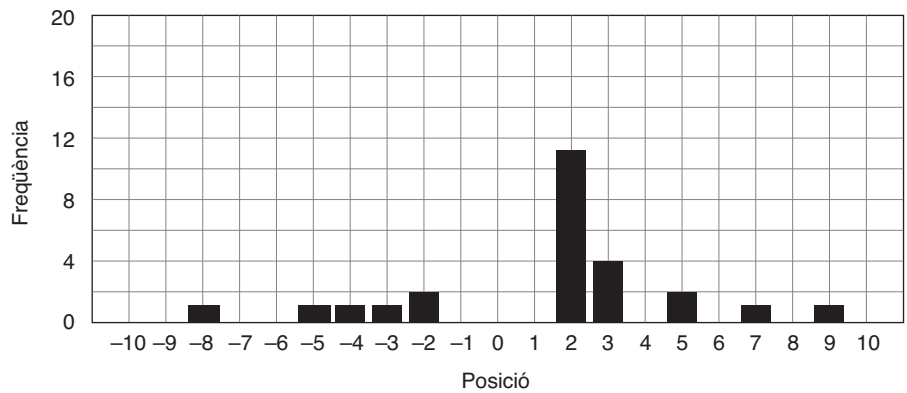


FIGURA 1. Histograma que caracteritza la coocurrència de la forma *platypus* ('ornitorinc', en anglès) i la forma *anatinus* (part de la seva denominació científica). Exemple 1: *…the **platypus** ornithorhynchus **anatinus** is a semiaquatic mammal endemic to eastern Australia, including…*

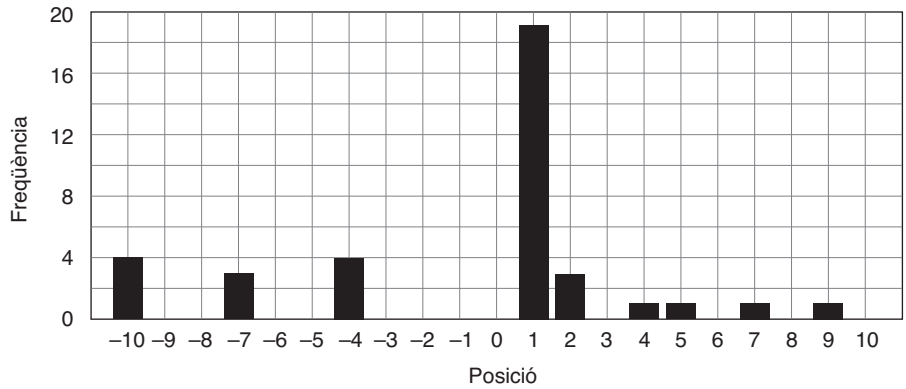


FIGURA 2. Histograma per les formes *platypus* i *has* ('té'). Exemple 2: *…the* ***platypus has*** *four legs which extend horizontally from its body…*

anglesa *platypus* (ornitorinc) en un corpus descarregat d'Internet. En la figura 1, observem que les ocurrències de la forma *anatinus,* una de les paraules amb les quals està associada, es reparteixen a esquerra i dreta de *platypus*. Comprovem aquí que les ocurrències de *anatinus* es concentren en la posició +2, és a dir que la majoria de les vegades la forma *anatinus* apareix dues posicions després de la forma *platypus*, com en l'exemple 1. En la figura 2, observem que el mateix passa amb la forma *has*, tot i que ara la forma es concentra en la posició +1, tal com ocorre en l'exemple [2].

En lingüística de corpus és habitual utilitzar mesures d'associació, però no tant per a falsar una hipòtesi nul·la, segons la qual els elements que estem estudiant es combinen per atzar, sinó més aviat per a ordenar combinacions d'elements a partir de la ponderació que obtenen a conseqüència de l'aplicació d'aquestes mesures. Podem establir diferents tipus de mesures d'associació en funció de la simetria o asimetria que presenten. Entre les mesures d'associació simètriques trobem el concepte d'informació mútua (fórmula [6]), derivat de la teoria de la informació. Representa la quantitat d'informació que ens dóna l'ocurrència de l'esdeveniment *i* sobre l'ocurrència de l'esdeveniment *j* (Church i Hanks, 1991; Manning i Schütze, 1999). Amb aquesta fórmula mesurem, en bits, com és de previsible un esdeveniment *i* en passar *j*, és a dir, quanta sorpresa ens causa *i* quan apareix *j*. En un cas extrem, una alta informació mútua seria que *i* només passa quan ha passat *j*, i en l'extrem oposat, que si passa *i* pot passar *j* o qualsevol altre esdeveniment. És simètrica per definició, és a dir, dóna un mateix valor a *i* donada *j*, que a *j* donada *i*. Aquesta mesura no és aplicable a esdeveniments que tenen poca fre-

3. Aquests histogrames es poden generar automàticament amb el programa Jaguar, accessible a través d'Internet (*http://jaguar.iula.upf.edu*).

qüència, ja que atorgaria una alta associació als que apareixen en conjunció per simple atzar.

$$\mathrm{MI}(i,j) = \log_2 \frac{P(i,j)}{P(i)P(j)} \qquad [6]$$

Entre les mesures d'associació asimètriques trobem la probabilitat condicional dels esdeveniments *i* i *j* (fórmula [7]). És una mesura asimètrica, perquè pot no ser igual la probabilitat *i* donada *j*, que la probabilitat de *j* donada *i*. Per exemple, si *j* és la paraula *auguri* i *i* és *mal* (o *bon*), la paraula *auguri* prediu *mal*, però *mal* no prediu en absolut *auguri*.

$$p(i \mid j) = p(i \cap j) \,/\, p(j) \qquad [7]$$

Fins ara hem vist exemples amb bigrames, és a dir, seqüències de dues paraules. Si estem estimant la probabilitat d'aparició d'un bigrama, podríem també tornar a la fórmula [1] i definir-ne la probabilitat com la freqüència d'aparició dividida per la quantitat total de bigrames que hem observat en un corpus.

Veurem en la secció 4 que és possible, estudiant només les freqüències d'aparició dels bigrames, reconèixer l'escriptura d'autors individuals. Això és possible, perquè el llenguatge és un sistema d'opcions i eleccions. El llenguatge ofereix al parlant o a l'autor diferents possibilitats de combinatòria, i aquest últim, amb les seves eleccions, es va construint a si mateix. Llavors hi comença a haver combinacions que són recurrents o típiques d'un autor en comparació amb altres. Però no parlem només d'autors, perquè també les variants dialectals dels diferents col·lectius o nacions tenen una determinada manera de combinar les paraules i conformen patrons que l'ordinador pot reconèixer mitjançant l'aplicació d'un senzill càlcul estadístic. Aquests patrons, no cal dir-ho, són completament imperceptibles per a l'ull humà.

3.2. *Distribució*

La secció anterior ofereix una visió del corpus com un espai continu on es pot donar la coocurrència d'esdeveniments-paraules, valent-se de la noció de *finestra de context* per a definir quan dues paraules apareixen juntes. Aquesta secció, en canvi, ofereix una perspectiva diferent del corpus, ja que el concebem dividit segons un criteri determinat. En primer lloc, comentarem alguns exemples de com podem estudiar —o, més aviat, visualitzar— la distribució d'unitats o de combinacions d'unitats en corpus dividits de manera diferent. Finalment, estudiarem la manera d'ordenar les unitats d'un corpus a partir del comportament que té la seva corba de distribució.

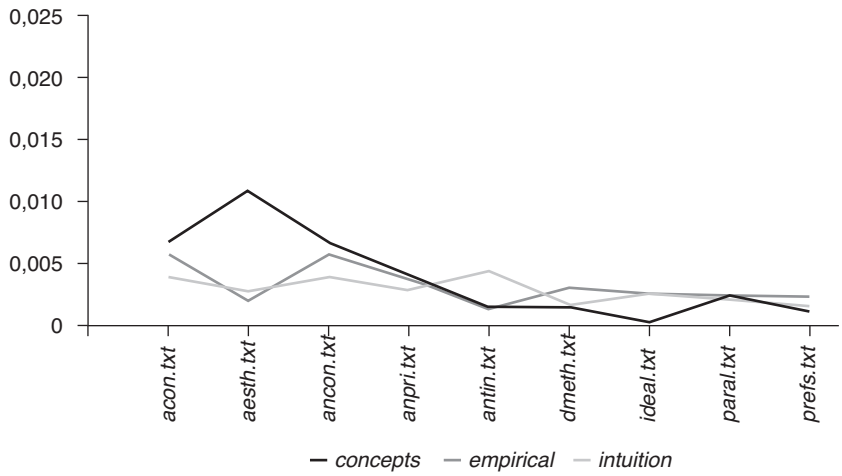


FIGURA 3. Distribució de les formes *concepts*, *empirical* i *intuition* al llarg dels diferents capítols d'una versió en anglès de la *Crítica de la raó pura*, de Kant. L'eix horitzontal representa els diferents capítols. L'eix vertical, la freqüència relativa

El primer exemple és l'anàlisi de la distribució de termes en un document concret. D'acord amb finalitats diverses, ja sigui l'anàlisi del discurs en el pla teòric o l'elaboració de sistemes d'indexació per a la recuperació d'informació, podem tenir interès a esbrinar com es distribueixen les ocurrències de determinats termes en l'obra d'un autor. És possible que existeixin termes clau en certes obres que es distribueixin d'una manera recurrent al llarg del text. També pot ocórrer que alguns termes es concentrin en determinats capítols de l'obra. Pot ser que es trobin en la introducció, per exemple, ja que la seva funció és introduir el lector en els conceptes que després presentarà el text, associats als coneixements que se suposa que té el lector. Però és possible també que aquests termes introductoris no siguin fonamentals en l'obra. La figura 3, per exemple, mostra que tres termes clau en l'obra de Kant, *concepts*, *empirical* i *intuition* es distribueixen de manera regular en l'obra, si bé *intuition* es concentra en el capítol dedicat a l'estètica.

Tanmateix, també és possible que una gran quantitat de paraules es distribueixi de manera regular al llarg de l'obra; però no perquè sigui important per al contingut, sinó perquè forma part del sistema de la llengua. Per això, per als estudis de distribució d'una obra concreta, cal tenir en compte la distribució de les unitats en un corpus. La figura 4 mostra un exemple de distribució d'unitats, aquesta vegada en un corpus diacrònic. Es tracta de les freqüències de les paraules[4] dels ar-

4. Vegeu *http://www.elpais.es*.

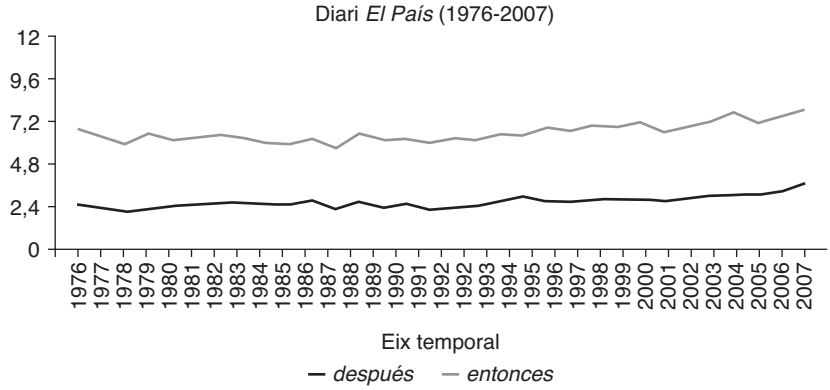


FIGURA 4. Distribució de les formes *después* i *entonces*, dues paraules del vocabulari central de la llengua castellana, en els arxius del diari *El País* en el període 1976-2007. L'eix horitzontal representa el temps i l'eix vertical, la freqüència relativa

xius del diari *El País*.[5] Cadascuna de les divisions en l'eix horitzontal representa totes les edicions d'un mateix any. L'eix vertical representa la freqüència relativa d'una paraula determinada o d'una combinació de paraules en cada any. Podem observar que, mentre que algunes paraules tenen un ús continu al llarg del temps, ja que són paraules del vocabulari central de la llengua (figura 4), altres unitats tenen un ús que fluctua, ja que fan referència a conceptes extralingüístics que tenen diferent vigència en funció de l'agenda temàtica dels mitjans de comunicació (figura 5).

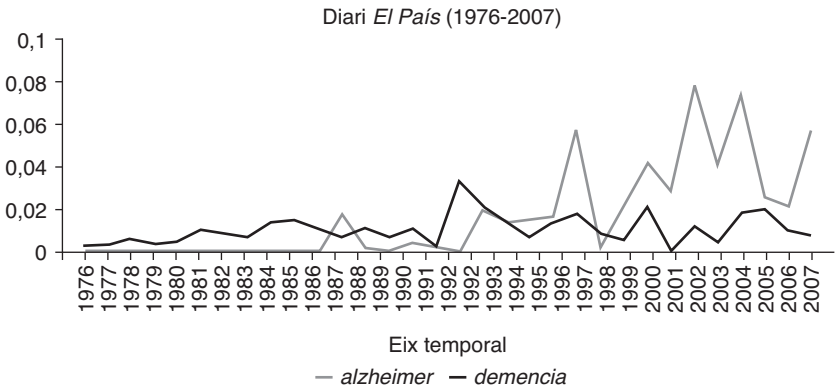


FIGURA 5. Distribució de les formes *demencia* i *Alzheimer* en el mateix corpus, dues unitats que fan referència a coneixement extralingüístic

5. El programa que genera aquestes gràfiques es pot consultar per Internet a l'adreça *http://melot.upf.edu/elpais*.

Un cas diferent és el de dues unitats que, si bé també estan implantades, presenten oscil·lacions a causa de l'evolució del sistema semàntic de la llengua. Ho exemplifiquem en la figura 6, amb les unitats *hombre* i *mujer*, que representen el desenvolupament ideològic d'una societat que pren consciencia del llenguatge sexista. Així, veiem que mentre l'any 1976 la paraula *hombre* és molt més comuna que la paraula *mujer*; aquesta diferència es va revertint amb el temps fins a assolir la mateixa freqüència d'ús l'any 2007.

Basant-nos en el comportament de les corbes de distribució de freqüències de les unitats en aquests corpus dividits, hi ha diversos coeficients que ens interessen per diferents finalitats. En alguns casos, ens interessaran les unitats o combinacions d'unitats que tinguin una freqüència d'ús ascendent, com en el cas de l'extracció de neologia (subsecció 4.3). Però en altres casos ens interessarà saber quin és el vocabulari consolidat d'una llengua, per contrast amb les unitats referencials, és a dir, aquelles que fan referència a coneixement extralingüístic. En aquest cas, ens interessen aquelles unitats que tinguin les corbes més horitzontals. En el cas oposat, podem caracteritzar la irregularitat d'una distribució mitjançant la fórmula [8] (Nazar, 2008) que mesura la dispersió D d'una unitat t per mitjà de la multiplicació del valor màxim de freqüència de t o $\max f(t)$, que seria la freqüència de t en la partició on és més freqüent, multiplicada per $\mathrm{Cr}(t)$, que seria la quantitat de particions en què t té freqüència 0 o una freqüència inferior a un paràmetre k.

$$D(t) = \max f(t) \cdot \mathrm{Cr}(t) \qquad [8]$$

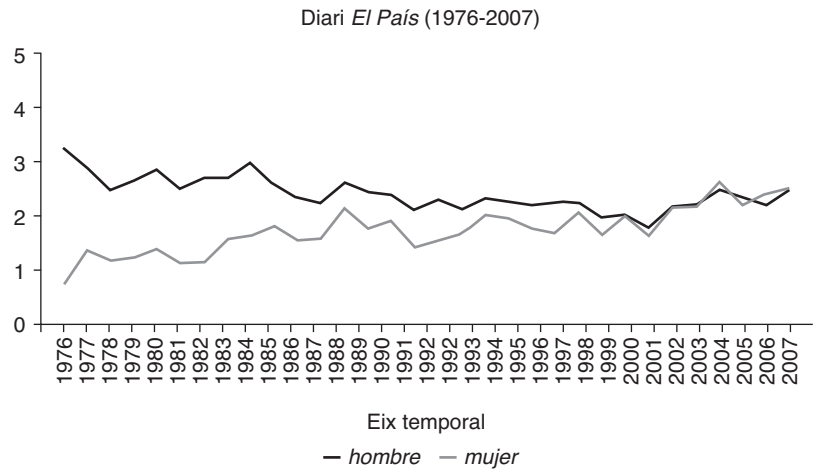


FIGURA 6. Distribució de les formes *hombre* i *mujer* en el mateix corpus

3.3. *Similitud*

En aquesta secció tractem el concepte de *similitud* des d'un espectre ampli. Podríem parlar exclusivament de *similitud entre entitats lingüístiques*, però cal saber que també és possible calcular la similitud entre diferents objectes complexos si som capaços de codificar-los com a vectors. Podem agrupar diferents objectes segons la similitud que tinguin, definida d'acord amb els atributs que comparteixin. Aquests atributs estaran definits per a cada objecte en forma de vector. Un vector pot representar diverses coses: un document, el feix de coocurrències d'un terme, els predicats amb els quals acostuma a aparèixer un nom, etc. La quantitat de valors d'un vector és el que determina la seva dimensionalitat, n, on els x_i en són els components (fórmula [9]).

$$\vec{x} = (x_1, x_2, x_3, \ldots, x_n) \qquad [9]$$

Un vector s'intueix amb facilitat com una fila d'una matriu. La taula 1 mostra, per exemple, una matriu de document per terme, mentre que la taula 2 mostra una matriu de terme per terme.

TAULA 1. *Matriu de document per terme*

	$Term_1$	$Term_2$	$Term_3$	...
Doc_1	1	0	1	...
Doc_2	0	1	1	...
Doc_3	0	1	0	...
...	...	...	...	...

TAULA 2. *Matriu de terme per terme*

	$Term_2$	$Term_3$	$Term_4$	...
$Term_1$	1	0	1	...
$Term_2$	–	0	1	...
$Term_3$	–	–	0	...
...	...	...	...	...

Si els objectes que estem comparant fossin termes i els components dels seus vectors representessin els n-grames de lletres que els conformen, llavors podríem

utilitzar les mesures de similitud entre cadenes de caràcters pel fet de tenir, entre altres coses, una forma de pseudolematització en el treball amb textos no etiquetats, ja que aquesta metodologia seria capaç de detectar la similitud que existeix entre cadenes com *malaltia* i *malalties*; o bé la identificació de variants terminològiques, com en el cas de *superfície pulmonar* i *superfície dels pulmons*.

Amb mesures de similitud com aquestes podem elaborar, per exemple, un programa que, a partir d'un terme d'entrada, indiqui una llista de termes en un corpus que presenten una similitud morfològica. El mateix es pot fer amb documents: a partir d'un document determinat, el programa ordenarà la resta dels documents del corpus d'acord amb la similitud. Però les possibilitats no es limiten a això. En la seva tesi, per exemple, Vanesa Vidal (en preparació) té un experiment en el qual compara diferents verbs especialitzats en funció dels noms amb els quals aquests verbs solen aparèixer.

TAULA 3. *Matriu de verbs per noms*

	Nom_1	$Norm_2$	Nom_3	Nom_4	...
$Verb_1$	0	0	0	1	...
$Verb_2$	1	0	0	0	...
$Verb_3$	0	0	1	0	...
...	...	...	...	...	...

La taula 3 mostra un fragment d'una matriu que té centenars de files i columnes que encreuen la informació de coocurrència de verbs (files) i noms (columnes) en un corpus de genoma. És una matriu binària, ja que codifica, en cada cel·la, l'aparició o la no-aparició de les combinacions verbonominals. La comparació automàtica de tots els verbs[6] entre si dóna una llista dels grups de verbs més similars, és a dir, aquells que es relacionen amb el mateix o gairebé amb el mateix grup de noms. D'aquesta manera, podrem veure que, sense tenir en compte cap tipus d'informació sobre la similitud morfològica i ortogràfica, trobem que, en castellà, en l'àmbit de genoma, els verbs *enrollar* i *desenrollar* són molt semblants perquè apareixen al costat dels noms *hélice*, *cadena*, *adn*, *hebra*, etc.; així com els verbs *beber, ingerir* i *reabsorber* s'assemblen perquè comparteixen els noms *agua*, *cantidad*, *cola, célula* i *glucosa*, entre d'altres. Diferents autors han adoptat estratègies més o

6. El programa que fa aquesta comparació (algorisme de *clustering*) es pot executar a través d'Internet a l'adreça *http://melot.upf.edu/clusteau*, però encara no està suficientment documentat.

menys semblants; no ja en l'estudi de combinacions verbonominals, sinó per al descobriment de sinònims, quasisinònims o bé equivalents en diferents llengües que posen en relació elements que comparteixen els mateixos veïns (Nazar, en preparació).

Entre altres mesures de similitud, la mesura Dice és apropiada per a la comparació de vectors amb valors binaris. El que fa és comptar la quantitat de dimensions en què en dos vectors el valor és superior a zero. Si X i Y són els dos vectors, la mesura queda expressada en la fórmula [10]. $|X|$ és el conjunt cardinal de X, és a dir, la quantitat de components. Es multiplica per dos per tenir un escala que va de 0,0 a 1,0, que seria la similitud total.

$$\text{Dice}(X, Y) = \frac{2\left|X \cap Y\right|}{\left|X\right| + \left|Y\right|} \qquad [10]$$

La mesura Jaccard (fórmula [11]) és similar a l'anterior, però introdueix una normalització: la divisió per la quantitat de dimensions dels vectors, és a dir, que introdueix una penalització quan hi ha poques dimensions compartides en proporció a la quantitat total de dimensions.

$$\text{Jaccard}(X, Y) = \frac{\left|X \cap Y\right|}{\left|X \cup Y\right|} \qquad [10]$$

4. Aplicacions pràctiques

Si bé la secció anterior ja suggereix alguns exemples d'aplicació pràctica, en aquesta secció presentem un espectre d'aplicació més ampli. Analitzarem l'aplicació de mesures de similitud i coocurrència en l'àmbit de la classificació automàtica de documents en les dues modalitats en què aquesta pràctica existeix actualment: la classificació amb aprenentatge supervisat i no supervisat. Finalment, comentarem breument l'aplicació de mesures de distribució aplicades al descobriment de neologia. La manca d'espai ens obligarà a deixar temes que hauria estat molt interessant comentar, com per exemple l'aplicació de metodologies estadístiques a l'extracció de terminologia especialitzada, així com les metodologies per a l'extracció de terminologia bilingüe de corpus no paral·lels, que són línies de recerca en curs.

4.1. *Classificació de documents*

Com és sabut, els algorismes de classificació automàtica de documents es divideixen en *supervisats* i *no supervisats* (Manning i Schütze, 1999; Sebastiani, 2002). En ambdós casos estem agrupant objectes (documents, en aquest context), però la diferència és que, en el primer, un algorisme de classificació té un coneixement previ sobre els objectes que ha de classificar, ja que ha passat per un procés d'«entrenament», en el qual un usuari li ha ensenyat exemples d'objectes classificats segons un criteri qualsevol. En el segon cas, en canvi, la tasca de classificació s'ha de fer sense aquest coneixement, és a dir que l'algorisme no sabrà quantes ni quines són les categories segons les quals els objectes han de ser agrupats, i per tant la classificació serà una propietat que sorgirà a partir de les similituds que tenen els objectes.

4.1.1. Classificació amb aprenentatge supervisat

L'any 2004 em vaig vincular a dos grups d'investigació que estaven treballant en àrees que en principi poden semblar dissímils. Un dels grups estava treballant en l'atribució d'autoria amb el propòsit d'aplicar-la a la lingüística forense. L'altre grup, més vinculat a la terminologia, tenia interès a trobar una manera sistemàtica de classificar un document, tant segons la temàtica com segons el grau d'especialitat. La filosofia de treball en ambdós grups era la mateixa: dissenyar estratègies fonamentades en el coneixement lingüístic, entesa com l'examen manual de la casuística i la identificació, d'acord amb la intuïció de l'investigador, d'aquells trets que podrien ser discriminants de les diferents categories. En ambdós casos es tracta d'un treball d'enorme complexitat i arrelat en el coneixement que l'investigador té de la llengua particular en què està escrit el text. En el cas de la lingüística forense, aquests trets poden ser, per esmentar alguns exemples, girs idiosincràtics que puguin delatar una pertinença a una zona geogràfica o a una condició social, o bé particularitats com els errors d'ortografia o gramàtica que tinguin en comú els textos d'autoria disputada amb aquells textos d'autoria indubtable (vegeu Turell, 2005, per a una introducció). En el cas de la classificació de documents per tema o per grau d'especialitat, l'estratègia consistia a trobar trets lingüístics d'un domini temàtic (la densitat de terminologia especialitzada en el text, per exemple) o bé altres trets morfològics i lèxics que poden ser característics de la literatura especialitzada (Cabré *et al.*, 2009).

En aquest context, va sorgir el programari Poppins.[7] Aquest programa representa una solució de classificació diferent, ja que es pot aplicar tant als problemes d'atribució d'autoria com a la classificació per tema, per grau d'especialitat i fins i

7. El programa Poppins pot ser executat a través d'Internet a l'adreça *http://www.poppinsweb.com*.

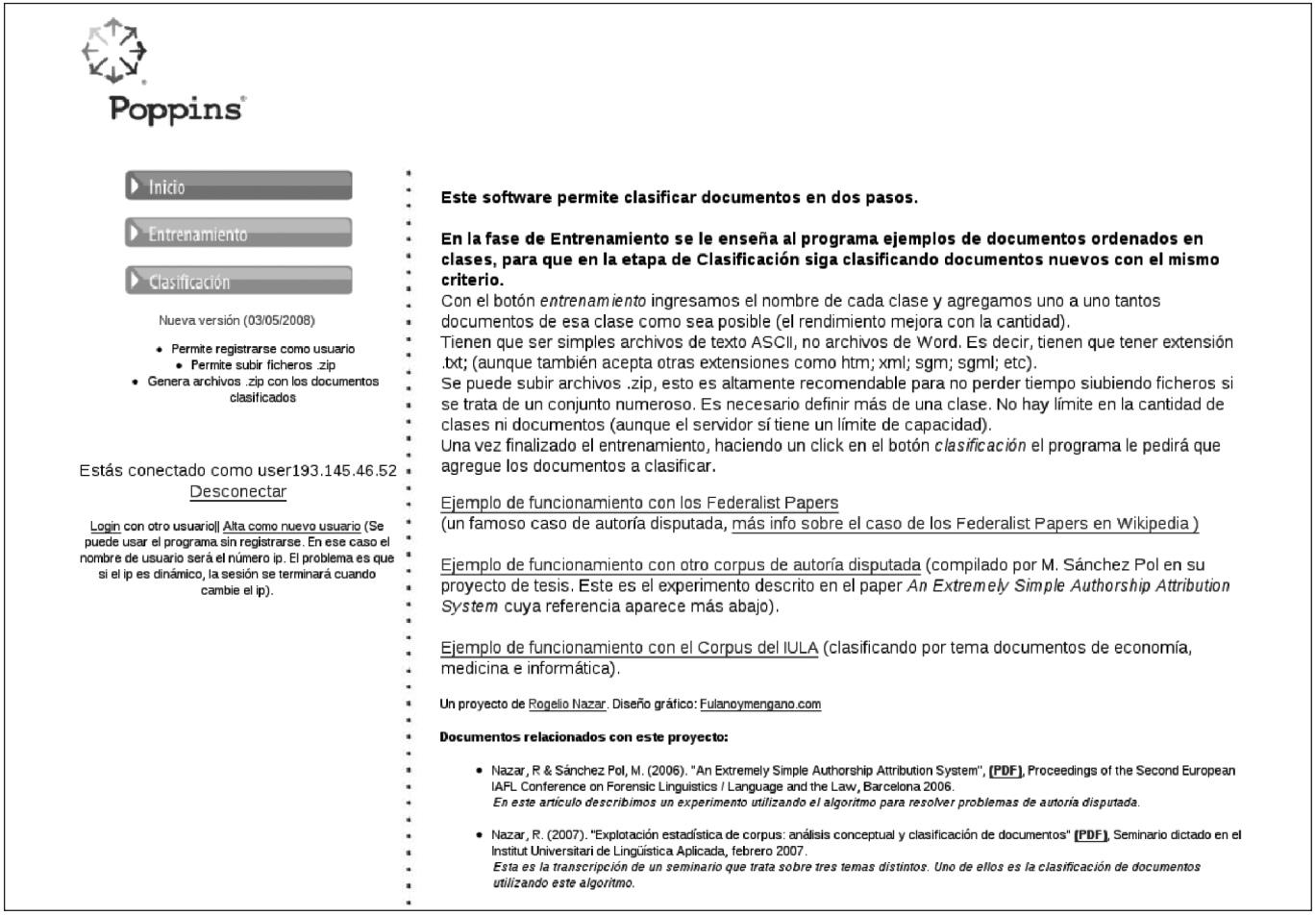


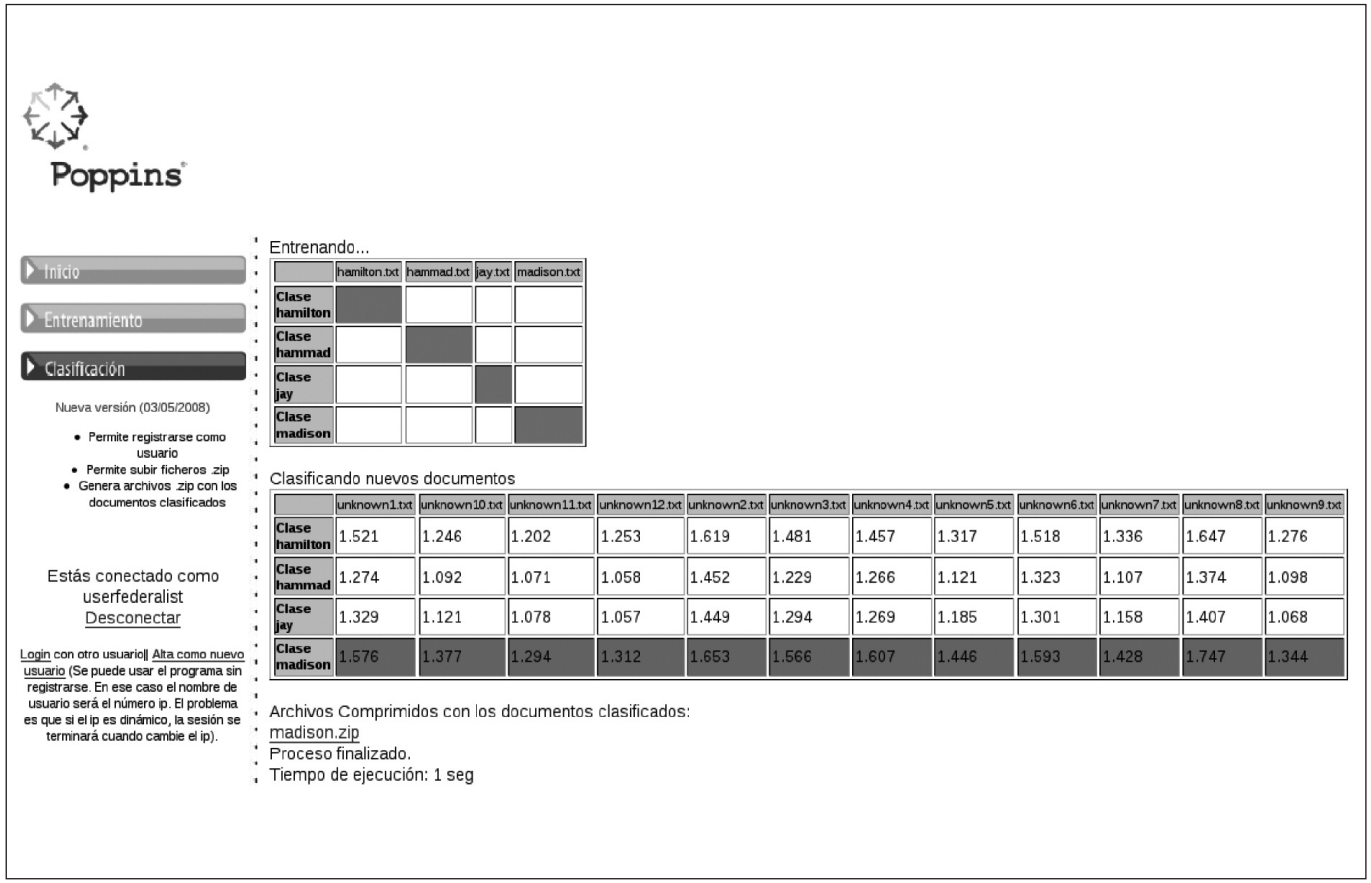
FIGURA 7. Interfície web del programa de classificació automàtica Poppins (*http://www.poppinsweb.com*)

tot per altres problemes de classificació en els quals l'algorisme sigui entrenat, i això amb independència de la llengua dels documents, del domini temàtic o del criteri de classificació. Com dèiem abans per al cas dels algorismes supervisats, la lògica d'aquest programa inclou dues fases principals. A la primera, la fase d'entrenament, un usuari «presenta» al programa exemples de documents ordenats en classes. Un cop acabada aquesta etapa, l'etapa de classificació consisteix a, partint d'un nou conjunt de documents, ordenar-los basant-se en la classificació que ha après durant la fase d'entrenament. El mode de funcionament és bàsic perquè els textos que són classificats no són sotmesos a cap tipus de processament. L'única operació que es fa és calcular les freqüències d'aparició dels diferents bigrames de paraules del corpus. Així, cada classe d'entrenament es converteix en un vector que té per atributs els bigrames, i, per valor, la freqüència d'aparició. D'aquesta manera, a partir d'un nou document, el que fem és computar una mesura de similitud que consisteix a sumar les freqüències dels bigrames que tenen en comú el document per classificar i cadascuna de les classes. La comparació que obté com a resultat la suma més gran és la classe escollida per a aquest document.

Amb Marta Sánchez Pol (Nazar i Sánchez Pol, 2006) vam descobrir que, amb aquest programa, podíem determinar correctament l'autoria d'un text amb una probabilitat del 90 %. La interfície del programa mostra experiments amb altres casos, com el dels Federalist Papers, un famós cas d'autoria disputada, i atribueix els textos d'autoria disputada a James Madison (figura 8) tal com han demostrat altres estudis duts a terme (Mosteller i Wallace, 1984). Pel que fa a la classificació per temàtica i per grau d'especialitat, experiments de classificació de documents del Corpus Tècnic de l'IULA van demostrar nivells de precisió semblants. L'experiment encara es pot repetir de diverses maneres, mitjançant la classificació dels documents per llengua, per variant dialectal o per altres criteris.

4.1.2. Classificació amb aprenentatge no supervisat

Com hem dit en la introducció d'aquesta secció, la classificació amb aprenentatge no supervisat és l'escenari en el qual l'algorisme no ha passat per una etapa d'entrenament i, per tant, no sap quines ni quantes són les categories en què han de ser classificats els documents. Si en el cas anterior relacionàvem la classificació de documents amb aplicacions concretes com l'atribució d'autoria, en aquest cas la classificació de documents amb aprenentatge no supervisat es relaciona amb la desambiguació de terminologia. Això és així perquè plantegem la classificació com un problema de desambiguació. En aquest experiment, reunim una col·lecció de documents en què apareix una forma ambigua, per exemple per mitjà de la descàrrega de documents d'Internet, i els classifiquem a partir dels diferents sentits que pot mostrar aquesta forma dins la col·lecció. Aquesta classifica-



	hamilton.txt	hammad.txt	jay.txt	madison.txt
Clase hamilton				
Clase hammad				
Clase jay				
Clase madison				

	unknown1.txt	unknown10.txt	unknown11.txt	unknown12.txt	unknown2.txt	unknown3.txt	unknown4.txt	unknown5.txt	unknown6.txt	unknown7.txt	unknown8.txt	unknown9.txt
Clase hamilton	1.521	1.246	1.202	1.253	1.619	1.481	1.457	1.317	1.518	1.336	1.647	1.276
Clase hammad	1.274	1.092	1.071	1.058	1.452	1.229	1.266	1.121	1.323	1.107	1.374	1.098
Clase jay	1.329	1.121	1.078	1.057	1.449	1.294	1.269	1.185	1.301	1.158	1.407	1.068
Clase madison	1.576	1.377	1.294	1.312	1.653	1.566	1.607	1.446	1.593	1.428	1.747	1.344

FIGURA 8. Resultat de la classificació de documents per autor mitjançant el programa Poppins en el cas de l'autoria disputada dels Federalist Papers

ció es duu a terme per mitjà dels grafs de coocurrència lèxica. Prenguem, en primer lloc, un exemple amb una forma ambigua com *ratón*, en castellà, que, en el Corpus Tècnic de l'IULA (Vivaldi, 2009) —que conté documents sobre informàtica i sobre genoma— pot ser utilitzada per a fer referència a un dispositiu perifèric de l'ordinador o bé a un animal de laboratori.

En els grafs de coocurrència hi ha un node principal, situat a la zona superior central, que és el terme que estem analitzant: *ratón*, en aquest cas. D'aquest node, en depenen tots els altres. Cada node representa una paraula o una combinació de paraules, i les connexions entre nodes expressen que les paraules que els nodes representen apareixen juntes en els mateixos contextos on apareix la unitat que estem analitzant. En la figura 9 s'aprecia l'existència de dues regions en el graf, una a la dreta i una altra a l'esquerra. Aquestes dues regions —atractors o clústers de nodes— es corresponen amb cadascun dels sentits que la forma presenta en el corpus. En un cas, les unitats amb les que apareixerà *ratón* seran *cromosoma, mamífero, rata, genoma, laboratorio, bacteria,* entre altres; mentre que, en l'altre cas, les unitats que es relacionen amb *ratón* són *usuario, pantalla, teclado, clic,* etcètera.

En la tesi (en preparació) presento, entre altres coses, un estudi de desambiguació de sigles, ja que aquestes són formes ambigües per naturalesa. Així, davant d'una col·lecció de documents descarregada d'Internet amb una forma ambigua com *NLP*, per exemple, un programa informàtic[8] és capaç d'obtenir dos clústers que representen els dos sentits d'aquesta paraula: d'una banda, documents referits a la forma expandida *natural language processing* i de l'altra, documents sobre *neuro-linguistic programming*. En el primer cas, *NLP* es relaciona amb unitats com *knowledge representation*, *language technology*, *functional grammar*, *machine translation*, *statistical NLP*, *computational linguistics*, entre altres; mentre que el segon clúster inclou unitats com *practitioner training*, *practitioner NLP*, *gestalt therapy*, *John Grinder*, *Richard Bandler*, *Robert Dilts*, etcètera.

4.2. *Descobriment de neologia*

En aquesta secció analitzarem l'aplicació d'algunes de les mesures de distribució que hem vist en la secció 3.2, amb el propòsit concret de fer un experiment d'extracció automàtica de neologia. Els resultats de l'aplicació d'aquestes tècniques per a l'extracció de neologia, així com de les tècniques de desambiguació automàtica presentada en el punt anterior (4.1.2) van ser presentades en un treball previ (Nazar i Vidal, 2008).

8. El programa que fa aquesta classificació de documents descarregats d'Internet mitjançant una forma ambigua també es pot executar a l'adreça *http://melot.upf.edu/mandinga*, encara que no existeix documentació per al programa i la interfície és encara rudimentària. El resultat de l'experiment de *NLP* es pot veure a la adreça següent: *http://melot.upf.edu/nlp*.

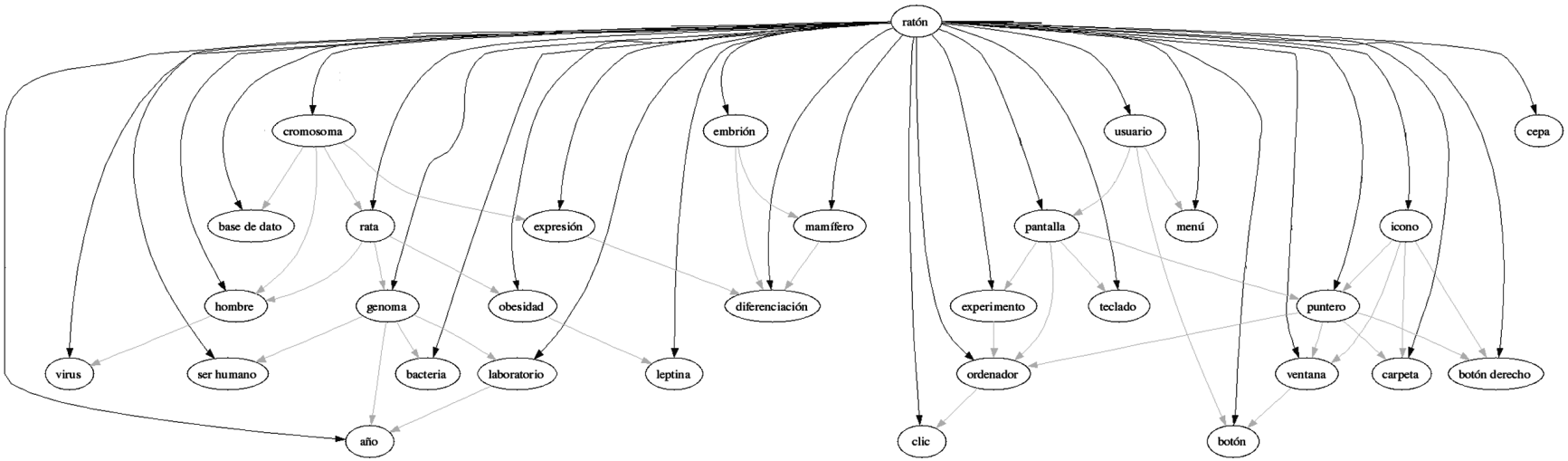


FIGURA 9. Graf de coocurrència de la forma ambigua *ratón*, utilitzada per a fer referència al dispositiu perifèric de l'ordinador en els documents d'informàtica i a l'animal utilitzat en laboratori en els documents de genoma

Les figures 10 i 11 ofereixen gràfiques que ja ens són familiars, perquè hem vist corbes semblants en la subsecció 3.2: seguiments de determinades unitats lèxiques al llarg del corpus diacrònic d'*El País*. Mostren exemples del comportament d'unitats que considerem neologismes, com ara *teléfono móvil*, *teléfono fijo* i *cambio climático*, unitats la freqüència d'ús de les quals mostra un increment acusat en la línia del temps.

$$f(x) = x^{10} \qquad [12]$$

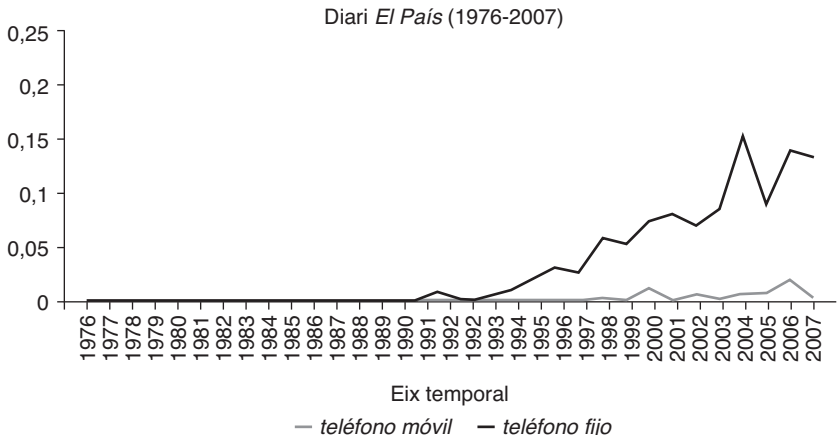


FIGURA 10. Distribució de les formes *teléfono móvil* (corba superior) i *teléfono fijo* (corba inferior) en el corpus diacrònic d'*El País*

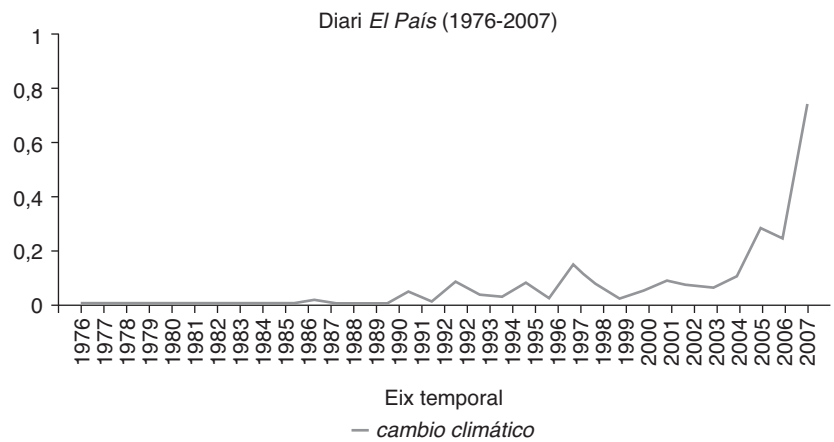


FIGURA 11. Distribució de la forma *cambio climático* en el mateix corpus

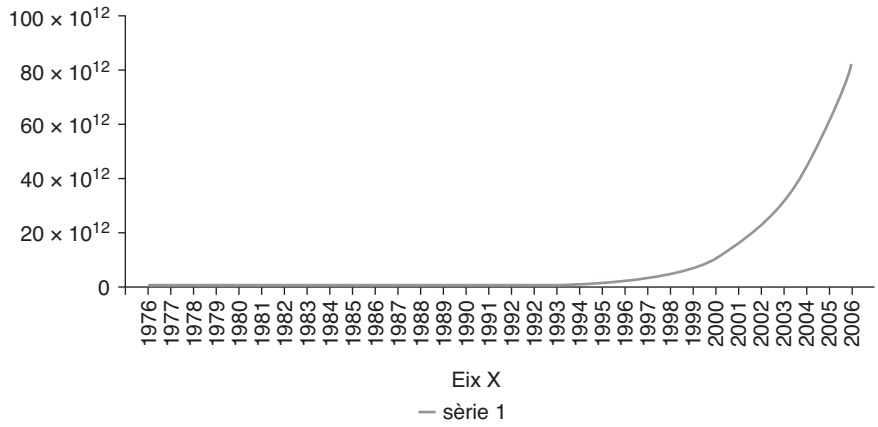


FIGURA 12. Gràfica del neologisme ideal

En l'esmentat treball sobre extracció de neologia vam definir el que seria la corba de comportament d'un neologisme ideal o teòric, representada en la figura 12 i definida en la fórmula [12]. Es tracta d'una corba exponencial en l'interval d'anys estudiat. L'experiment va consistir a prendre una mostra de *n* unitats del corpus (les unitats eren tant paraules aïllades com seqüències de fins a cinc paraules de longitud) i a ordenar-les d'acord amb la distància euclidiana de les seves corbes de freqüència amb la corba d'aquest neologisme ideal. D'aquesta manera, vam obtenir les unitats que s'han anat incorporant a la llengua en els darrers anys, unitats que després s'han de filtrar, ja que inclouen formes que no són neologismes, com és el cas de noms propis o referents que han adquirit notorietat en els darrers anys.

Naturalment, aquest senzill mètode no resultava eficaç en el cas dels neologismes semàntics, unitats que si bé són formalment idèntiques a altres formes de la llengua, es comencen a fer servir amb un significat diferent. Aquestes formes representen un desafiament per a l'extracció automàtica amb els mètodes tradicionals, però aquest mateix escenari és el que trobàvem en la subsecció 4.1.2, en la qual classificàvem contextos d'aparició d'unitats polisèmiques. És el cas, per exemple, de la forma *palabra de honor*, que si bé té un ús literal, en el sentit de 'fer una promesa verbal', en els darrers anys és cada vegada més freqüent utilitzar-la per a designar un determinat tipus d'escot. Si bé la seva condició de neologisme per a aquest segon sentit és discutible, ja que aquest tipus d'escot no és nou, sí que és nova la massificació d'aquest ús del terme, i, per tant, l'exemple segueix sent útil. Un algorisme de clusterització (*clustering*) similar al descrit en la subsecció 3.2 és capaç de classificar tots els contextos d'aparició de la forma *palabra de honor* en els arxius d'*El País* i oferir dos clústers amb un nom per a cadascun. El clúster 1 és ano-

menat *empeñar* i el clúster 2 és anomenat *escotes*. Cada un d'aquests clústers conté una sèrie d'unitats lèxiques que conformen l'entorn típic de les ocurrències de l'expressió en un sentit i en l'altre. Així, en el clúster 1, tenim unitats com: *Astarloa*, *Barrionuevo*, *confederal*, *consentido*, *credulidad*, *empeñar*, *esclarece*, *Escudero*, *Fusté*, *Herrero*, *incité*, *inocencia*, *proclamar*, *quebrantamiento*, *reiterado*, etc. Aquestes formes es relacionen amb el sentit literal. Veiem que es tracta de noms propis de personatges públics, per als quals la credibilitat no hauria de ser irrellevant. En el cas del segon clúster, en canvi, els veïns típics tenen relació amb el món de la moda: *cubren*, *drapeados*, *escotes*, *Gucci*, *marrón*, *modista*, *ojito*, *organza*, *Swarovski, tonos*, etcètera.

5. Conclusions

Aquest article presenta una visió àmplia de la cruïlla entre la lingüística i l'estadística, i inclou alguns exemples de tècniques que es poden utilitzar per a l'estudi del llenguatge. Aquestes tècniques s'han acompanyat, a més, amb exemples d'aplicació concreta, com és el cas de la classificació de documents amb supervisió o sense, així com la desambiguació de signes polisèmics i el descobriment de neologia. Hauria estat interessant esmentar altres exemples d'aplicació pràctica d'aquestes tècniques, com la utilització de mesures de similitud per a la comparació entre unitats lèxiques de diferents llengües, és a dir, l'extracció de terminologia bilingüe des de corpus no paral·lels, o bé per a la comparació d'unitats lèxiques de diferents varietats dialectals.

A priori, pot semblar que es tracta d'àrees d'aplicació completament diferents, sobretot per a qui està acostumat a enfrontar tasques d'aquest tipus amb la incorporació de regles explícites que codifiquen coneixement de la llengua o del domini temàtic, així com informació semàntica extreta de diccionaris i ontologies, en el cas de l'extracció de terminologia, o corpus d'exclusió lexicogràfics, en el cas de l'extracció de neologia. L'estadística, per contra, possibilita una manera diferent de concebre la llengua. Una investigació de la complexitat, però des d'una perspectiva integradora i simplificadora. Des del punt de vista estadístic, tasques i dades dissímils comencen a semblar relacionades. De vegades, els mateixos mètodes o les mateixes formes de pensar es poden aplicar a problemes que en principi semblaven completament diferents. Concebem, doncs, l'estadística com una «trans-disciplina».

Per a tancar aquest article, és important remarcar que cal no perdre de vista l'aspecte teòric. No estem parlant només de «trucs enginyerils» per resoldre problemes pràctics que no tenen una relació intrínseca amb la lingüística, com si aquestes solucions estiguessin desproveïdes de teoria. Està per veure si l'estadística i la lingüística conformen disciplines diferents o si hi pot haver alguna cosa que anomena-

ríem una «sensibilitat estadística» en l'anàlisi lingüística, una manera d'aproximarnos a les dades, d'advertir patrons, regularitats o tendències en el cúmul dels casos individuals en què l'ull humà no pot veure sinó quantitat i diversitat.

Agraïments

Aquest treball ha estat possible gràcies al finançament per al projecte RICOTERM3 (Ministeri d'Educació i Ciència: HUM2007-65966-C02-01/FILO. Investigadora principal: doctora Mercè Lorente). Voldria agrair també l'ajuda d'Amor Montané i Alba Coll en la redacció en català.

7. Referències

Cabré, M. T.; Bach, C.; Da Cunha, I.; Morales, A.; Vivaldi, J (2009). *Comparación de algunas características lingüísticas del discurso especializado frente al discurso general: el caso del discurso económico*. XXVII Congreso de AESLA (Ciudad Real, 26-28 març 2009).

Castoriadis, C (1975). *La institución imaginaria de la sociedad.* Buenos Aires: Tusquets.

Church, K.; Hanks, P (1990). «Word Association Norms, Mutual Information and Lexicography». *Computational Linguistics*, vol 16, núm. 1, p. 22-29.

Dilthey, W (1986). *Introducción a las Ciencias del Espíritu*. Madrid: Alianza.

Evert, S (2004). *The Statistics of Word Coocurrences*. Tesi (doctorat). Stuttgart: Universitat de Stuttgart. Institut für Maschinelle Sprachverarbeitung, 2004.

Herdan, G (1964). *Quantitative Linguistics*. Washington: Butterworths.

Mandelbrot, B (1961). «On the theory of word frequencies and Markovian models of discourse». A: *Structure of Language and its Mathematical Aspects.* Symposia on Applied Mathematics. American Mathematical Society. Vol. 12, p. 190-219.

Manning, C.; Schütze, H (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

Mosteller, F.; Wallace, D (1984). *Applied Bayesian and Classical Inference: the Case of the Federalist Papers*. Nova York: Springer.

Muller, C. (1973). *Estadística Lingüística*. Madrid: Gredos.

Nazar, R. (2008). Diferencias cuantitativas entre referencia y sentido. Actas del XXVI Congreso de AESLA. (Universitat d'Almeria, 3-5 d'abril de 2008).

— ([en preparació]). *Quantitative Approach to Concept Analysis*. Tesi (doctorat). Barcelona: Universitat Pompeu Fabra. Institut Universitari de Lingüística Aplicada.

Nazar, R; Sánchez Pol, M. (2006). *An Extremely Simple Authorship Attribution System.* Second European IAFL Conference on Forensic Linguistics / Language and the Law (Barcelona, 2006).

Nazar, R.; Vidal, V. (2008). *Aproximación cuantitativa a la neología*. I Congreso Internacional de Neología en las lenguas románicas (Barcelona, 7-10 maig 2008).

SEBASTIANI, F. (2000). *Machine earning in automated text categorization*. ACM Press, vol. 34, núm. 1.

SHANNON, C. E. (1948). «A mathematical theory of communication». *Bell System Technical Journal*, vol. 27 (juliol), p. 379-423.

SNOW, C. P. (1959 [1993]). *The Two Cultures*. Cambridge: Cambridge University Press.

TURELL, M. (2005). «Presentación». A: *Lingüística forense, lengua y derecho: conceptos, métodos y aplicaciones*. Barcelona: Universitat Pompeu Fabra. Institut Universitari de Ligüística Aplicada, p. 13-16.

VICKERS, B. (2002). *Counterfeiting Shakespeare*. Cambridge: Cambridge University Press.

VIDAL, V. (en preparació) *Combinatoria verbo-nominal en el discurso de especialidad. Delimitación, caracterización y soluciones terminográficas.* Tesi (doctorat). Barcelona: Universitat Pompeu Fabra. Institut Universitari de Lingüística Aplicada.

VIVALDI, J. (2009). *Corpus and exploitation tool*: *IULACT and bwanaNet*. I Congreso Internacional de Lingüística de Corpus (Múrcia, 7-9 maig 2009).