

CABRÉ, M. T.; DOMÈNECH, O.; ESTOPÀ, R.; FREIXA, J.; SOLÉ, E. (2004) «La lexicografia i la identificació automatitzada de neologia lèxica». En: Battaner, P.; DeCesaris, J. (eds.) (2004) *De Lexicografia. Actes del I Symposium Internacional de Lexicografia (Barcelona, 16-18 de maig de 2002)*. p. 287- 294. (Sèrie activitats, 15). ISBN: 84-96367-06-1 (CL).

La lexicografia i la identificació automatitzada de neologia lèxica

M. T. Cabré, M. Domènech, R. Estopà, J. Freixa, E. Solé

Observatori de Neologia (IULA)

Introducció

L'objectiu d'aquesta comunicació és mostrar, a partir de l'experiència adquirida amb els anys de treball a l'Observatori de Neologia, com es podria refinar el reconeixement i la detecció de neologismes introduint una sèrie de filtres que matisen el criteri d'exclusió lexicogràfica.

Demostrem que el criteri lexicogràfic és útil en la identificació de candidats a neologismes, però a la vegada insuficient en alguns casos, ja que permet l'obtenció de dades neològiques amb graus de neologicitat molt diferents. De manera que caldrà establir filtres de neologicitat que donin compte de les diferents percepcions dels parlants davant dels neologismes seleccionats en el procés de buidatge a partir del criteri lexicogràfic. Començarem analitzant la relativitat del concepte de *neologisme* avalada per la multiplicitat de paràmetres que entren en joc a l'hora de definir-lo.

El concepte de neologisme

El significat del mot *neologisme* és difícil de delimitar amb paràmetres objectivables, com reflecteixen els diccionaris generals de llengua en definir-lo. Segons el *Diccionari de la llengua catalana*, un neologisme és una "unitat lèxica nova, formalment o semànticament, creada en una llengua per les pròpies regles de formació de mots o manllevada a una altra llengua". El *Diccionario de la lengua española* de la Real Academia Española (2001) ens ofereix una definició molt semblant ("vocablo, acepción o giro nuevo en una lengua"). I igualment els diccionaris d'altres llengües que podem consultar.

Són definicions que remetent sistemàticament al concepte de *nou*, concepte que, d'acord amb el diccionari normatiu de la llengua catalana, significa "originat o ocorregut fa poc, aparegut de poc, que apareix per primera vegada". Ens trobem, doncs, una altra vegada, amb un concepte difícil de delimitar de manera objectiva, amb la dificultat d'establir uns paràmetres clars que

ens permetin saber quines unitats de la llengua es poden considerar neològiques, perquè ens apareixen interrogants com els següents, que no tenen una resposta satisfactòria: Què vol dir "fa poc"? "Fa poc" per a qui? "Fa poc" respecte de què?

Diversos autors han remarcat la relativitat del concepte *neologisme* i, tradicionalment, per definir-lo i delimitar-lo, han recorregut als paràmetres utilitzats per identificar o detectar les paraules noves. Alain Rey (1976), per exemple, considera que una unitat lèxica és neològica segons quin sigui el paràmetre d'identificació del qual es parteix. Aquest autor estableix tres paràmetres de diferent naturalesa que ens permeten establir conjunts de mots nous diferents: temporal, psicolingüístic i lexicogràfic. Així, si el criteri de selecció és temporal, considerarem neologisme qualsevol paraula que ha aparegut en un període recent; si el paràmetre és psicolingüístic, un neologisme serà aquella unitat lèxica que els parlants perceben com a nova; i, finalment, si el criteri és lexicogràfic, una unitat lèxica es pot considerar neològica si no es troba documentada en un determinat corpus lexicogràfic.

El problema que presenta, al nostre entendre, aquesta classificació i d'altres de semblants és que parteixen del supòsit d'identificar el concepte de neologisme amb els criteris de detecció de neologismes. Si en lloc de partir d'aquesta identificació, mirem de tractar les dues qüestions separadament, podem dir:

— d'una banda, que un neologisme és una "unitat lèxica nova", és a dir, una unitat lèxica "originada o ocorreguda fa poc "; el *criteri temporal*, doncs, esmentat per Rey, no és un criteri de detecció de neologismes, sinó que és pròpiament allò que els defineix, perquè és inherent a la seva essència mateixa: la "novetat";

— i d'altra banda, que els criteris que tenim per detectar o identificar aquestes paraules noves són dos: un de *tangible*, que seria molt objectivable i inclouria el *criteri lexicogràfic* de què parla Rey, i un de *cognitiu*, molt més difícil d'objectivar i força coincident amb l'anomenat *criteri psicolingüístic* de Rey.

El criteri tangible correspon a la utilització de paràmetres documentals que permeten considerar neològiques paraules o expressions que no es troben documentades en un corpus, ja sigui lexicogràfic o textual, prèviament establert. Aquest criteri de selecció ha estat, com veurem més endavant, el que ha utilitzat des dels seus orígens l'Observatori de Neologia.

El criteri cognitiu, en canvi, no es basa en el fet que un mot aparegui o no en un corpus de referència determinat, sinó que respon a diversos factors lligats a la percepció de novetat que un parlant té respecte d'una unitat de la llengua. Es tracta, doncs, d'un criteri més subjectiu perquè està condicionat per paràmetres de diferent naturalesa com el coneixement lingüístic i enciclopèdic del parlant, el seu nivell sociocultural o el context comunicatiu.

El concepte de neologisme a l'Observatori de Neologia

L'Observatori de Neologia analitza el fenomen de l'aparició de paraules noves o neologismes en els mitjans de comunicació de gran difusió (premsa, ràdio i televisió) i en textos escrits *espontanis* (és a dir, que no passen per un servei de correcció) en català i en castellà. Aquest grup està format per un col·lectiu molt ampli i diversificat de col·laboradors, tant pel que fa a la seva formació com a la seva activitat professional.

Aquesta diversitat dels membres de l'Observatori, però també els objectius aplicats del projecte i les possibilitats d'automatització del procés de detecció i selecció dels neologismes ha condicionat des dels orígens la tria del criteri de buidatge de neologismes, que s'ha basat en el criteri lexicogràfic per assegurar la sistematicitat: es considera neologisme qualsevol paraula que no apareix en un corpus lexicogràfic d'exclusió prèviament establert, format per diccionaris prescriptius i per diccionaris descriptius. Cal esmentar, però, que aquest criteri lexicogràfic es restringeix o es refina per mitjà de dos filtres, un de positiu i un altre de negatiu:

a. Quant al primer, es consideren neologismes, encara que es documentin al corpus d'exclusió, les paraules amb marques de neologicitat en el corpus lexicogràfic (normalment amb un asterisc). Per exemple: *best seller*, *lehendakari*, *mountain bike* i *western*.

- b. Pel que fa al filtre negatiu, no es consideren neologismes, encara que no es documentin al corpus d'exclusió:
- els augmentatius, diminutius i superlatius, per la seva potencialitat derivativa gairebé il·limitada
 - els adverbis en *-ment*
 - les paraules formades amb el prefix *ex-* quan s'adjunta a radicals simples o derivats que fan referència a càrrecs, oficis o relacions personals
 - els gentilicis, amb l'excepció dels compostos del tipus *hispanofrancès*, *sinotibetà*
 - les unitats lèxiques sintagmàtiques altament especialitzades

- les sigles i abreviatures (excepte quan són la base d'un procés de formació —*ugetista*— o s'han lexicalitzat de tal manera que es consideren noves unitats lèxiques —les *pimes*).

Utilitat del criteri lexicogràfic

Hem dit que els objectius aplicats del projecte han condicionat també la tria del criteri lexicogràfic. Un dels objectius prioritaris de l'Observatori ha estat la constitució d'un banc de dades de neologia no especialitzada que pugui servir de font per a l'actualització de diccionaris de llengua general o per a l'elaboració de diccionaris de paraules noves. Des d'aquesta perspectiva aplicada, el criteri utilitzat per l'Observatori és totalment vàlid.

La utilitat d'aquest banc de dades en el camp de la lexicografia és evident, perquè ofereix al lexicògraf un gran corpus de paraules que s'utilitzen en els textos de llengua general (orals i escrits), però que no estan incloses en el diccionaris de llengua més utilitzats i poden ser candidates a formar-ne part.

El corpus de neologismes de l'Observatori, per exemple, ha estat una de les fonts d'informació utilitzades per actualitzar el *Diccionari de la Llengua Catalana* (1995) de l'IEC, i el *Gran Diccionari de la Llengua Catalana* (1998) d'Enciclopèdia Catalana. D'altra banda, l'any 1998, l'Observatori de Neologia, juntament amb l'empresa editorial Enciclopèdia Catalana, va publicar el *Diccionari de paraules noves*, que conté les mil entrades corresponents als neologismes més freqüents en la premsa escrita durant el període comprès entre 1989 i 1996.

Eficiència del criteri lexicogràfic en la identificació automàtica de neologismes

A banda de la diversitat dels membres del grup de recerca i dels objectius aplicats del projecte en l'àmbit de la lexicografia, les possibilitats d'automatització del procés d'identificació i de buidatge de neologismes afavoreixen la tria del criteri lexicogràfic.

L'Observatori de Neologia disposa d'un programa automàtic de detecció de neologismes creat a l'IULA, el SEXTAN (sistema d'extracció automàtica de neologismes), que reaprofitja les eines de processament lingüístic del Corpus Tècnic de l'IULA. Un cop processats els textos de buidatge, la cadena de detecció, selecció i buidatge de neologismes del SEXTAN es basa en un corpus de diccionaris digitalitzats i desplegats morfològicament, de manera que el programa compara totes les ocurrències del corpus textual que s'ha de buidar amb els diccionaris i

proposa com a candidates a neologisme totes les paraules que no hi són registrades. En una interfície amigable, el programa presenta al neòleg tots els candidats a neologisme perquè els verifiqui o desestimi; cadascun dels neologismes verificats s'aboquen directament a una fitxa de buidatge que el neòleg només ha de completar amb aquelles informacions que no provenen directament del corpus textual.

La identificació automatitzada dels neologismes presenta, però, inconvenients perquè resulta insuficient per detectar determinats tipus d'unitats, concretament els neologismes semàntics i els neologismes formats per sintagmació. I, naturalment, necessita disposar del corpus de buidatge digitalitzat per poder aplicar-se. Això fa que l'Observatori utilitzi encara el buidatge manual per a aquests casos.

El criteri lexicogràfic en el buidatge manual d'aquests corpus és útil i permet un grau força elevat de sistematicitat, però també presenta algunes limitacions, perquè la seva eficiència depèn en gran mesura de les habilitats i els coneixements lingüístics que té cada persona que buida el textos de les obres lexicogràfiques de referència. En canvi, el criteri lexicogràfic és totalment eficient en el buidatge automàtic de neologismes formals, perquè permet objectivitat, sistematicitat i rapidesa en la detecció d'aquest tipus de paraules. A més, els neologismes resultants permeten o bé actualitzar el diccionari màquina o bé crear un mòdul lexicogràfic nou amb el conjunt de neologismes detectats, que s'activa o desactiva segons les necessitats de detecció.

Actualment, l'Observatori compta amb un banc de dades de 80.771 ocurrències de neologismes que no es troben en els diccionaris d'exclusió. Es tracta d'un material que permet fer un seguiment de la vitalitat lingüística de la llengua, però que no és prou elaborat o preparat perquè alguns neologismes de l'Observatori no es considerarien com a tals si, en lloc del criteri lexicogràfic, se seguís un criteri més cognitiu.

Limitació del criteri lexicogràfic i necessitat de filtres de neologicitat

Qualsevol parlant de la llengua s'adona, de manera espontània, que al banc de neologismes de l'Observatori hi ha unitats que, de manera intuïtiva, no semblen noves i, per tant, no serien considerades neològiques. Així, paraules com *antiavalots*, *centrecampista*, *càmera*, *descoordinació*, *postelectoral* o *subcampionat* no són percebudes com a noves per la majoria de parlants de la llengua catalana. Tampoc tots els parlants considerarien neològiques formes

com *a posteriori*, *ad hoc*, *de facto* o *ego*; o bé mots com *afició*, *hortera* i *carinyo*, o com *casting*, *catering*, *skin* i *abertzale*; o sintagmes com *fecundació in vitro* o *tràfic d'influències*. Segurament molt pocs parlants del català pensarien que la falta d'adaptació ortogràfica de *cassette* o la presència del guionet a *cotxe-bomba* són característiques suficients per considerar aquestes formes com a neològiques. I, malgrat això, tots aquest exemples estan recollits en el banc de dades de l'Observatori perquè no apareixen documentats en el corpus lexicogràfic d'exclusió.

Des del nostre punt de vista, l'ús del criteri lexicogràfic continua sent l'únic eficient per a la detecció automatitzada (i manual) de neologismes, però per tal que els neologismes recollits per l'Observatori puguin tenir aplicacions lexicogràfiques més refinades o altres aplicacions que les estrictament lexicogràfiques cal establir una sèrie de filtres formulats en forma de condicions que permetin atribuir a cada unitat inicialment neològica un valor determinat de neologicitat. Aquest valor matisaria el concepte de neologisme de cada unitat.

Creiem que alguns filtres es poden aplicar a qualsevol unitat amb independència del procés de formació, i que d'altres serveixen per mesurar la neologicitat d'un tipus concret de neologisme (sigui formal, semàntic o manllevat). Presentem a continuació alguns dels filtres a partir d'aquesta dicotomia; anomenem *filtres generals* els primers, i *específics* els segons.

Filtres generals

En relació als filtres generals, creiem que en podem establir almenys tres per distingir la neologicitat de les unitats independentment del procés de formació: la presència en altres fonts, la freqüència d'aparició en el banc de dades i la variació (ortogràfica o categorial) menor respecte al corpus lexicogràfic d'exclusió.

Ens referim en primer lloc a la **presència dels neologismes en altres fonts**. En principi, aquelles unitats considerades neològiques tindran un valor superior en la mesura que no apareguin en altres fonts, siguin lexicogràfiques o textuales, més enllà de les fonts utilitzades com a corpus d'exclusió. Existeixen fonts lexicogràfiques, en suport paper o electrònic, on podem documentar alguns neologismes, i podem relacionar l'aparició dels neologismes en altres fonts amb un grau de neologicitat menor. Així, per exemple, documentem les formes *datàfon* i *discjòquei* a la Neoloteca del Termcat; formes com *prió* al Diccionari Enciclopèdic de Medicina; formes com *decepcionar* i *desaire* al DCVB; i formes com *carinyós*, *medir* i

colmado en diccionaris d'interferències com el de Ruaix. A banda de les fonts lexicogràfiques, les unitats neològiques podran ser considerades més neològiques en la mesura que no apareguin en corpus textuais.

La **freqüència** amb què els neologismes han aparegut en el banc de dades de l'Observatori és també un criteri per establir la neologicitat de les unitats. La freqüència d'aparició es pot establir amb el control i creuament de dos paràmetres diferents: d'una banda, el nombre d'anys que han aparegut i, d'altra, el nombre d'ocurrències de cada any. Com més elevada és la freqüència d'aparició, més estabilitzat està el neologisme i, per tant, més s'ha perdut el sentiment de neologicitat.

Igualment, creiem que les unitats considerades neològiques d'acord amb el criteri lexicogràfic però que presenten un **canvi menor** respecte d'una variant repertoriada tenen un grau de neologicitat menor que aquelles que representen realment una unitat nova o que presenten un canvi més important respecte de la unitat documentada. Es tracta, en definitiva, de vacil·lacions dels parlants respecte a la norma ortogràfica establerta. Així, considerariem menys neològiques aquelles unitats que presenten canvis ortogràfics respecte d'una unitat documentada en el corpus lexicogràfic d'exclusió: alternances entre guionet i espai blanc, entre presència i absència d'*e* protètica o epentètica, etc. I també, sempre que no impliquin, a més, un canvi en el significat, les variacions categorials menors com l'alternança entre substantius i adjectius o l'alternança entre la transitivitat i la intransitivitat d'un verb.

Filtres específics

Més enllà d'aquests filtres generals podem identificar-ne d'altres si analitzem els neologismes agrupats segons el procés amb què s'han format o amb què han passat a ser unitats neològiques. Presentem a continuació aquests filtres, que hem denominat filtres específics, segons el tipus de neologisme.

Neologismes formals

Per atribuir un grau de neologicitat major o menor als neologismes formals cal avaluar el grau de predictibilitat de la regla amb què s'han format. Com més predictable és la regla de formació menys neològica es pot considerar. Creiem que la predictibilitat de la regla es pot mesurar tenint en compte, d'una banda, les característiques dels afixos i formants que hi entren en joc i de les bases implicades, i la predictibilitat del significat del neologisme resultant, de l'altra.

Per determinar el grau de neologicitat de les unitats formades per **prefixació**, per exemple, caldria tenir en compte filtres com els següents:

- En primer lloc, si el prefix està documentat com a entrada en el corpus lexicogràfic d'exclusió (com a *anticorrupció*) té un grau de neologicitat més baix que si no té entrada pròpia (com a *cibercafè*), en què el grau de neologicitat és més alt.
- En segon lloc, si el prefix s'adjunta a una base que no està documentada en el corpus lexicogràfic d'exclusió (com a *contraopa*), el grau de neologicitat és més alt que si la base està documentada.
- En tercer lloc, si el prefix s'adjunta a una base que no té les característiques “esperables” d'acord amb les restriccions categorials i semàntiques descrites en diccionaris i gramàtiques, el neologisme té un grau de neologicitat més alt que si la regla és absolutament previsible. Per exemple *extraverge* respecte d'*extracinematogràfic* o d'*extracomunitari*, i *antiningú* respecte d'*antitabaquista*.

Creiem que cal també tenir en compte altres aspectes com el fet que el prefix o el neologisme prefixat adquireixi matisos connotats des d'un punt de vista pragmàtic (siguin pejoratius o positius), o que sigui sinònim d'una entrada prefixada ja documentada en el corpus d'exclusió.

Amb criteris idèntics o semblants podem establir filtres de neologicitat per a la resta de processos de formació com la sufixació, la composició, la truncació, etc. En canvi, per a l'anàlisi de la neologia semàntica i dels manlleus cal proposar altres filtres específics.

Neologismes semàntics

Pel que fa als neologismes semàntics, considerem en primer lloc que la utilització de noms propis com a noms comuns dóna com a resultat unitats perceptivament molt neològiques, és a dir, amb un grau de neologicitat alt (per exemple, *barbie* o *bollicao* per referir-se a persones).

En segon lloc, que el canvi radical de significat en relació a la unitat documentada en el corpus lexicogràfic d'exclusió també comporta un grau de neologicitat major que si només es tracta d'una ampliació o restricció de l'accepció (per exemple, *pentinar* un terreny).

I, finalment, creiem que el salt d'un àmbit temàtic a un altre (independentment del grau de canvi de significat) es percep com un grau de neologicitat major que si el neologisme semàntic i la unitat documentada respecte a la qual és neològica pertanyen al mateix àmbit temàtic (per exemple, *vaselina* en esports).

Manlleus

Pel que fa als manlleus, proposem un grau de neologicitat més alt per als neologismes que no es troben documentats en obres lexicogràfiques de referència en la llengua d'origen; és a dir, també neològics en aquelles llengües segons el criteri lexicogràfic (per exemple, *bluff* / *caddie*). Proposem també un grau de neologicitat inversament proporcional al grau d'adaptació del manlleu al sistema lingüístic: com més adaptat està, menor és el grau de neologicitat, ja que demostra ser una unitat més estabilitzada (per exemple, *discjòquei* / *copyright*).

Finalment, aquells manlleus considerats neologismes pel criteri lexicogràfic que ja tenen en la llengua d'arribada un equivalent normativitzat o normalitzat documentat en les obres de referència, tenen un grau de neologicitat menor que si es tracta de neologismes manllevats que no tenen cap equivalent documentat en el corpus lexicogràfic (per exemple, *raclette* / *rentrée*).

Conclusions

La primera conclusió que volem destacar és que, malgrat les limitacions, el criteri lexicogràfic és realment l'únic criteri vàlid per identificar sistemàticament els neologismes dins dels textos (orals i escrits), si volem aconseguir una certa objectivitat en la selecció de candidats a neologismes, dur a terme una sèrie d'aplicacions lexicogràfiques, automatitzar al màxim la cadena de buidatge i reaprofitar els resultats en diccionaris màquina.

En segon lloc, hem vist com el criteri lexicogràfic presenta limitacions bàsicament perquè considera com a neologismes unitats que difícilment un parlant de la llengua consideraria com a neològiques (o en què, com a mínim, no hi hauria consens).

En tercer lloc, i per tal d'aconseguir que el banc de dades de neologismes de l'Observatori sigui multifuncional d'acord amb l'aplicabilitat que se li vulgui donar, cal establir diferents filtres de neologicitat, formulats com a condicions, aplicables a cadascun dels neologismes, que permetin d'establir per a cadascun dels filtres què es correspon a un grau alt de neologicitat i què a un grau baix de neologicitat.

En quart lloc, hem mostrat com, dels diversos filtres que entren en joc n'hi ha de menes diverses: filtres documentals, filtres pragmàtics, filtres de productivitat, filtres estrictament lingüístics que posen l'èmfasi en la forma o el significat del neologisme, etc. Com hem pogut comprovar, la presència dels corpus lexicogràfics en l'establiment de molts d'aquests filtres ha estat recurrent; caldrà trobar, doncs, mecanismes d'incorporació d'aquests filtres en el procés d'identificació automàtica dels neologismes en els textos per tal de refinar el sistema d'extracció a partir del criteri lexicogràfic.

Finalment, la combinació i el creuament de tots aquests filtres potser permetrà anar establint zones de neologicitat majors o menors, de manera que el criteri lexicogràfic complementat amb l'aplicació dels diferents filtres, i sempre tenint en compte l'aplicabilitat de les dades, escurci la distància entre els resultats obtinguts exclusivament per criteris tangibles o materials i per criteris cognitius.

Bibliografia

Cabré, M. T.; Freixa, J.; Solé, E. (1997) "A la limite des mots construits possibles". A: *Sillexicales*. 1997. Núm. 1, 65-78. [*Mots possibles et mots existants*. Forum de morphologie (1ères rencontres). Actes du colloque de Villeneuve d'Ascq (28-29 avril 1997).]

Cabré, M. T.; Bayà, R.; Bernal, E.; Freixa, J.; Solé, E.; Vallès, T. (2000) "Evaluación de la vitalidad de una lengua a través de la neología: A propósito de la neología espontánea y de la neología planificada". A: Jean-Claude Chevalier, Marie-France Delport (dir.). *La fabrique des mots. La néologie ibérique*. París: Presses de l'Université Paris-Sorbonne, 2000, p. 91-130.

Cabré, M. T.; Freixa, J.; Solé, E. (ed.). (2000) *La neologia en el tombant de segle. Actes del I Simposi sobre Neologia (1988) i Actes sobre el I Seminari de Neologia (2000)*. Barcelona: Universitat Pompeu Fabra, Institut Universitari de Lingüística Aplicada.

Cabré, M. T.; Freixa, J.; Solé, E. (2001) "Anàlisi contrastiva de la innovació lèxica en català i en castellà". *Caplletra*, 30, 199-212.

Cabré, M. T.; Domènech, M.; Estopà, R.; Freixa, J.; Solé, E. (en premsa) "L'Observatoire de Néologie: conception, méthodologie, résultats et nouveaux travaux". *Actes de L'innovation*

lexicale. Limoges: Université de Limoges, Faculté des Lettres et des Sciences Humaines. 1-3 de febrer de 2000.

Freixa, E.; Solé, E. (2000) “La identificació i el tractament dels neologismes”. A: Marí, I. (ed.). *Jornades per a la cooperació en l'estandardització lingüística*. Barcelona. Institut d'Estudis Catalans, 2000, p. 187-193. (Sèrie jornades científiques, 9).

Guilbert, L. (1975) *La créativité lexicale*. París: Larousse.

Rey, A. (1976) “Néologisme : un pseudo-concept ?”. *Cahiers de lexicologie*, 28, p. 3-17.

Vivaldi, J. (2000) “Sextan: prototip d'un sistema d'extracció de neologismes”. A: Cabré, M. T.; Freixa, E.; Solé, E. (ed.). (2000) *La neologia en el tombant de segle. Actes del I Simposi sobre Neologia (1988) i Actes sobre el I Seminari de Neologia (2000)*. Barcelona: Universitat Pompeu Fabra, Institut Universitari de Lingüística.