Aproximación cuantitativa a la neología

Rogelio Nazar

Institut Universitari de Lingüística Aplicada
Universitat Pompeu Fabra
Pl. de la Mercè 10-12
08002 Barcelona

Vanesa Vidal

Fundació Barcelona Media
Universitat Pompeu Fabra
Ocata, 1
08003 Barcelona

{rogelio.nazar; vanesa.vidal}@upf.edu

ABSTRACT: El criterio lexicográfico en la detección de neología presenta algunos problemas interesantes para nuestro estudio: (i) muchos de los candidatos a neologismo distan considerablemente de la percepción que posee el hablante de una lengua de lo que es una "palabra nueva" (por ejemplo, candidatos como *neurofarmacología* o *prejubilado*); (ii) las unidades no neológicas de estructura sintagmática raramente aparecen como lemas en el corpus lexicográfico de exclusión, por lo que dentro de los candidatos de este grupo encontramos unidades como *a contrarreloj* de bajo grado de neologicidad junto a otros neologismos más prototípicos del tipo *familia monoparental* y, finalmente, (iii) existe el problema de los neologismos semánticos, cuya forma es idéntica a la que existe documentada en los diccionarios. Las variables cuantitativas en el análisis de los datos ofrecen una información importante sobre el comportamiento de las unidades candidato-a-analizar en el tiempo que permite evaluar el estatuto neológico de una unidad. Por tanto, el método cuantitativo se concibe aquí como un complemento del tradicional método lexicográfico. En el presente artículo realizamos una serie de experimentos para asignar a una expresión un grado de neologicidad con un fundamento empírico, utilizando como corpus los archivos del diario EL PAÍS en el período 1976-2007. En algunos experimentos tomamos como base una muestra de los neologismos del Observatorio de Neología y evaluamos su correlato estadístico (en el caso de que aparezcan en nuestro corpus) con la finalidad de obtener conclusiones relevantes sobre su grado de neologicidad. En otros experimentos planteamos la extracción automática de candidatos a neologismo, incluyendo la clase más compleja, representada por los neologismos semánticos.

1. Introducción

1.1. El método estadístico como complemento del tradicional método lexicográfico en la detección de neologismos.

En tanto organismo viviente, el lenguaje se encuentra en constante mutación, por ello la tarea del especialista en neología es advertir las unidades que pasan a formar parte del sistema de la lengua. El ritmo al que se produce esta renovación del vocabulario es vertiginoso. Terminologías de diversos ámbitos científico-técnicos se suman al lenguaje, surgen nuevos conceptos, nuevos usos y costumbres y con ellos la necesidad de su denominación y de su percepción social (Cabré, 2000; 2002). Pero por otro lado surgen también nuevos nombres de personajes y entidades que comienzan a circular en los textos de cada época. Entonces la complejidad del problema consiste precisamente en distinguir cuáles, de todas las unidades nuevas, tienen que pasar a ser tenidas en cuenta desde el punto de vista lexicográfico.

La aplicación del criterio lexicográfico (Vivaldi, 2000) consiste en detectar la presencia de una determinada unidad en un corpus de exclusión compuesto por diccionarios representativos de una lengua dada. Se trata de un criterio categórico y objetivo que puede ser fácilmente automatizable, basta con contrastar una lista con las palabras del corpus analizado con la lista de los lemas de las obras lexicográficas que integran el corpus de exclusión. El principal problema (1) de este criterio es que muchos de los candidatos a neologismo (unidades que no están recogidas en los diccionarios) distan de lo que el hablante de una lengua percibe como palabra nueva; porque incluyen tecnicismos o nombres de las distintas entidades del mundo que no se consideran parte del

vocabulario propio de una lengua. Muchas de estas unidades carecen de valor semántico fuera de su función referencial, y, si tuviesen que ser registradas, su lugar correspondería a la enciclopedia y no al diccionario. Aún así, muchas de estas unidades acaban formando parte del sistema de la lengua cuando designan categorías generales u objetos. Es el caso de nombres de marcas como *Kleenex* para designar a los pañuelos de papel, cuyo uso ya está implantado en la lengua. Otro problema (2) se da en el caso de las unidades de estructura sintagmática, a las que no corresponde una entrada propia en los diccionarios. Existe también el problema (3) de los neologismos semánticos, que sí aparecen, aunque con otro sentido, en el corpus de exclusión. El vaciado manual, por su parte, también es problemático (4), porque un informante puede no conocer un vocablo que existe desde hace décadas en un circuito que no es el suyo, sea por motivos generacionales o culturales.

El método que planteamos aquí se concibe entonces como un complemento del tradicional método lexicográfico y el vaciado manual. En este trabajo planteamos diversas líneas de investigación sobre la base de estos problemas enumerados. En cuanto al corpus utilizado en los experimentos, se trata en este caso del archivo de 30 años del diario EL PAIS, que ha sido suficiente para este experimento. De cualquier modo, si esta metodología se quisiera aplicar a la extracción real de neología, entonces debería alimentarse con datos de distintas publicaciones periódicas, atendiendo a las posibles variaciones regionales que pueda presentar la lengua analizada.

1.2. Líneas de investigación

1.2.1. El método estadístico, filtro que permite establecer diferentes grados de neologicidad: En relación al problema (1), se hace necesario filtrar los candidatos a neologismo. Para ello trabajamos en el establecimiento de filtros de neologicidad que nos permitan establecer grados de neologicidad y aproximarnos a lo que se percibe como nuevo en la lengua. Para ello, el estudio de la distribución de las frecuencias de aparición de las unidades a lo largo del tiempo se percibe como el recurso idóneo.

1.2.2. El método estadístico y la detección de neologismos sintagmáticos: En relación con el problema (2), trabajamos en la detección de combinaciones de palabras o concurrencias, si definimos este término como la ocurrencia conjunta de los elementos en forma adyacente en el sintagma (Vidal et al., 2006). Así, la palabra *familia* no es un neologismo pero su combinación recurrente con *monoparental* sí es nueva.

1.2.3. El método estadístico y la detección de neologismos semánticos: En relación con el problema (3), podríamos trabajar también en la detección de neologismos semánticos. Podemos hacer esto mismo también por medio del estudio de las concurrencias, definiendo esta vez *concurrencia* como la aparición conjunta de dos unidades a una distancia flexible, y sin importar el orden, dentro de una ventana de contexto de *n* palabras. Normalmente una unidad léxica posee uno o más perfiles de concurrencia léxica en un tiempo *t*. La unidad *X*, unidad cuyo sentido se describe en el diccionario de exclusión, comparece típicamente con las unidades *Y*, *Z*, *W*, etc. Pero ese perfil puede variar con el tiempo, la misma unidad *X* puede comenzar a presentar concurrentes distintos y es entonces cuando estamos ante un neologismo semántico *X'*.

1.3. Objetivos del trabajo

1.3.1. Presentar esta investigación por su interés teórico desde el punto de vista de la lingüística cuantitativa, ya que presenta una metodología de estudio y caracterización del comportamiento de unidades léxicas.

1.3.2. Aplicar los resultados de la investigación como metodología complementaria de validación de candidatos a neologismo (obtenidos con cualquier método).

1.3.3. Proponer una metodología para la extracción automática de candidatos a neologismo.

1.4. HIPÓTESIS DE TRABAJO

Existe una relación entre el grado de neologicidad de una expresión y los siguientes parámetros:

1.4.1. La distribución de la frecuencia de dicha unidad en el corpus: La curva de distribución de frecuencias de un *neologismo ideal* (neologismo de alto grado de neologicidad) del corpus tendrá una pendiente ascendente hacia el año 2007.
1.4.2. El nivel de implantación de dicho neologismo en la lengua. Los neologismos de mayor grado de neologicidad presentan una pendiente más acusada y el período temporal en el que se utiliza es menos extenso.
1.4.3. En el caso de los neologismos semánticos, observaremos un cambio en el perfil de concurrencia. Cuando estamos ante una unidad que adquiere un nuevo sentido, podremos advertir la existencia de un grupo de contextos de ocurrencia de esa unidad con una serie de palabras en común.

2. METODOLOGÍA DE TRABAJO: LOS EXPERIMENTOS

Como una breve introducción a los experimentos aquí realizados, aclararemos que nuestro trabajo se orienta en dos direcciones: establecer diferentes grados de neologicidad entre los candidatos obtenidos después de haber aplicado un filtro lexicográfico (como en los experimentos 1 y 2, secciones, 2.1. y 2.2, respectivamente) e identificar neologismos automáticamente, diferenciando neologismos de alto grado de neologicidad de otros de grado bajo o nulo. En cuanto a esta segunda dirección, trabajamos simultáneamente desde dos puntos de vista: por un lado formulamos hipótesis para validarlas o refutarlas a partir del análisis de los datos del corpus, planteando el comportamiento que debería tener un neologismo *prototípico* (como en el experimento 3, sección 2.3) y por otro lado analizamos primero el corpus para llegar a regularidades sobre lo que esperamos obtener, como cuando detectamos neologismos semánticos por medio de la clasificación de sentidos directamente a partir de los datos (experimento 4, sección 2.4).

Las gráficas que aparecen a continuación representan los archivos del diario EL PAIS en el período 1976-2007. El eje horizontal representa los años y el vertical la frecuencia relativa de la unidad (relativa para compensar las posibles diferencias en el tamaño de las muestras correspondientes a los años). Cada posición, es decir cada año, representa todas las ediciones del diario en ese año.

2.1. PRIMER EXPERIMENTO: EXPLORACIÓN DE UNIDADES SIMPLES Y SINTAGMÁTICAS EN EL CORPUS

Este primer experimento consistió en introducir algunas unidades que pensamos podrían describir una curva ascendente y ser neologismos desde este punto de vista, como *teléfono móvil* y *teléfono fijo* (figura 1). Curvas ascendentes similares son las de otros ejemplos como *familias monoparentales* o *cambio climático* (figura 2).

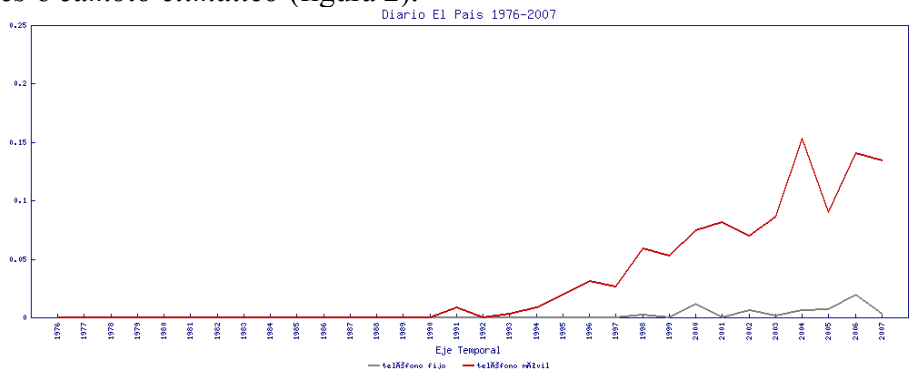


Figura 1: "teléfono móvil" (curva superior) y "teléfono fijo" (curva inferior).

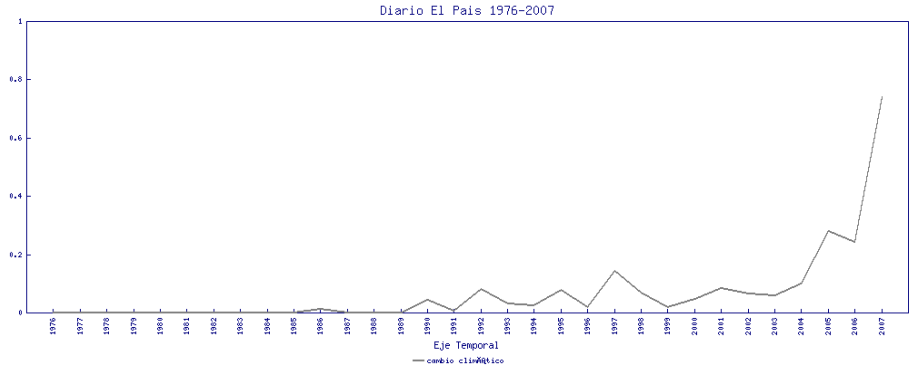


Figura 2: cambio climático

2.2. Segundo experimento: evaluación de candidatos del Observatori de Neologia

Este segundo experimento consistió en tomar una muestra aleatoria de 100 neologismos del conjunto de los registrados por el *Observatori de Neologia* en el 2007. Entre ellos hay algunos que consideraríamos neologismos de alto grado de neologicidad, como *sms*, *spam* o *blog* (figura 3) caracterizados por una curva ascendente y una pendiente alta.

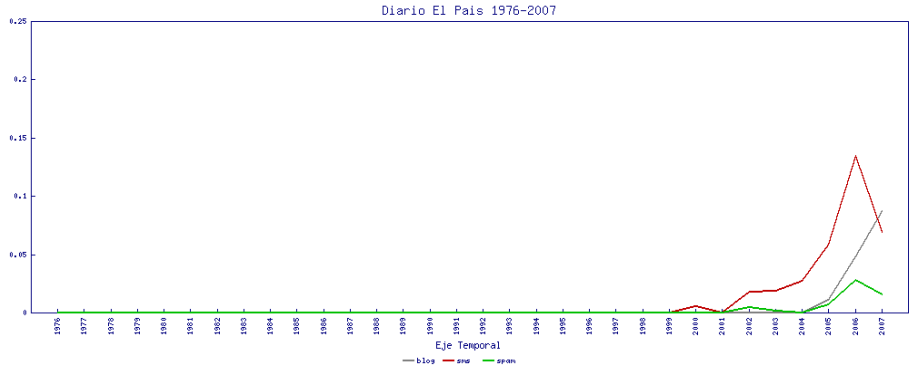


Figura 3: sms, spam y blog.

Hay casos que muestran menor grado de neologicidad, como *crooner* (figura 4) y otros, como el de *palabra de honor* (figura 5), del que nos ocuparemos en el cuarto experimento (sección 2.4). Este último es un caso de neología semántica ya que, además de su significado recto, en los últimos años esta expresión también se utiliza para designar un tipo de escote. Este último ejemplo pone de manifiesto las limitaciones de esta parte de nuestra metodología, pero como veremos más adelante, podemos resolver el problema de la polisemia con técnicas de clustering. Más de allá de este problema específico, los resultados del segundo experimento sugieren que podríamos utilizar la técnica descrita para filtrar los neologismos propuestos con el criterio lexicográfico o con el vaciado manual para quedarnos sólo con aquellos que tengan una curva prototípica, en caso de aparecer en el corpus de trabajo.

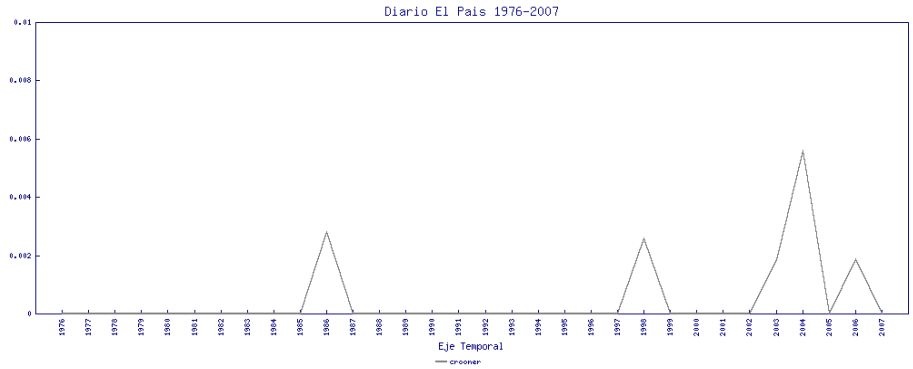


Figura 4: crooner

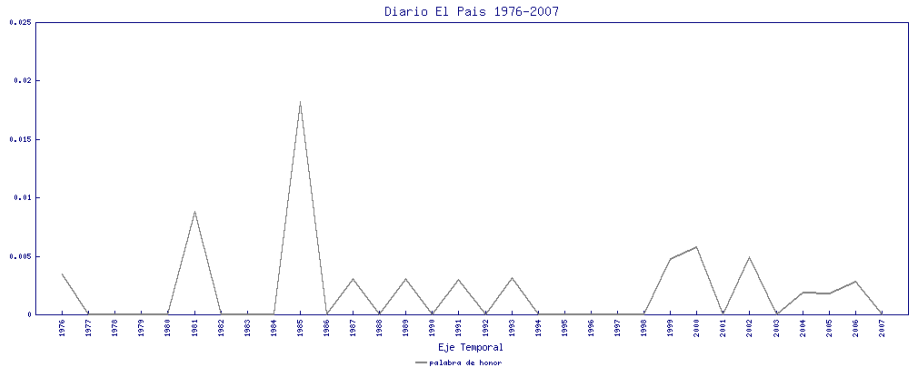


Figura 5: palabra de honor

2.3. TERCER EXPERIMENTO: EXTRACCIÓN AUTOMÁTICA DE CANDIDATOS A NEOLOGISMO

Este experimento consistió en extraer automáticamente los candidatos a neologismo a partir del corpus. Para ello tomamos una muestra aleatoria de unidades del año 2007 y seleccionamos aquellas cuya curva se parezca más a la de un neologismo ideal o prototípico. Suponiendo que el intervalo [1976-2007] corresponde a los valores de *x* [1-31], este neologismo teórico se puede expresar con la fórmula 1 y representar gráficamente con la figura 6.

$$f(x) = x^{10}$$

Fórmula 1: definición del neologismo ideal o teórico

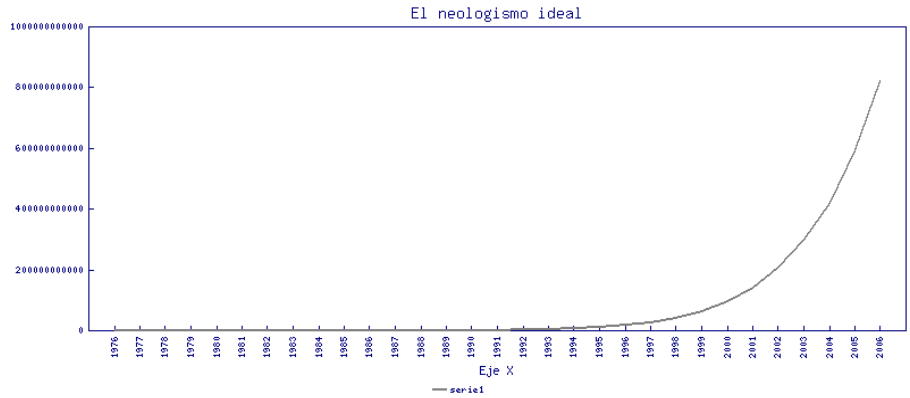


Figura 6: Gráfica del neologismo ideal

Podemos definir cada unidad del corpus con su respectiva curva de frecuencia diacrónica, seleccionando así las *n* curvas más próximas. Para comparar las curvas de cada unidad con la de nuestro neologismo prototípico medimos la distancia euclideana, expresada en la fórmula 2:

$$\sqrt{\sum \left(X_i - Y_i\right)^2}$$

Fórmula 2: distancia euclideana.

El símbolo *X* sería nuestro neologismo teórico e *Y* la unidad candidato-a-analizar. Antes de proceder así es necesario llevar las distintas curvas a la misma escala[1]. La manera de hacerlo es normalizar los datos de las frecuencias tal como se indica en la fórmula 3:

$$x'_i = \frac{x_i}{argmax\left(X\right)}$$

Fórmula 3: procedimiento de normalización.

Con esta normalización, los valores máximos de dos curvas de distinto tamaño pasan a tener un valor 1, de modo que el resto de los valores se vuelven proporcionales a esa unidad, y por lo tanto las curvas se vuelven comparables. Algunas curvas extraídas del mismo corpus y que tienen un aspecto relativamente similar al de nuestro neologismo teórico son las de *ley de igualdad* o *enaltecimiento del terrorismo* (figura 7).

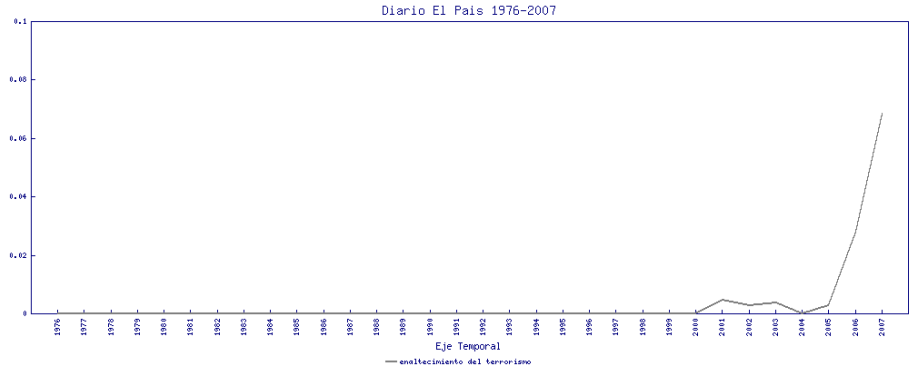


Figura 7: enaltecimiento del terrorismo

Si ordenamos algunos de los ejemplos que hemos extraído del corpus de acuerdo a la distancia euclideana que tienen respecto al neologismo ideal, obtenemos el listado de la tabla 1, que muestra a simple vista que hay un correlato entre el grado de neologicidad y similitud con el ideal.

Vector:	**Distancia:**
neologismo ideal	0.00000
ley de igualdad	0.56512
cambio climático	0.56977
sms	0.61388
spa	0.63034
blog	0.65528
enaltecimiento del terrorismo	0.72075
spam	0.73037
vida de los otros	0.83607
teléfono fijo	1.07240
teléfono móvil	1.13533
kale borroka	1.40737
familias monoparentales	1.44058
crooner	1.52033

1 Recordemos que los datos de frecuencias estaban ya expresados en valores relativos para compensar las posibles diferencias de tamaño entre las distintas muestras anuales.

palabra de honor	1.76737
multiétnico	1.81140
teléfono	2.69050

Tabla 1: Ejemplos ordenados de acuerdo al grado de neologicidad propuesto.

Si hiciéramos lo mismo con todas las unidades que aparecen en el año 2007 seríamos capaces de extraer aquellas que tengan el mayor índice de neologicidad. Para demostrarlo llevamos a cabo el siguiente experimento: tomamos una muestra aleatoria de 450 palabras o cadenas de palabras de hasta cuatro unidades de largo y con una frecuencia absoluta de 30 apariciones en el año 2007 en nuestro corpus. Examinando manualmente esa lista de palabras descubrimos sólo cuatro palabras que podríamos considerar, al menos relativamente, neologismos: *disco duro*, *hipotecas subprime*, *vivienda protegida* y *terrorista suicida.* Ordenando estas 450 palabras de acuerdo a la distancia que tienen con el neologismo ideal, los cuatro neologismos que hemos señalado quedan entre las 30 primeras posiciones, rodeados de otras palabras que son mayoritariamente nombres de personajes o entidades que han pasado a formar parte o han tenido prominencia en la agenda temática de los medios recién en los últimos años, como *Lavrov*, la revista *Forbes*, *Stamford*, *Heidfeld*, etc. Estos neologismos espurios se pueden eliminar con facilidad ya que su probabilidad de aparición con mayúsculas en el corpus es mucho mayor. Con este filtrado de los nombres propios el grupo de las 30 primeras posiciones donde se concentran los neologismos verdaderos se reduce prácticamente a la mitad, mejorando así el filtrado.

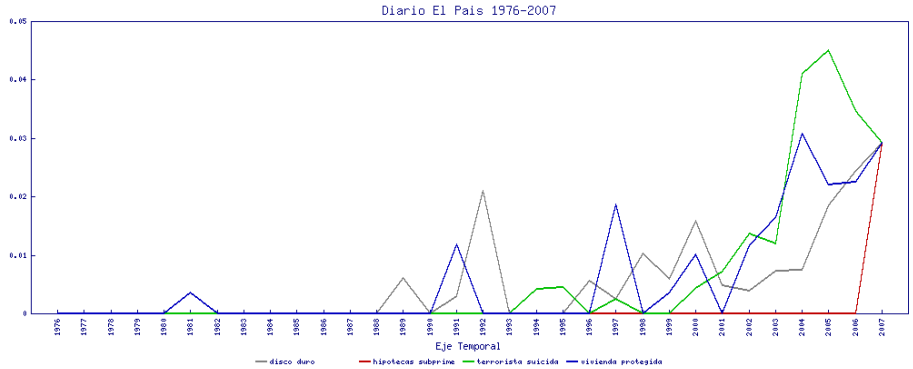


Figura 8: *disco duro*; *hipoteca subprime*; *terrorista suicida* y *vivienda protegida*.

Podemos apreciar también la relación que existe entre las curvas de la figura 8 y el grado de implantación de estos neologismos, tal como esperábamos según nuestras hipótesis (sección 1.4.2). Una unidad como *disco duro* presenta alta implantación, mientras que una unidad como *hipotecas subprime* muestra un alto grado de neologicidad.

2.4. Cuarto experimento: extracción automática de neologismos semánticos

Seguramente los neologismos semánticos son el tipo de neologismo que más dificultades ofrece para su vaciado automático (Tebé, 2002). Sin embargo no creemos que sea una tarea imposible, como se pensaba hace algunos años (Vivaldi, 2000). Es fácil intuir que lo que distingue a los distintos sentidos de una expresión es el contexto, tal como muestran los histogramas que se exhiben en la tabla 2. En estos histogramas la unidad analizada ocupa la posición 0 en el eje horizontal, mientras que el vertical representa la frecuencia. Así, en el primer caso, con la forma *bajo*, observamos que en la mayor parte de los casos ocupa la posición -1 respecto a *palabra de honor*, como se muestra en el ejemplo.

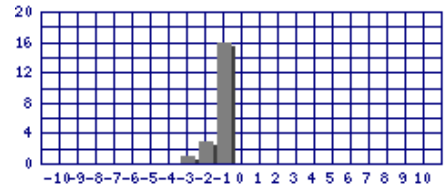
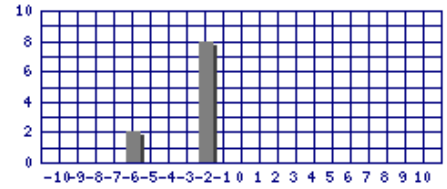
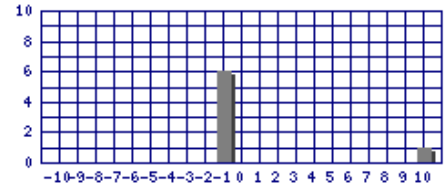
Forma:	*Casos:*	*Ejemplo:*	*Histograma:*
bajo	16	*...de los gal , que afirma lo que afirma **bajo palabra** de honor . 6 . que en este contexto ,...*	
dado	8	*...divulgar el nombre del donante , porque había **dado** su **palabra** de honor . tuvo que pagar una multa para evitar...*	
escote	6	*....la actriz llevaba un vestido verde agua con **escote palabra** de honor , diseñado por oscar de la renta ....*	

Tabla 2: Ejemplos de concurrencias léxicas significativas de *palabra de honor*.

Dado este comportamiento, y con el propósito de caracterizar los distintos sentidos de una unidad polisémica aplicamos en este experimento un algoritmo de clustering que sigue la pauta de lo planteado en nuestras hipótesis (sección 1.4.3). El clustering se lleva a cabo mediante la construcción de grafos que caracterizan las concurrencias léxicas más significativas de la unidad analizada. La forma en que se construyen estos grafos se explica en Nazar et al. (2007) y Nazar et al. (en prensa). La implementación del presente experimento recorre la línea del tiempo del corpus de El País y extrae los contextos de ocurrencia (15 palabras a un lado y al otro) de una unidad analizada (en este caso el neologismo semántico *palabra de honor*). De estos contextos elimina las palabras menos informativas utilizando un modelo de la lengua castellana. El modelo consiste en un listado de frecuencias de palabras sobre un corpus de lengua general de tres millones de palabras que hemos extraído del dominio *Lenguaje General* del Corpus del IULA (Cabré et al., 2006). Las palabras con una frecuencia superior a 100 en ese modelo se consideran prescindibles y son eliminadas del análisis. Un módulo de reconocimiento de similitud morfológica agrupa distintas unidades (como *empeñada, empeñó, empeñar*) bajo una sola forma, lo que beneficia el conteo de frecuencias sin necesidad de disponer de un lematizador. Finalmente, a las restantes unidades (simples, no sintagmáticas) se les asigna un *nodo*, que es simplemente un apuntador a los contextos de ocurrencia de esa unidad léxica, y un *arco* entre los nodos de las unidades léxicas que aparecen en los mismos contextos, fortaleciendo la ponderación (un valor numérico) del arco entre los nodos que coocurren más frecuentemente. Uno de los modos de llevar a cabo esta ponderación entre un nodo *i* y un nodo *j* es la fórmula 4, donde *x* representa la unidad analizada en cuyos contextos ocurren *i* y *j*. La variable *N* es el número total de contextos donde aparece *x*.

$$R_{ij}(x) = \log\left(\frac{F_{ij}(x)}{N}\right)$$

Fórmula 4: Ponderación de un arco entre un nodo *i* y un nodo *j*.

El resultado es una matriz que implícitamente encierra grupos de documentos. Los documentos son contextos de ocurrencia de la unidad analizada, asociados a la fecha de su aparición. Para explicitar esta agrupación en clases recorremos el grafo desde los nodos centrales hacia los periféricos, siendo los centrales los que tienen mayor cantidad de arcos. Cada vez que llegamos a un nodo, incluimos en una misma clase a los documentos vinculados a este nodo. Si en determinado punto del proceso dos clases comparten más del 40% de los documentos, entonces fundimos las dos clases en una sola. Finalmente, y si tenemos como condición que se retenga únicamente grupos de más de dos

contextos, obtenemos en el caso de *palabra de honor* solo dos grupos o clusters, expuestos en la tabla 3. Los nombres que etiquetan los clusters son asignados automáticamente a partir del nodo que tiene mayor cantidad de arcos.

Cluster 1: **empeñar**		Cluster 2: **escotes**	
Términos	***Contextos:***	***Términos***	***Contextos:***
1) ap 2) astarloa 3) barrionuevo 4) confederal 5) consentido 6) credulidad 7) cuan 8) empeñar 9) esclarece 10) escudero 11) fusté 12) herrero 13) incité 14) inocencia: 15) proclamar 16) quebrantamiento 17) reiterado 18) tejero	1) 19790420.txt 2) 19810501.txt 3) 19850524.txt 4) 19850913.txt 5) 19880331.txt 6) 19970520.txt 7) 19970814.txt 8) 19980908.txt 9) 20041119.txt	1) copresidente 2) cubren 3) drapeados 4) escotes 5) gucci 6) marrón 7) modista 8) ojito 9) organza 10) swarovski 11) tonos	1) 19981019.txt 2) 20020203.txt 3) 20020930.txt 4) 20070118.txt 5) 20071201.txt

Tabla 3: clustering de concurrencias léxicas de *palabra de honor*

La tabla 3 nos muestra, del lado derecho, un grupo (*Cluster 1*) de palabras que suele aparecer con la unidad analizada. En este caso son palabras que relacionaríamos con el sentido recto de la expresión. En la columna de *Documentos*, dentro de este mismo grupo, observamos que las fechas de los contextos de ocurrencia se extienden a lo largo de un período extenso de tiempo (1979-2004). Lo contrario ocurre con el grupo de la derecha de la tabla. Las palabras en este caso pertenecen al ámbito de la moda y los fechas de los contextos de ocurrencia se concentran en la última década (1998-2007).

3. Conclusiones

Conclusiones parciales sobre cada experimento:

Experimento 1: Existe una relación apreciable entre las unidades que como hablantes de la lengua consideramos de un alto grado de neologicidad y la curva ascendente que caracteriza su frecuencia de uso en el tiempo.

Experimento 2: Es posible utilizar estas mediciones para filtrar los neologismos propuestos con el criterio lexicográfico o con el vaciado manual, resaltando aquellos cuya curva más se parezca a la de un neologismo prototípico.

Experimento 3: Las curvas más similares a la curva prototípica que hemos definido a priori para los neologismos de alto grado de neologicidad corresponden efectivamente a neologismos de este tipo, por lo cual podemos extraer candidatos a neologismo automáticamente.

Experimento 4: A partir de las mediciones que hemos hecho en el experimento de extracción de neología semántica será posible para un algoritmo advertir el nuevo uso de una expresión y almacenarlo en un diccionario como neologismo semántico, con la opción de incluir también ejemplos de palabras que conforman su nuevo entorno.

Conclusiones generales:
Hemos presentado en este artículo una metodología de análisis en lingüística cuantitativa para el estudio del fenómeno de la neología. No quisiéramos que el presente trabajo sea categorizado sólo como una técnica informática o una forma de automatización de un proceso. Creemos que el cambio de enfoque que implican estos métodos de análisis cuantitativo nos permite ir más allá de la solución de un problema técnico concreto. La oportunidad que tenemos hoy de extraer con rigor científico datos de los textos del mundo representa una fuente de información que trasciende el conocimiento que nos puede proveer el hablante individual de una lengua. Por eso el método representa un giro importante en el punto de vista ya que nos permite plantear la neología como un problema puramente geométrico. El interés desde el punto de vista puramente ingenieril también existe, ya que esta metodología se ofrece como una técnica eficiente y poco costosa de implementar a escala masiva. El acceso a los archivos digitales de distintos periódicos es cada día más fácil, por tanto la disponibilidad de corpus ya no es un problema.

4. Referencias

CABRÉ, T. (2000). "La neologia com a mesura de la vitalitat interna de les llengües". en M. T. Cabré, J. Freixa, E. Solé (ed.). *La neologia en el tombant de segle: I Simposi sobre Neologia* (18 de desembre de 1998), *I Seminari de Neologia* (17 de febrer de 2000), pp. 85-108. Barcelona: Observatori de Neologia, IULA, 2000. ISBN: 84-477-0712-1.

CABRÉ, T. (2002). "La neologia, avui: el naixement d'una disciplina", en M. T. Cabré, J. Freixa, E. Solé (ed.). *Lèxic i neologia.* Barcelona: IULA, DOCUMENTA UNIVERSITARIA, 2008. ISBN: 978-84-96742-48-2.

CABRÉ, T., BACH, C. y VIVALDI, J. (2006) "10 anys del Corpus de l'IULA", Serie Papers de l'iula, Informes, Barcelona, IULA/INF044/06.

NAZAR, R.; VIVALDI, J. y WANNER, L. (2007) "Towards Quantitative Concept Analysis", Procesamiento del Lenguaje Natural, n. 39, pp. 139-146. Septiembre, 2007. ISSN 1135-5948.

NAZAR, R.; VIVALDI, J. y WANNER, L. (en prensa) "Hacia una analítica cuantitativa de los conceptos", Serie Papers de l'IULA, Universidad Pompeu Fabra.

TEBÉ, C. (2002) "Bases pour une sélection de neologismes". en Cabré, M. T.; Freixa, J. y Solé, E. *Lèxic i neologia.* p. 43-50. Barcelona. Institut Universitari de Lingüística Aplicada.

VIDAL, V.; ESTOPÀ, R.; PALACÍN, V. y SOUTO, M. (2006). "Neologismes formats per composició i sintagmació". III Seminari de Neologia, Barcelona (España), Observatori de Neologia.

VIVALDI, J. (2000). "Sextan: prototip d'un sistema d'extracció de neologismes", en M. T. Cabré, J. Freixa, E. Solé (ed.). *La neologia en el tombant de segle: I Simposi sobre Neologia* (18 de desembre de 1998), *I Seminari de Neologia* (17 de febrer de 2000), p. 85-108. Barcelona: Observatori de Neologia, IULA, 2000. ISBN: 84-477-0712-1.