

**Resumen:** En el seminario de hoy presentaré los resultados preliminares de un trabajo en curso sobre detección de neologismos semánticos desde el enfoque de la lingüística cuantitativa. Este trabajo se enmarca en el proyecto APLE<sup>1</sup> y en el conjunto de otros trabajos con colegas del IULA. En esencia, se trata de la detección de neología semántica por medio del análisis de las relaciones de coocurrencia de las palabras. Lo que se describe es un experimento con el corpus diacrónico Google Books Ngrams consistente en un algoritmo que, a partir de una lista de verbos que aparecen en el diccionario, selecciona aquellos que han experimentado recientemente un cambio semántico. El corpus de Google está conformado por secuencias de hasta cinco palabras y la metodología de este experimento consiste en estudiar cambios en la combinatoria de cada verbo en el período comprendido entre 1975 y 2008. Si bien en este experimento se analizan verbos, no hay condición que impida repetirlo con otras categorías gramaticales.

## 0. Introducción

En general, la visión que tenemos de los textos es que son objetos definidos y concretos, como monumentos que trascienden históricamente al autor. Pero lo cierto es que las palabras cambian su significado con el tiempo y adquieren nuevas connotaciones debido a los más variados hechos históricos, de manera tal que el texto que el autor tenía intención de expresar se interpreta mejor en el contexto imaginado por el autor mismo. Inexorablemente, el texto que hoy producimos con el tiempo va a desfigurarse gradualmente perdiendo parte de su sentido original. Por eso, y sea o no un pensamiento optimista, posiblemente sea cierto que sólo se escribe para los contemporáneos, aunque le quede un papel a la filología en su esfuerzo por la recuperación del sentido original de los textos.

Con el término *neología semántica* nos referimos al fenómeno por el cual palabras formalmente conocidas adquieren un significado nuevo. Podemos decir que el estudio de la neología semántica forma parte del estudio del cambio semántico, y es por tanto una clase particular de lingüística diacrónica, donde existe ya una tradición y una clasificación de tipos de cambio de significado. Este campo de estudio es de una complejidad mayor al de la lingüística sincrónica, y posiblemente sea también más fascinante. Desde el punto de vista teórico, las posibilidades son muy diversas. Además del caso particular de la neología semántica, podríamos también abordar el estudio de la obsolescencia semántica o de distintos tipos de cambio que se examinarán en las siguientes secciones.

Paralelamente a la teoría, también existe una diversidad de posibilidades en cuanto a la aplicación práctica. Por poner un ejemplo, en la industria editorial existe una gran necesidad de automatización de

---

1 Este trabajo ha recibido el apoyo del Ministerio de Ciencia e Innovación de España a través del proyecto APLE (Procesos de actualización del léxico español a partir de la prensa editada en Cataluña FFI2009-12188-C05-01 subprograma FILO. Investigadora principal: Teresa Cabré). Para poder procesar un corpus del tamaño de Google Books, fue necesario utilizar la plataforma de cálculo del Departamento de Tecnologías de la Información y la Comunicación de la Universidad Pompeu Fabra. Agradezco la gentil ayuda de Silvina Re y Milton Hoz de Vila para operar esta maquinaria. Agradezco también la colaboración de Irene Renau, de quien he tomado muchas ideas para este trabajo.

los procesos de actualización de diccionarios, y hasta ahora esta industria no ha podido resolver de forma suficientemente satisfactoria, entre muchos otros factores, el problema de la neología semántica. No encontramos, al menos, publicación alguna que contradiga esto en la literatura sobre lexicografía, como las actas de congresos de Euralex, el *International Journal of Lexicography* o los *Cahiers de Lexicologie*. En esta presentación comentaremos, sin embargo, algunas publicaciones en las que se intenta plantear el problema en términos de la lingüística computacional.

A la pregunta por el interés de estudiar la neología semántica, podemos responder que hay distintas formas de aproximarse al objeto. El primer paso, que es el más elemental, es la recopilación de las instancias de un neologismo. Un paso subsecuente es su clasificación. El tercer paso sería la interpretación o la explicación del cambio semántico, que es también una consecuencia de la clasificación. Finalmente, y lo que resulta más interesante desde el punto de vista científico, es que una teoría tenga poder predictivo, y por tanto, nos permita disponer de un método objetivo para saber qué unidades léxicas están desarrollando un nuevo sentido. Encontrar de manera automática casos de neología semántica ejemplifica una de las diferencias fundamentales entre los métodos cuantitativos y cualitativos en lingüística. Con un método empírico y cualitativo podemos analizar los contextos de aparición de un determinado vocablo y comprobar que en él se produce el surgimiento de un significado nuevo. Los métodos cuantitativos, en cambio, nos permiten descubrir cuáles son esas unidades y esto es, evidentemente, un avance importante.

La estructura de la presentación es la siguiente. Sección 1: un análisis exploratorio del corpus, en el que intentaremos medir el crecimiento del vocabulario castellano. Sección 2: un repaso de distintos métodos de análisis de neología, de manera muy esquemática por razones de espacio. Sección 3: la hipótesis y la metodología de este experimento. Sección 4: resultados preliminares que apoyan la hipótesis de partida. Sección 5: conclusiones y líneas de trabajo futuro.

## **1. Análisis exploratorio**

### **1.1. El material estudiado**

Cualquier aproximación al estudio de la dinámica del vocabulario castellano requiere una muestra representativa del objeto, y para ello resulta fundamental la selección del corpus. En trabajos previos sobre neología (Nazar & Vidal, 2010; Nazar & Azarian, 2010; Cabré & Nazar, en preparación) utilizamos corpus de prensa. Sin embargo, este tipo de muestras tienen el problema de que el género, pese a ser un gran indicador de las novedades de una lengua, no puede considerarse plenamente representativo de la lengua castellana normal o neutral, por diversos factores, como la restricción temática que imponen las secciones, la abundante aparición de la terceras personas en detrimento de las primeras, el orden y estructura especiales en que los periodistas se expresan en sus diversos géneros textuales, etc. Los géneros no informativos de la prensa (crónica, reportaje, perfil del personaje, etc.), además, producen una lengua de gran expresividad que no siempre representa neologicidad generalizable al idioma. Oraciones del tipo de (1), por ejemplo, no son del todo infrecuentes<sup>2</sup>.

(1) *...fue una fashion-victim, con bulimia incluida, una barbie incómoda pero desafiante...*

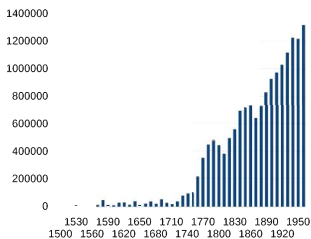
2 La frase es de la periodista Margarita Rivère, en "El primer mártir del imperio de la realidad virtual" (Diana, el Mito, Revista de La Vanguardia, domingo 7 de setiembre de 1997).

Esta es una de las razones por las que resulta atractivo un corpus de libros como el de Google Books (Michel et al. 2010). En comparación, en un corpus de estas características cabe esperar que haya un sesgo menor principalmente debido a la heterogeneidad de las fuentes. La segunda característica fundamental es el tamaño que tiene: alrededor de 45.000 millones de palabras en el caso del castellano, una escala totalmente distinta a lo habitual en lingüística de corpus. Por el momento, esta es la muestra más representativa del castellano culto de la que podamos disponer. Como limitación, se puede mencionar que se trata de un corpus de enigramas (en este caso, secuencias de hasta cinco palabras). Es natural que una ventana de contexto tan reducida resulte un problema para distintos tipos de análisis, pero la ventaja es que se registra la frecuencia de aparición de cada enigma en cada año, posibilitando así el seguimiento cronológico. Que el corpus contenga sólo texto editado por profesionales, por otra parte, puede ser también una limitación para quienes deseen estudiar el registro más informal de la lengua, que es uno de los más ricos en neologismos semánticos. También se puede considerar una limitación el que, a menos que exista el compromiso de mantener el corpus actualizado, este ya está envejeciendo y pronto dejará de servir para detectar neologismos reales, ya que sólo contiene texto editado hasta el año 2008. Finalmente, la última limitación que podemos añadir es que se trata de un corpus de formas, sin lematización ni etiquetado morfosintáctico. Por conveniencia práctica, los datos exploratorios que vamos a obtener están basados en el análisis de frecuencias de formas flexionadas tal como aparecen en el corpus, sin lematización ni extracción de nombres propios o expresiones polilexémicas. En consecuencia, se trata sólo de una estimación y no de un registro preciso de la cantidad de lexemas.

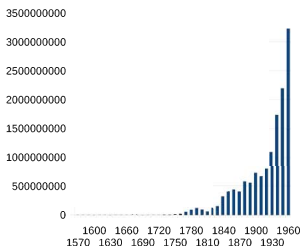
## 1.2. El crecimiento del vocabulario castellano

Hay personas que poseen una extraordinaria creatividad para inventar nuevas palabras y ponerlas en circulación. Una nueva palabra, para que sea buena, y para que sea adoptada en el vocabulario de los demás hablantes, tiene que ser interpretable cuando se escucha por primera vez. Necesita ser fácilmente asimilable en el plano fonológico, utilizando los mecanismos morfológicos de la lengua. Las palabras también surgen por composición o por resemantización, en el caso de los procesos metafóricos, metonímicos, etc. Estos procesos generan una corriente constante de neologismos y, sin embargo, no parece una consecuencia probable que se desate una especie de carrera inflacionista en la lengua. La lengua es un sistema autorregulado, y es capaz de encontrar un equilibrio—inestable— en el continuum de aparición de nuevas unidades léxicas. ¿Es posible medir de alguna forma este equilibrio? La cuestión no es sencilla porque para medir el vocabulario de una determinada época se debe adoptar una medida de referencia. El factor más claro sería la cantidad de texto producido, es decir, la relación entre el número de palabras distintas (vocabulario) y de palabras en total (extensión). Sin embargo, otras variables podrían ser de tipo extralingüístico, como la cantidad de hablantes o la posición que ocupa esa lengua en la cultura universal como difusora de nuevas ideas, realidades y denominaciones.

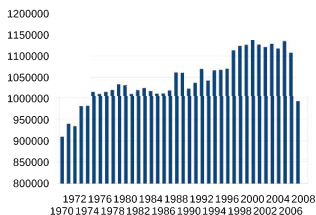
Una serie de datos exploratorios, antes de cualquier consideración teórica, puede darnos una primera idea de a qué nos atenemos cuando estudiamos la dinámica del vocabulario castellano. Es posible obtener una medición objetiva de dicha dinámica calculando su crecimiento, y esto nos permite saber si, más allá de la renovación del léxico, la cantidad de palabras en circulación va en aumento o si su número se mantiene relativamente estable. La figura 1 presenta la cantidad de palabras diferentes (eje Y) que hay en cada década desde 1500 hasta 1960 (eje X). No se trata del vocabulario acumulado a través de las épocas sino el vocabulario activo en cada década. Así, podemos ver que en la década de 1960 hay muchas más palabras vivas que antes. Lógicamente esto va en correlación con el tamaño del corpus, que también va creciendo (Figura 2).



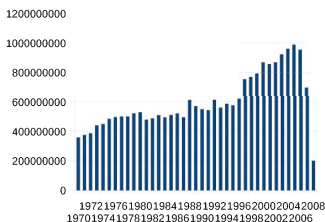
**Figura 1:** Vocabulario por año en Google Books entre 1500 y 1960



**Figura 2:** Extensión por año en Google Books 1500 y 1960



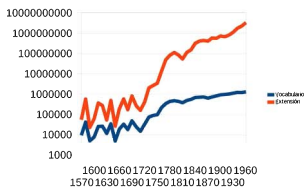
**Figura 3:** Vocabulario (Google Books 1970-2008)



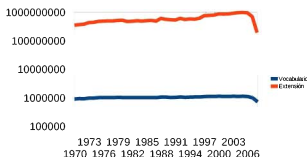
**Figura 4:** Extensión (Google Books 1970-2008)

Algo similar ocurre en el período 1970-2008, en la Figura 3 para el vocabulario y en la Figura 4 para la extensión, ahora por años y no por décadas, ya que el tamaño del corpus es en este período mucho mayor. En esta gráfica se ve que el vocabulario es bastante estable, en torno al millón de palabras. Hay un crecimiento, pero ese crecimiento también está en relación con el tamaño del corpus, y no es tan acusado. Se trata, por tanto, de una muestra equilibrada. Las gráficas de las Figuras 5 y 6, para el primer y segundo período respectivamente, expresan este paralelismo. El eje Y está en escala logarítmica para que podamos apreciar la relación porque de lo contrario las diferencias serían muy grandes. La Figura 7 expresa la correlación entre vocabulario y extensión en el período 1970-2008: mientras más grande es el corpus, mayor es la cantidad de palabras distintas. En escala logarítmica en los dos ejes (Figura 6) el crecimiento se ve como una línea recta, cumpliendo lo predicho por Heaps (1978). Podemos interpretar la pendiente de la línea recta como una medida de riqueza léxica, de

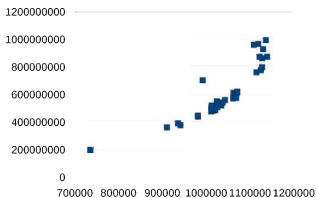
manera tal que menor pendiente de la recta significa mayor riqueza de vocabulario (con el tamaño del corpus aumentaría más rápidamente la cantidad de palabras distintas).



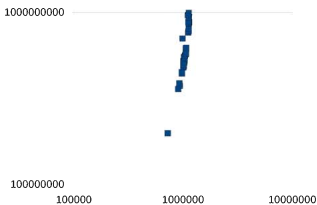
**Figura 5:** Vocabulario frente a extensión (Google Books 1500-1960)



**Figura 6:** Vocabulario frente a extensión (Google Books 1970-2006)



**Figura 7:** Correlación Extensión- vocabulario (Google Books 1970-2008)



**Figura 8:** Ídem anterior, ahora en escala logarítmica en los dos ejes

## 2. Antecedentes

### 2.1. La detección y clasificación de neologismos en general

En neología, los enfoques teóricos y empíricos se desarrollan en paralelo pero también se retroalimentan. La clave de la retroalimentación es que para un enfoque empírico o cuantitativo necesitamos una teoría de la neología que pueda ser expresada en términos computacionales (Janssen, 2009). Cada tipo de neología debe ser definido de una manera distinta, y de ello depende la estrategia de detección en cada caso. No es posible, por ejemplo, detectar neologismos semánticos con la misma técnica con la que se detectaría la neología formal. Por este motivo, toda estrategia de detección implica una forma de clasificación, por lo que esta, fundamentada en presupuestos teóricos, deberá ser siempre anterior a la detección empírica. Naturalmente, esto no significa que la clasificación no pueda modificarse a la vista de las muestras obtenidas.

En cuanto a los sistemas de clasificación de neologismos, además de las clasificaciones de cambio

semántico en general ofrecidas por Bloomfield (1933), Ullman (1957) y, más recientemente, Blank (1999), disponemos de distintos sistemas de clasificación específicos fundados en las matrices generadoras –conocidas también como mecanismos de formación– (Guilbert, 1975; Sablayrolles, 1997; Makri, 2009; Cabré, 2009). Cabré clasifica a los neologismos semánticos como reducción, ampliación o cambio de significado, y los ubica en la categoría más general de cambios gramaticales, junto con los casos de conversión de categoría gramatical (p. ej., un sustantivo se convierte en verbo) o bien de subcategorización (p. ej., un verbo transitivo se convierte en intransitivo). Otros miembros del mismo equipo han propuesto nuevas clasificaciones para introducir distinciones más granulares entre los neologismos semánticos, de manera que los neologismos se puedan clasificar, por ejemplo, según si se inspiran en ciertos tipos de nombres propios, como marcas comerciales (Feliu et al., 2009) o ciertos campos semánticos como los colores (Freixa et al., 2010).

En el caso de la neología formal, la detección es comparativamente más simple que la requerida para la neología semántica, sin embargo también entraña dificultades de orden teórico y práctico. En diversas herramientas de detección de neología, tales como *Sextan* (Vivaldi, 2000), *Buscanejo* (desarrollado por Laura López por encargo del IULA, ver Cabré, 2009), *NeoTrack* (Janssen, 2010), *Telanaute* (Issac et al. 2010), entre otros, la detección se realiza aplicando el criterio lexicográfico, que está fundado en una concepción del neologismo como una palabra que no aparece en los diccionarios. En la práctica, se utiliza un conjunto de diccionarios de referencia como corpus de exclusión y, de esta manera, toda palabra de un texto analizado que no esté en el corpus de exclusión es considerada como candidato a neologismo.

El criterio lexicográfico sobre el cual está fundada esta definición de neologismo no está libre de inconvenientes. Fundamentalmente, no hay garantías de que el corpus lexicográfico sea representativo del acervo léxico de una comunidad de habla y, en consecuencia, el hecho de que una palabra no esté en el diccionario no significa que sea un neologismo. El problema está en que no todas las palabras son diccionarizables. Las innumerables expresiones que utilizamos para designar entidades del mundo son palabras que no tienen lugar en el diccionario porque tienen referencia en lugar de significado. Y no se trata únicamente de expresiones referenciales en la forma de nombres propios o terminología técnica, otro tipo de expresiones, como *a capella*, *alien*, *aloe vera* o *barbie* no aparecen en los diccionarios generales de la lengua castellana y, sin embargo, su condición de neologismos es por lo menos dudosa, como veremos a continuación.

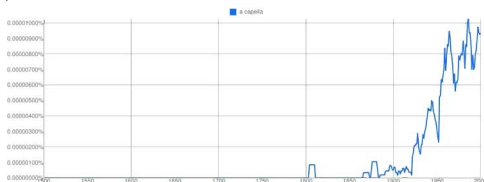


Figura 9: *a capella* (Google Books 1500-2008)

*A capella*, una locución italiana que, en su variante adjetiva, se aplica a las composiciones para canto

que se interpretan sin acompañamiento instrumental, aparece de forma frecuente en ejemplos como (2) durante al menos todo el siglo xx (Figura 9).

(2) “El Palacio de Liria contiene dos composiciones polifónicas, *a capella*, de Hotz” (GBooks, 1927).

El caso de *alien*, primo de nuestro *alienígena* y derivado del extraterrestre de la película *Alien: el octavo pasajero* y sus secuelas, es el de un uso que se puede haber extendido metafóricamente para aplicarse a cualquier tipo de extraterrestre o de entidad que, humorísticamente, toma sus rasgos (3).

(3) “...la serie de Tommy Lee Jones en la piel de ‘un *alien* que va a investigar a Japón’ ” (*El País*, 31/08/2011)

Ahora un cantante pop lo utilizó en el nombre de un nuevo disco (“Canciones para aliens” de Fito Páez) lo cual contribuirá a la difusión de este uso. La pregunta es en qué punto estos usos dejan de ser expresivos y pasan a convertirse en un ítem del vocabulario y, en cualquier caso, en qué fases de este proceso conviene llamar a esta unidad en formación un neologismo, puesto que cuando llega a su plena institucionalización deja de serlo. No se puede decir que el uso expresivo que se hace del vocabulario modifique el significado del vocabulario en el acto, ya que la analogía en el discurso es un cambio que se produce en el orden pragmático y no en el semántico. Un cambio de tipo pragmático no altera el significado de la unidad léxica sino que la ajusta a un determinado contexto. En el caso de un uso expresivo del lenguaje, se podría considerar una “explotación” de un uso normal, en términos de Hanks (2004). Sin embargo, con el uso y a través de un proceso histórico, algunas metáforas pasan a ser parte de la lengua, y lo hacen de manera casi inadvertida porque los hablantes olvidan que se trata de una metáfora.

En el caso de *aloe vera*, el nombre científico *Aloe vera* existe en la taxonomía científica de las plantas desde mucho antes de que se pusiera de moda esta especie carnosa con propiedades diversas. Este ejemplo es también una muestra de la complejidad de la noción teórica del neologismo. En este caso, se trataría de un neologismo por ser un término de especialidad que comienza a ser conocido por el resto de la población (nótese el cambio ortográfico). Los diccionarios contienen, sin embargo, muchos nombres de plantas. ¿Cuál es el límite, entonces, para que el nombre de una planta deje de ser un término especializado y pase a convertirse en vocabulario general?

El ejemplo de *barbie* es parecido al de *alien/Alien*. Nos habla también del paso de las expresiones referenciales a la lengua general, el mismo recorrido que en su momento hicieron otras como *lolita*, *donjuán* o *casanova*, que ya se han ganado su lugar en los diccionarios. ¿Son unidades del vocabulario o debemos considerarlas expresiones referenciales? *Lolita* y *Barbie* son nombres propios (respectivamente, de la protagonista de la novela homónima de V. Nabokov y de una muñeca) que representan estereotipos. A diferencia de *lolita*, es más probable encontrar el nombre de *barbie* en la prensa que en una novela. Las dos describen una frecuencia ascendente en Google Books (Figura 10). Debe tenerse en cuenta, además, que por la condición de marca comercial de *Barbie*, los diccionarios serán más reacios a incluirla, y si lo hacen deberán indicar en nota su origen de marca registrada. En el caso de *donjuán*, el proceso de lexicalización de la unidad también se ve en la evolución de su ortografía. En el mismo corpus (Figura 11), observamos que distintas grafías van compitiendo en el tiempo hasta que finalmente una se impone frente a la otra. Mientras decae el uso de *Don Juán* escrito separado, crece la forma *donjuán* escrito junto. En la misma zona de tránsito podríamos incluir también el caso de los alfonimos, que son las siglas que se convierten en palabras, tales como *oenegé* o *devedé*

(Martínez de Sousa, 2005).

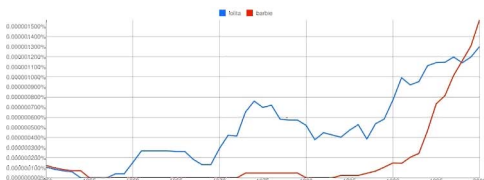


Figura 10: lolita frente a barbie (Google Books 1500-2008)

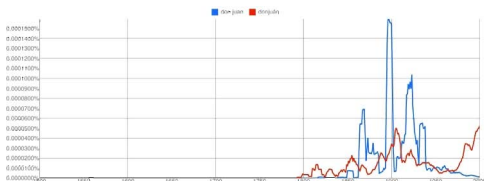


Figura 11: donjuán frente a don juan (Google Books 1500-2008)

**Comentario Amor Montané:** ¿En el corpus se distingue entre mayúsculas y minúsculas?

**Re:** El corpus original sí, pero la versión utilizada en el experimento que luego voy a explicar está filtrada bajando todas las letras a minúsculas y eliminando enagramas que, entre otras cosas, incorporan signos de puntuación, debido a que en este experimento no eran necesarios. Su eliminación facilitaba el procesamiento ahorrando memoria al reducir la cantidad de enagramas distintos.

**Comentario Teresa Cabré:** Las mayúsculas y los signos de puntuación te hubieran podido servir para identificar los nombre propios.

**Re:** Efectivamente, con un estudio concreto de los nombres propios este estudio habría sido más fino. Para ello hubiera sido necesario respetar la integridad del corpus original e implementar además una técnica de reconocimiento de entidades.

Para cerrar la discusión sobre los casos límite de palabras que se encuentran fuera del diccionario, tenemos que tratar el caso de las faltas de ortografía. Esto vale, por ejemplo, para la variante gráfica errónea de *a contrarreloj* (*a contrareloj*). Justamente la Fundeu lo recordó hace poco<sup>3</sup>. ¿Hasta qué

3 04/02/2011 – Vademécum contrarreloj/contra reloj. <http://www.fundeu.es/vademecum-C-contrarreloj-contra-reloj->

punto, entonces, se tiene que generalizar un uso incorrecto para que empecemos a considerar esta ocurrencia un neologismo? Lo mismo puede pasar con otras expresiones que son de un registro coloquial o que se utilizan en un contexto determinado como una metáfora. Resulta evidente que no puede incorporarse una nueva entrada al diccionario cada vez que un periodista ofrezca un uso nuevo del vocabulario. Esta situación refleja la tensión existente entre lengua y habla, o entre sistema y proceso. Como dije antes, en un lento proceso histórico, el habla altera el código de la lengua hasta que un uso que primero era original o era un ocasionalismo lentamente se va incorporando al vocabulario.

## 2.2. La detección neologismos con métodos cuantitativos

Cuando no se aplica el criterio lexicográfico, problemas como los que discutíamos antes o bien se plantean de otro modo, o bien no se plantean en absoluto. Un ejemplo de enfoque empírico no basado en diccionarios es el presentado por Baayen & Renouf (1996) en el campo de estudio de la productividad morfológica en la neología. Se trata de un estudio de productividad de una serie de afijos en inglés (-ness, -ity, un-, in- y -ly) en un corpus diacrónico conformado por los archivos del periódico *The Times* entre 1989 y 1993. El método está basado en el estudio de los *hapax legomena*, que en el contexto de un corpus diacrónico denotan cualquier palabra que aparece en un determinado período por primera vez con respecto a la parte precedente del corpus. Cabe imaginarlo como un barrido temporal del corpus en el que se registran todas las palabras que aparecen por primera vez. El recurso de los hapax resulta útil en este contexto porque aquí lo que se estudia es la cantidad de palabras diferentes (o nuevas) que incorporan un determinado afijo y, de esta manera, se puede saber qué afijo es más productivo.

Los autores son críticos con los estudios sobre productividad morfológica basados en el análisis de diccionarios, como el de Cannon (1988), quien registra los afijos más frecuentes en el leuario y ello le lleva a argumentar que el sufijo -ly está en decadencia en inglés, cuando del estudio basado en corpus de Baayen & Renouf (*op. cit.*) se desprenden conclusiones opuestas. Esto se puede apreciar en la Figura 12 (en la mitad derecha para el caso del sufijo -ly), donde el orden temporal se representa en el eje X y la cantidad de hapax en el eje Y. Lejos de decaer, se ve claramente que la cantidad de palabras nuevas con el sufijo -ly va en aumento. La razón de esta diferencia en las conclusiones de uno y otro estudio es que la selección del leuario de los diccionarios sigue pautas del criterio editorial que no tienen por qué corresponder fielmente a lo que ocurre en el lenguaje. En el caso del sufijo -ly, por ejemplo, el lexicógrafo posiblemente haya considerado innecesario incluir en el diccionario palabras con ese morfema por creer que se trata de derivaciones morfológicas transparentes en significado.

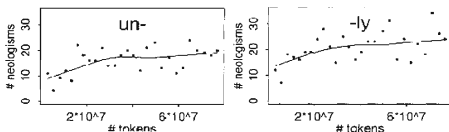
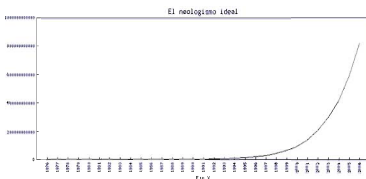


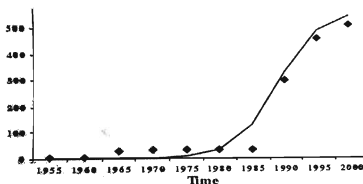
Figura 12: Estudios de productividad morfológica (Baayen & Renouf, 1996)

También en una línea de investigación cuantitativa, que incorpora tanto el estudio de la neología como del desuso y fenómenos que forman parte de lo que podríamos llamar el ciclo de vida de las palabras, se encuentran algunos trabajos previos (Nazar & Vidal, 2010; Nazar & Azarian, 2010; Nazar, 2011; Cabré & Nazar, en preparación). En ellos propusimos una estrategia de detección consistente en la postulación de una especie de neologismo ideal para representar el comportamiento típico de una unidad neológica, que muestra un aumento exponencial en su frecuencia de uso (Figura 13). La extracción de los candidatos a neologismo a partir de un corpus diacrónico de archivos de prensa se producía por medio de un cálculo de similitud geométrica (distancia euclidiana) entre la curva de distribución de frecuencias de cada candidato y la de este neologismo ideal.



**Figura 13:** Extracción de neologismos por medio de la similitud con un neologismo “ideal” (Nazar & Vidal, 2008; Nazar & Azarian, 2010; Nazar, 2011)

Hay otros autores trabajando en paralelo con un enfoque parecido. Un ejemplo es el trabajo de Schierz (2007) en el campo de la terminología especializada, en el que describe la evolución de la frecuencia de uso de los términos en corpus diacrónicos conformados por textos sobre inteligencia artificial. Compara un corpus de textos académicos con otro de patentes y esto le permite hacer predicciones sobre cuáles van a ser los temas sobre los que se crearán más patentes en el futuro. Ella sostiene que, en realidad, la curva de los neologismos –o más precisamente, en su contexto, de los términos que se ponen de moda en la disciplina– es una función sigmoide, es decir, que se parece más bien a una curva que describe una S (Figura 14). Esta curva, frecuente en diversos ámbitos como el ritmo de las innovaciones tecnológicas en la historia o el crecimiento de poblaciones en biología, se conoce también con el nombre de función de Gompertz, en honor a Benjamin Gompertz, matemático francés que la describió en el siglo XIX.



**Figura 14:** la neología como función sigmoide (Schierz, 2007)

### 2.3. Antecedentes del estudio de la neología semántica

La neología semántica es tanto o más difícil de definir que la neología formal. La definición utilizada al principio, según la cual los neologismos semánticos son palabras formalmente conocidas que adquieren un significado nuevo (Guilbert, 1975), entraña la dificultad de precisar el significado de la expresión “palabras conocidas”. La expresión quiere decir “conocidas por la comunidad de habla” y no por la competencia individual del sujeto hablante, que puede desconocer parte del vocabulario debido a condicionamientos etéreos, sociales, culturales o bien diatópicos. Pero en la neología semántica confluyen diversos fenómenos que tienen distintas causas. En general, podemos hablar de neología semántica como de un proceso de polisemización, pero este proceso a su vez puede tener distintos orígenes. Puede ser resultado de un proceso de especialización debido a usos técnicos, como en el caso de textos de informática en donde, por un préstamo semántico del inglés, se utiliza el verbo soportar en un sentido que los diccionarios del castellano no registran (4).

(4) ...en las versiones no *soportadas* se dejarán de publicar nuevos paquetes...

Además de los tecnicismos o de la invasión de otras lenguas, podemos observar otras causas como la variación dialectal. Si nos centramos en el estudio del castellano, encontramos que en las distintas variantes geográficas existe un proceso de neología semántica singular. Palabras que primero circulan como expresiones coloquiales, lentamente van ganando terreno en el vocabulario de la lengua por medio del uso y de los medios de comunicación hasta que finalmente acaban por transfigurar el uso normal del lexema desde la perspectiva del hablante medio. En el caso de la variante argentina de la lengua, algunos ejemplos de palabras que están haciendo este recorrido son *colgado*, *tildado*, *nave*, *máquina*, *tubo*, *bobo*, *bobazo*, *trucho*, *producirse*, *camión*, *cuento*, *manija*, *corralito*, *arrugar*, *corrida*, *arbolito*. Lo mismo ocurre en otros círculos como en el argot marginal. Por ejemplo, palabras como *limado*, *chapita*, *canuto*, *copado*, *manso*, *forro*, *guacho*, *zapato*, etc., posiblemente provengan del argot carcelario argentino, pero comienzan a ampliar su ámbito de circulación y ser utilizadas por jóvenes de distintas clases sociales.

Desde el punto de vista teórico, la primera referencia sobre la neología semántica es seguramente el trabajo de Bréal (1904), en donde esta se entiende como un proceso de polisemización que puede tener origen, por ejemplo, en truncamientos. Cuando una expresión sintagmática empieza a utilizarse muy a menudo se contrae y esto puede producir cambios de categoría gramatical por los cuales, por ejemplo, un adjetivo pasa a ser un nombre, como en *droite perpendiculaire* → *une perpendiculaire*, y por lo tanto esa forma pasa a ganar un sentido, es decir, a hacerse (más) polisémica. Las tipologías estudiadas por Bréal son diversas: otro tipo de cambio de significado es la diferenciación, por la cual si en una lengua coexisten dos unidades con el mismo significado, estas tienden a repartirse distintos rasgos del significado que antes compartían, adquiriendo cada una connotaciones diferentes. Según uno de los ejemplos que menciona el autor, cuando el francés invadió uno de los *patois* de Suiza, la palabra francesa *jardin*, que tenía el mismo significado de la variante suiza *courtill*, acabó por desplazar esta denominación, que fue adquiriendo una carga despectiva debido a que el francés se consideraba una lengua de prestigio.

Los autores que trabajan en neología semántica y en el campo más general del estudio del cambio semántico consideran que el trabajo de Bréal no ha perdido vigencia y, de hecho, las categorías que

manejan están basadas en él (Bloomfield, 1933; Blank, 1999; Traugott & Dasher, 2002; Cabré, 2009; Sánchez Manzanares, 2009; Steffens & Baiwir, 2009). Se puede decir que las categorías en las que debe entrar un fenómeno de cambio de significado son las expuestas en la Tabla 1, aunque existen diferencias de matices y discusiones sobre cómo subdividirlas: por ejemplo, se podría distinguir entre las distintas figuras retóricas (como la sinécdoque, hipérbole, lítote, etc.), así como emplear los procesos de gramaticalización que se estudian en lingüística cognitiva.

| Proceso         | Definición                                                                                                                                                                                                                                                                                                                                                                                     | Ejemplo                                                                                                                                                                                                                                                                          |
|-----------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Generalización  | Se produce cuando el nombre de un elemento de un conjunto pasa a designar el conjunto entero.                                                                                                                                                                                                                                                                                                  | En inglés antiguo, <i>dogge</i> era una raza particular de perro. El actual <i>dog</i> , sin embargo, terminó designando la clase entera.                                                                                                                                        |
| Especialización | El nombre de una categoría pasa a designar un elemento en particular.                                                                                                                                                                                                                                                                                                                          | La palabra <i>deor</i> en inglés antiguo designaba una amplia categoría de animales, como en alemán <i>Tier</i> , que se emplea para los animales en general, pero en el inglés actual, <i>deer</i> pasa a designar una especie en particular, el ciervo.                        |
| Metáfora        | Una palabra toma los rasgos semánticos de otra a través de una comparación tácita. Ocupa un vacío designativo en la lengua de forma inmediata, al utilizar, por analogía o transposición, las características de un dominio distinto.                                                                                                                                                          | La palabra <i>montaña</i> designa prototípicamente el accidente geográfico, pero por derivación metafórica ha pasado a significar otros elementos, como una gran acumulación de cosas que recuerdan por su forma a este accidente ("una montaña de basura, de papeles"...).      |
| Metonimia       | Los significados son contiguos en el espacio o en el tiempo.                                                                                                                                                                                                                                                                                                                                   | Actualmente, en inglés un <i>paper</i> es un artículo (científico), pero deriva del material o soporte tradicional, el papel.                                                                                                                                                    |
| Empeoramiento   | La palabra adquiere connotaciones negativas.                                                                                                                                                                                                                                                                                                                                                   | En inglés antiguo, <i>awful</i> significaba <i>full of awe</i> , es decir, algo que inspira admiración, fascinación e incluso miedo o reverencia; en cambio, en el inglés moderno ha pasado a significar <i>horrible</i> .                                                       |
| Mejora          | La palabra adquiere connotaciones positivas.                                                                                                                                                                                                                                                                                                                                                   | Con un origen similar a <i>awful</i> , la palabra <i>awesome</i> en cambio, en inglés actual significa algo así como <i>buenísimo</i> .                                                                                                                                          |
| Eufemismo       | Una palabra que se refiere a un concepto tabú es sustituida por otra (el eufemismo) que esconde o disimula dicho concepto considerado negativo. Con el paso del tiempo, las palabras que refieren a temas tabú pueden adquirir una carga semántica negativa, de manera tal que lo que inicialmente era un eufemismo acaba por convertirse en impronunciable y se tiene que volver a sustituir. | En inglés antiguo, <i>stincan</i> no significaba 'oler mal', como ahora lo es <i>to stink</i> , sino simplemente 'oler'. Sin embargo, el ser usado como eufemismo le supuso una carga negativa, y así aparecen <i>smell</i> y luego <i>odor</i> para pasar por el mismo proceso. |

Tabla 1: Principales tipos de cambio de significado

En el ámbito específico de la neología semántica, destaca el trabajo de Bastuji (1974) en su análisis de las relaciones entre cambios semánticos y cambios observables en el plano sintáctico mediante el uso de rasgos sintáctico-semánticos (por ej., ± concreto). En el caso de una expresión como *acceptiser le climat social*, lo que se produce es un cambio del verbo *acceptiser* donde un argumento que se supone debe ser algo concreto como una herida, para a ser algo abstracto como la economía o el clima social. La misma autora admite, sin embargo, que muchos casos de neología semántica no comportan cambios

en los rasgos semánticos sino que se fundan en las presuposiciones y en el conocimiento del mundo que tienen los hablantes. La idea de buscar evidencias de cambio semántico en los cambios sintácticos, sin embargo, resulta atractiva, ya que si bien no fructificará en una estrategia de detección exhaustiva, sí puede servir sin embargo para obtener casos con cierta seguridad, ya que los cambios sintácticos que acusan cambio semántico pueden ser muchos y de diversos tipos. Para mencionar otro ejemplo, esta vez en el caso de los adjetivos, un indicio puede ser la aparición de cuantificadores. Es el caso de *limado*, uno de los ejemplos que mencioné antes. Originalmente, el sustantivo *limado* es un término técnico proveniente de la mecánica y la industria, derivado participial del verbo *limar*, y designa la acción o el efecto de limar, mientras que, en su nuevo uso coloquial, refiere a una persona en la que se observa daño cognitivo crónico o irreversible por efecto de la toxicomanía. En el uso normal no admite cuantificadores. En el uso nuevo, en cambio, aparece un paradigma de cuantificación, ya que una persona puede estar {*muy, medio, más o menos*} *limada*. Otros trabajos que van en la línea de buscar evidencia del cambio semántico en los cambios de tipo sintáctico, son los de Lorente (2011) y, más específicamente en el estudio de cambio semántico por conversión, el de Cabré et al. (2011).

Cerrando esta parte, añadiría que la reflexión sobre la neología semántica existe también en el estudio de la terminología especializada (como estudio de la neonomía). Un ejemplo en esa línea son los trabajos compilados por Groult et al. (1988) sobre el estudio de los procesos migración de vocabulario entre distintas ciencias. Otro es el de Meyer & Mackintosh (2000) acerca de los procesos de fluctuación de significado por términos que se originan en la lengua general y adquieren un uso nuevo en la lengua especializada o viceversa, es decir, casos de des- y re-terminologización. La neología semántica como recurso dentro de la terminología especializada no era alentada inicialmente por los terminólogos, pero aparentemente esto está empezando a cambiar, ya que una denominación basada en una metáfora, si es acertada, puede ser útil desde el punto de vista cognitivo (Estopá, 2009).

## 2.4. La detección de neologismos semánticos con métodos cuantitativos

La dificultad de la detección de neología semántica radica en que para ella no se puede aplicar el criterio lexicográfico tal como se ha venido haciendo en la neología formal. Y tampoco podemos utilizar filtros de frecuencia, ya que, si bien parece razonable sospechar que el aumento de frecuencia de una unidad puede estar asociado a un nuevo significado, en realidad esto no tiene por qué ser así. El hecho de que la palabra aumente o disminuya su frecuencia de uso puede tener una diversidad de causas, y además puede haber cambio semántico sin aumento de frecuencia. Un ejemplo de ello se muestra en la Figura 15 para el caso de *descargar* en el período 1950-2000, donde curiosamente decae su frecuencia en lugar de aumentar, contrariamente a lo que cabía esperar tras la aparición del neologismo semántico *descargar* como 'grabar un archivo en una computadora proporcionado por un servidor web'. En la Figura 16, para el caso de *navegador* (azul) frente a *navegante* (rojo), vemos que *navegador* tiene un aumento de la frecuencia debido al uso más reciente de 'programa que sirve para navegar por la red'. Sin embargo, en términos relativos este uso no es superior al que ya se conocía en el siglo XIX. Es interesante preguntarse por qué eligieron *navegador* como traducción de *browser*. El origen seguramente está en el popular en su época Netscape Navigator, y el motivo por el que no se emplea la forma *navegante* es que está ya muy utilizada y connotada como la 'persona que navega', mientras que *navegador* es menos frecuente y, así, está más disponible para ser utilizada como el instrumento, además de porque el sufijo *-dor* es muy productivo para la creación de sustantivos que denotan máquinas, aparatos o dispositivos.

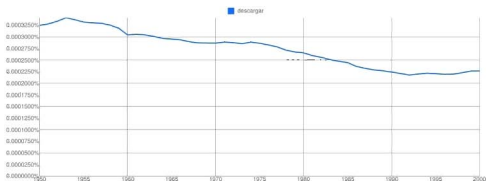


Figura 15: descargar (Google Books 1950-2000)

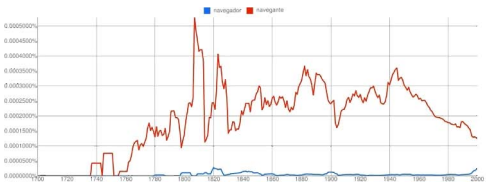


Figura 16: navegador frente a navegante (Google Books 1700-2000)

Se han presentado propuestas para la detección de la neología semántica que son empíricas pero no esencialmente cuantitativas. En esta línea podemos mencionar el trabajo de Mejri (2006; 2010), que estudia la neología semántica a través de las relaciones entre predicados y argumentos siguiendo la teoría de las clases de objetos de Gross (1994). Gross presenta una teoría lexicográfica en la que se incorpora la taxonomía al estudio de las relaciones entre predicados y argumentos. Recuerda en parte a la teoría del Corpus Pattern Analysis de Hanks (2004) ya que para una misma palabra se crean diferentes entradas como patrones típicos de ocurrencia. Por ejemplo, en el caso de un verbo como *volcar*, se observa que en uno de sus patrones típicos de ocurrencia tiene como argumento la clase de objetos de los vehículos ya que, normalmente, lo que puede volcar es una camioneta, un camión, un automóvil, etc. Por lo tanto, aquí ya no se está hablando de rasgos del tipo  $\pm$ concreto sino que se afina mucho más el análisis. Aplicando la teoría de las clases de objetos al estudio de la neología semántica, nos podemos preguntar, por ejemplo, qué es la cosa típica que se puede devorar. Un corpus normal o corpus de referencia nos dirá que lo que se devora es comida y no libros, como en (3). La idea de Mejri es que si en el corpus analizado comienza a aparecer un argumento que no es propio de la clase de objeto esperada, entonces allí hay una indicación de cambio semántico.

(3) *J'ai dévoré livre après livre*

En los últimos años se ha producido una corriente de análisis de una naturaleza distinta, fundada en una perspectiva distribucional, que, al menos en la literatura revisada hasta ahora, está basado en el estudio de la coocurrencia de palabras más que en el análisis semántico profundo, como el análisis de dependencias entre predicados y argumentos. Los dos puntos de vista no son en absoluto contrapuestos,

lo que quiere decir que todavía queda mucho por hacer aplicando el enfoque estadístico al análisis profundo de la semántica. Muy probablemente vaya a ser así como evolucionará este campo, aunque por el momento la mayoría de los autores revisados se limita a registrar coocurrencia de palabras.

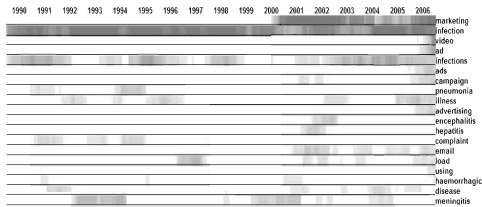


Figura 17: Cambios en la formación de colocaciones de *viral* (Renouf, 2008)

Posiblemente la primera en hacer estudios de neología semántica por medio del registro de las coocurrencias, Renouf (2010) presenta una metodología basada en el estudio de los cambios en las colocaciones. Lo que se ve en la Figura 17 son los colocativos de la palabra inglesa *viral*. La intensidad del color representa la frecuencia, y de esta manera advertimos que se produce un nuevo sentido con la entrada de *viral marketing*, en el año 2000, o de *viral video*, en contraste con otros colocativos más estables como *viral infection*.

Del mismo equipo que Renouf, Collier (1998) describe un enfoque similar en el estudio del cambio en las palabras tópicas, utilizando para ello los archivos del *Financial Times*. El interés del autor no es lexicográfico ni lingüístico. Su objetivo es desarrollar un sistema que permita actualizar automáticamente el glosario del periódico, que contiene nombres de cargos públicos e información general. La metodología consiste en analizar los colocativos de una determinada unidad (definiendo colocativos como unidades que ocurren a izquierda y derecha de la unidad analizada con una frecuencia estadísticamente significativa). De esa forma, compara distintos períodos de tiempo y registra cambios en la distribución de colocativos en cualquiera de las siguientes cuatro categorías: 1) son nuevos, 2) aumentan su posición en la lista de colocativos frecuentes, 3) pierden posiciones, 4) desaparecen. De esta forma, el sistema de Collier registra los cambios en los colocativos de, por ejemplo, la palabra *cult* ('culto'), que es una de esas voces que se van actualizando con el tiempo porque se utiliza para hacer referencia a distintos cultos cuando estos son relevantes para los medios de comunicación.

Sagi et al. (2009) aplicaron las estadísticas de coocurrencia al estudio de dos tipos específicos de cambio semántico, la generalización y la especialización, en un corpus de inglés de los siglos *xii* al *xviii*. Lo que estos autores hacen es calcular un índice que llaman *densidad* y que definen como la probabilidad de una palabra de aparecer en contextos similares. Los autores explican que pueden medir la pérdida de densidad de palabras como *dog*, por ejemplo, a medida que se va haciendo más general (desde designar una raza de perro a la especie completa). Cook & Stevenson (2010) adoptan una línea parecida, concentrándose en la aplicación de estadísticas al estudio del cambio en el plano axiológico de la unidad léxica, que corresponde a otras dos de las categorías mencionadas en la sección 2.3, el empeoramiento (*pejoration*) y la mejora (*amelioration*). El método consiste en organizar dos grupos de

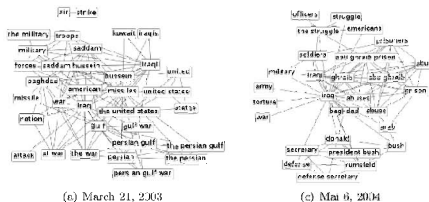
palabras “semilla”, que tienen valores negativos y positivos, es decir, por un lado palabras como *good*, *nice*, *excellent*, etc., y por el otro *bad*, *nasty*, *poor*, etc. De esta forma, una palabra analizada recibirá una ponderación determinada en función de la frecuencia con que aparezca en los mismos contextos con palabras de cada uno de estos dos polos. Como estos contextos provienen de un corpus diacrónico (desde el siglo xvii hasta la actualidad), esta ponderación va variando con el tiempo.

Otros estudios que se basan en la visualización de procedimientos de clustering son los de Rohrdantz et al. (2011) y Boussidan & Ploux (2011). Junto con las anteriores, son propuestas que tienen muchos elementos en común, entre ellos la aplicación de técnicas de clustering de los colocativos de una unidad analizada, para posteriormente medir la distribución de frecuencias de los miembros de cada cluster en el tiempo. De esa manera, tal como podemos ver en la Figura 19, tomada de Rohrdantz et al. (2011), en el corpus del *New York Times* el verbo *to browse*, hasta 1995, podía admitir argumentos como *bookstore*, *book*, etc., pero hay otro grupo de palabras que empieza a aparecer a partir de ese año: *web*, *internet*, *site*, etc., y esto nos permite advertir que esa palabra está mutando su sentido.



**Figura 18:** Visualización de procedimientos de clustering para detectar cambios semánticos (Rohrdantz et al., 2011)

En la línea del estudio de las palabras tópicas y también basándose en estadísticas de coocurrencia, en el trabajo de Holz & Tersniak (2010) se utilizan grafos de coocurrencia léxica aplicados a un corpus diacrónico igualmente de archivos de prensa (de nuevo con el *New York Times*). Estos autores se dedican al análisis de palabras tópicas pero la idea que proponen se podría aplicar también a otras unidades del vocabulario. Por ejemplo, en el caso de *Iraq*, toman distintas muestras de su uso en el periódico entre 2003 y 2004, y demuestran que en el año 2004 se produce una alternación del grafo de coocurrencia de *Iraq* en el que comienzan a aparecer nuevos nodos o temas. Concretamente en este caso, el nombre de *Abu Grahیب*, que apareció con frecuencia en la prensa en esa época debido a una filtración sobre cómo las tropas norteamericanas torturaban a los prisioneros de guerra iraquíes. Estos hechos alteraron el significado o las connotaciones de *Iraq*, tal como queda reflejado en el grafo de coocurrencia de la palabra en ese momento particular. Cambios de esta naturaleza determinan lo que los autores llaman la *volatilidad* de una unidad.



En un trabajo previo (Nazar & Vidal, 2010), proponíamos la utilización de técnicas de clustering para el estudio de la neología semántica. El clustering estaba basado en grafos de coocurrencia, de la misma manera que en mi tesis doctoral (Nazar, 2010), donde trataba de manera general la aplicación de grafos a estudios de semántica distribucional. Una de las aplicaciones de estos grafos sería la inducción de significado, que consiste en agrupar contextos de aparición de una determinada unidad léxica de acuerdo con un criterio de similitud en el vocabulario de los contextos. Para una expresión polisémica, se agrupan sus contextos de aparición reflejando cada uno de los distintos sentidos. El estudio de la neología semántica parece, entonces, un ámbito natural para la aplicación de esta técnica, ya que esto nos permitiría determinar y aislar un uso neológico de una expresión en una muestra de contextos de aparición ordenados cronológicamente. En la figura 20, por ejemplo, analizamos la unidad “palabra de honor” en un corpus de prensa (el diario *El País*). Se trata de una expresión que tiene dos usos: un uso recto (‘palabra que da una persona confianza lo que dice es verdadero’) y otro figurado, aplicado al escote de prenda femenina que no tiene tiras.

**Figura 20:** Inducción de sentidos por medio de clustering (Nazar & Vidal, 2008; Nazar, 2010)

mientras que el otro, en cambio, recién comienza a aparecer en 1998. Esta técnica, si bien resulta interesante, no se podría aplicar sin embargo al corpus de Google porque no habría suficiente contexto. Por otro lado, también hay que decir que el ejemplo de *palabra de honor* no es ideal en realidad porque el uso como tipo de escote no es neológico. Únicamente se puede decir que es un uso nuevo en el corpus analizado.

**Comentario Mercè Lorente:** Lo que ha cambiado es la prensa. La unidad no es neológica, sino que una expresión como *palabra de honor* 'escote' ha empezado a incluirse en la prensa cuando se ha comenzado a incorporar la moda, que antes era exclusiva de las revistas.

**Comentario Lluís de Yzaguirre:** Yo quería hacer un comentario similar. Cuando representabas la curva del vocabulario, has obviado algo importante que es que en los libros del <sup>xvi</sup> no se hablaba de ciertos temas, porque en los libros el registro era el de la lengua culta. ¿Cuántas palabras había en el siglo <sup>xvi</sup> que no estaban en los libros? Me parece que en esa visión se ha simplificado demasiado. Incluso cabe esperar que ahora la inclusión de libros exclusivamente digitales afecte también la curva del vocabulario, porque es probable que en ese género se hable de ciertos temas de los que no se hablaría en los libros de papel.

**Re:** Es cierto, pero este es el material del que disponemos por el momento. No queda otra alternativa que asumirlo como una limitación de este análisis.

Se puede advertir que todos los trabajos sobre neología semántica desde el enfoque cuantitativo que se han producido recientemente son esencialmente similares. De momento, una de las limitaciones más notables es que prácticamente no existen evaluaciones a gran escala de las distintas metodologías que se proponen, ya que se suele analizar un conjunto limitado de unidades elegidas arbitrariamente. Sería deseable dar un paso hacia el desarrollo de sistemas de clasificación de las palabras según su probabilidad de ser neológicas. La visualización es útil, pero más esfuerzo se ahorra si es el propio sistema el que decide cuáles son las unidades léxicas que vale la pena analizar. Son pocos los trabajos en los que se puede decir que este sistema de clasificación realmente existe. Se han presentado dos que, además, utilizan también el corpus de Google Books Ngrams (Renau & Nazar, 2011; Gulordava & Baroni, 2011). En Renau & Nazar (2011) analizamos la evolución de los verbos pronominales estrictos y constatamos que, con el tiempo, los verbos entran y salen de esta categoría. Antes era frecuente decir, por ejemplo, que uno puede "jactar valentía", sin embargo ahora lo normal es que uno "se jacte de" algo. Si este cambio puede ser considerado semántico o meramente sintáctico dependerá, sin embargo, de cada verbo en particular. El trabajo de Gulordava & Baroni (2011), en cambio, es más parecido a la propuesta de este seminario. Estos autores trabajan únicamente con los bigramas del corpus de Google Books Ngrams y proceden llevando a cabo una comparación de vectores que están compuestos por los colocativos de una misma palabra por un lado en la década del sesenta y por otro lado en la década del noventa, entendiendo por colocativos todas aquellas unidades léxicas que comparten bigramas con la unidad analizada. Utilizando una medida de similitud, pueden advertir si una palabra ha experimentado un cambio semántico cuando la comparación acusa que los vectores son muy disímiles. Según el sistema de evaluación que proponen, sabemos que su metodología de detección es buena porque demuestra una correlación de 0.38 con el juicio humano, que se puede considerar alta ya que los humanos entre ellos sólo tienen una correlación del 0.51.

Cook & Hirst (2011) son críticos con el trabajo de Gulordava & Baroni, particularmente con su sistema de evaluación. Efectivamente, ese sistema hace difícil la comparación de los resultados con los

obtenidos por otros métodos. El modelo que presentan Cook & Hirst es también esencialmente similar, ya que comparan las coocurrencias de una unidad analizada en dos subcorpus que representan distintos períodos de tiempo de un mismo corpus. La diferencia es que se interesan por fenómenos de baja frecuencia, y trabajan por tanto en un corpus más pequeño (50 millones de palabras del *New York Times*). Además, utilizan una metodología más compleja, que incluye análisis morfosintáctico y análisis de dependencias en lugar de tratar únicamente con estadísticas de coocurrencia. Las figuras de precisión y cobertura de las estrategias de detección de neologismos semánticos que presentan son muy altas, de entre 48% y 97%, en función de la frecuencia del nuevo sentido ya que, como es natural, aquellos sentidos que reúnen pocos contextos de uso son más difíciles de detectar. Lo que se le puede criticar a este método de evaluación es que utiliza ejemplos artificiales, una práctica que es común en ámbitos computacionales relacionados con la desambiguación y la extracción de sinónimos. Concretamente, dada una palabra polisémica y distintos conjuntos de aparición en los distintos sentidos de esa palabra, lo que Cook & Hirst hacen es simular un proceso de neología semántica separando los contextos de uno de los sentidos como si se tratara de un corpus nuevo. De esta manera, el ejercicio consiste en comprobar si el algoritmo es capaz de detectar este nuevo sentido. Es indudable que la evaluación de metodologías en el ámbito de la neología semántica es muy difícil y, al ser los enfoques cuantitativos relativamente recientes, es natural que aún no exista un consenso. Sin embargo, no parece haber una razón para no trabajar con casos reales en todo momento.

La presente propuesta de detección sigue esencialmente la misma línea de los trabajos comentados, ya que se basa también en el estudio de las relaciones sintagmáticas entre palabras por medio de las frecuencias de coocurrencia. En contraste, el presente trabajo destaca por estar especialmente adaptado a las condiciones que impone el corpus analizado, con una ventana de contexto reducida, por su sistema de evaluación objetiva y por su simplicidad conceptual y computacional.

### 3. Hipótesis y metodología

La hipótesis que formulamos es que el cambio semántico de una unidad analizada queda evidenciado en el cambio de las palabras que suelen coocurrir con ella, lo cual no es sino una consecuencia lógica de la hipótesis distribucional de Harris (1954). El objetivo es entonces detectar el cambio semántico a través de este cambio en la combinatoria de una unidad analizada. La metodología del experimento preliminar, con el que se pretende apoyar esta hipótesis, consiste en el estudio de las palabras que coocurren con cada uno de una lista de verbos del dominio de la informática (Tabla 1), que es un campo rico en la producción de neologismos semánticos. La idea es tomar cada verbo y otorgarle una ponderación de neologicidad, es decir, indicar qué tan probable es que haya incorporado un nuevo sentido por medio del cálculo de la cantidad de palabras nuevas que aparecen en los contextos de la unidad analizada. En el experimento descrito aquí se analizan sólo verbos por una decisión arbitraria, sin embargo, no hay ninguna razón para no intentar repetir el experimento con otras categorías gramaticales e incluso con otras lenguas. Otra decisión arbitraria ha sido analizar sólo una parte del corpus. Puesto que el año 2008 es el más reciente, el estudio de la evolución de los verbos se limita al período comprendido entre los años 1975 y 2008.

Entrando un poco más en detalle, para poder llevar a cabo este estudio fue necesario extraer los enigramas en los que la palabra analizada ocurre en el período comprendido entre 1975 y 1995 por un lado, y de 1996 hasta 2008 por otro. Para ello, y debido a que el corpus no está lematizado, primero se buscaron todas las formas de cada verbo en una base de datos léxica (OSLIN). Se evitaron todas

aquellas formas que podían resultar ambiguas, es decir, unidades que fueran idénticas a la de otra categoría gramatical (en el caso de *activar*, por ejemplo, *activa*, *activo*, *activas* quedarían excluidos porque se podría confundir con sustantivos o adjetivos). Por cada colocativo de cada verbo, comparamos la frecuencia que tenía en el primer período en relación con el segundo. Por cada verbo analizado, buscamos los enagramas (de 2 a 5) en los que el verbo aparece y registramos la frecuencia de aparición de todas las formas que comparten enagramas con la expresión analizada, con la excepción de palabras funcionales que se eliminan por medio de una lista de exclusión. Una vez contabilizadas las frecuencias de los coocurrentes en los dos períodos, calculamos un coeficiente de neologicidad tal como se muestra en la Ecuación 1, donde  $j_{i,1}$  sería la frecuencia del colocativo  $i$  del verbo  $j$  en el período 1 correspondiente a 1975-1985, que separamos del período 2 (1986-1995) sólo para visualizar resultados.

$$\text{Neo}(j) = \sum_{i=1}^{|J|} \frac{j_{i,3}}{j_{i,1} + j_{i,2} + 1} \quad (1)$$

En esencia, se trata simplemente de una división, y el resultado es lo que va a dar la posición de ese verbo en la clasificación final. Como veremos luego en los ejemplos, los elementos que más a menudo acompañan a los verbos en los enagramas son los sustantivos que funcionan como argumentos de estos verbos, y es el cambio de argumento el que tomamos como medida del cambio de significado.

|            |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    |
|------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| actualizar | DUEAE: “2 Introducir o aportar los datos más actuales o recientes: actualizar una base de datos; actualizar los contenidos de una enciclopedia; la prensa debe ofrecer una información actualizada y precisa”                                                                                                                                                                                                                                                                                                                                                                                                                                      |
| almacenar  | DRAE: “3. tr. Inform. Registrar información en la memoria de un ordenador”                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         |
| arrastrar  | Wikcionario: “2: Desplazar el ratón de un computador para mover objetos de la pantalla”                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            |
| descargar  | DRAE, el Clave, el DUEAE y el DEM (en este último: “9. tr. Inform. Transferir información desde un sistema electrónico a otro”                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |
| copiar     | sin documentar                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |
| ejecutar   | DRAE: “6. tr. Inform. Realizar las operaciones especificadas por un programa de un ordenador”<br>DUEAE: “6 inform Hacer que una instrucción o un programa sea realizado por la computadora: ejecutar un programa para enviar un correo”                                                                                                                                                                                                                                                                                                                                                                                                            |
| enlazar    | sin documentar                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |
| navegar    | DRAE: “5. intr. Inform. Desplazarse a través de una red o de un sistema informático”<br>Clave: “3 Referido esp. a una red informática, utilizarla o desplazarse a través de ella: Mi tía es usuaria de una red informática internacional y se pasa horas navegando por ella”<br>DUEAE: “4 Desplazarse de una página o documento a otro en una red informática, a través de ciertos vínculos entre ellos: navegando por internet puedes encontrar información interesante sobre todo tipo de temas”<br>DUE: “3 Inform. Desplazarse a través de una red mediante un navegador”<br>Wikcionario: “2: Explorar la red internet utilizando un navegador” |
| pegar      | sin documentar                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |
| reiniciar  | DRAE: “2. tr. Inform. Cargar de nuevo el sistema operativo en una computadora”<br>Clave: “reiniciar v. En informática, volver a iniciar: Reinicia el ordenador para que arranque correctamente”                                                                                                                                                                                                                                                                                                                                                                                                                                                    |
| soportar   | Clave: “3 Referido a un sistema informático, reunir las características necesarias para su                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         |

|          |                                                                  |
|----------|------------------------------------------------------------------|
|          | funcionamiento: Este ordenador soporta cualquier entorno de red" |
| vincular | sin documentar                                                   |
| visitar  | sin documentar                                                   |

**Tabla 2:** Muestra arbitraria de verbos en proceso de neología semántica relacionados con el mundo de la informática

Como se puede apreciar en la Tabla 1, algunos de los verbos ya tienen entradas en el sentido informático en algunos de los diccionarios de la lengua castellana consultados, tales como el *Diccionario de la lengua española*, de la Real Academia Española (DRAE), el *Diccionario de uso del español de América y España*, dir<sup>a</sup> P. Battaner (DUEAE), el *Diccionario Clave*, dir<sup>a</sup> C. Maldonado, El *Diccionario del español actual*, dir. M. Seco (DEA), el *Diccionari de uso del español*, de M. Moliner (DUE), El *Diccionario del español de México*, dir. L. F. Lara (DEM) y el *Wikcionario* (obra colaborativa). El verbo *almacenar* tenía una acepción independiente en el DRAE para el sentido informático del vocablo, pero la acepción es eliminada en el avance de la vigésima tercera edición. En *actualizar*, el DUEAE incluye el ejemplo *actualizar una base de datos* dentro de una definición general. La primera conclusión que se desprende de los datos expuestos en la Tabla 1 es que existe una gran disparidad en relación con la actualización de datos semánticos en la lexicografía castellana, lo cual no tiene por qué ser en sí mismo negativo o producto de la arbitrariedad, pues cada diccionario tiene sus propios criterios basados en concepciones teóricas, público o necesidades distintas –el diccionario es un “producto cultural”, según recuerda Lara (1997). Sin embargo, no es menos cierto que ninguna de estas obras está plenamente basada en procedimientos científicos como el análisis de corpus y las técnicas de lingüística cuantitativa, que deberían sostener objetivamente las decisiones de cada uno de los equipos lexicográficos. Un proyecto así podría enmarcarse en las líneas de la lexicografía de Gross o de Hanks, pero lo importante es que debería ser sistemático y capaz de seguir los cambios semánticos en tiempo real.

Con el objetivo de evaluar esta estrategia, la metodología consiste en mezclar estos verbos neológicos con una muestra aleatoria de verbos tomados de la base de datos OSLIN hasta llegar a un total de 250 verbos (Tabla 3) y comprobar entonces si este algoritmo es capaz de volver a separarlos.

abarrotar, abofetear, abreviar, abroquelar, acabangar, acarrazar, accionar, achicopalar, aclarar, acopar, acototar, acrisolar, activar, adjuntar, afietar, afiliar, aflojar, afluir, agachar, agramillar, alechugar, alfabetear, almacigar, almogavarear, aluminizar, ameritar, amoldar, ampliar, añadir, anisar, apocar, aporrillar, apostillar, arbolar, asediar, asordar, asperger, bermejar, bocadejar, borrajear, bracear, braquilar, broncear, calosfrir, canbar, cangallar, cantar, cardar, cargar, carpintear, cebar, cernir, cerverear, cesar, chacalear, chapotear, charlatanear, charquear, chicolear, chicotear, chufilar, cifrar, coagular, colicar, coludir, columpiar, combinar, comisionar, compadrear, confeccionar, congelar, congraciarse, conmemorar, conmutar, consagrar, construir, contramarcarse, copalear, cuerear, culear, deflagrar, denostar, deputar, desaparecer, desbonetar, descalandrajear, descansar, descorazonar, descubrir, desembanastar, desemepear, desempotrar, desengañillar, desenvainar, desgarrar, deshilar, desinflamar, desleír, desmotivar, desparejar, desposeer, desuñar, desuñar, detener, deteriorar, dialogar, dimidir, diputar, diseccionar, dislocar, disolver, embalsamar, embestir, embolar, empalagar, empaquetar, empastar, empericuetar, empuntar, enamorar, encarroñar, encatalejar, encerruar, endoblar, endoblar, engarbar, enhuevar, enrielar, enterrar, entrecerrar, entrechocar, entremediar, entremeter, entrujar, entubar, envinar, escabechar, esclavizar, esculpir, esparcir, estatuir, estropear, estropear, existir, expeditar, fagocitar, farrear, ferrificar, fichar, fortificar, fundir, galguezar, ganar, helenizar, hipotecar, huestear, intercambiar, intrigar, jochar, litar, machucar, madrugar, mangonear, mantener, marmullar, masar, merar, milonguear, mimar, montazgar, morrear, mudar, odiar, ominar, opilar, ovular, paginar, papear, paporear, parapetear, parcelar, parvificar, pavonar, pavordear, peñear, percutir, pirquinrear, pispilear, priorizar, potar, predisponer, presidir, priorizar, propiciar, proseguir, puntuar, quitar, rastrear, realzar, rebalsar, recercar, recomenzar, recovar, redolarcopiar, reencamar, refrigerar, regañir, regentear, reingresar, remangar, renunciar, repacer, repantigar, repoblar, respingar, resplandecer, retajar, reutilizar, revelar,

seccionar, sembrar, sexualizar, silabear, singar, sobrescribir, sollamar, sonetizar, sorber, tasar, teatralizar, terciar, tirantear, tiritar, transcurrir, transfigurar, trascurrir, trastear, trocar, turbar, viaticar, zabuquear

**Tabla 3:** Muestra aleatoria de 250 verbos de la base de datos OSLIN (Janssen, 2005)

#### 4. Resultados

De los 250 verbos iniciales, vemos que en las primeras 19 posiciones del ranking aparecen ya varias de las unidades que teníamos en nuestra lista inicial. Entre ellos hay otras que, sin estar en la lista inicial sino en la muestra aleatoria, han obtenido una alta puntuación como neologismos pero, de hecho, resultan dudosos, como comentaré más adelante. En cuanto a la evaluación del resultado, si somos rigurosos y sólo tomamos como buenos los verbos que aparecían en la lista inicial, el algoritmo fue capaz de captar 8 de los 13 verbos neológicos en las primeras 20 posiciones del ranking, es decir, se obtiene un 40% de precisión y un 61% de cobertura. En las primeras 50 posiciones del ranking aparecen ya 11 de los 13 verbos (el verbo número 12 aparece en la posición 51). Esto es, 22% de precisión y 84,6% de cobertura.

| RANGO | VERBO      | VALOR   | LISTA INICIAL |
|-------|------------|---------|---------------|
| 1     | activar    | 7394.91 |               |
| 2     | actualizar | 3251.33 | X             |
| 3     | arrastrar  | 1914.5  | X             |
| 4     | abreviar   | 1689    |               |
| 5     | existir    | 1650.25 |               |
| 6     | visitar    | 1582.5  | X             |
| 7     | añadir     | 1373.25 |               |
| 8     | copiar     | 1348.83 | X             |
| 9     | quitar     | 1345.08 |               |
| 10    | mantener   | 1097.92 |               |
| 11    | construir  | 1042.5  |               |
| 12    | combinar   | 853.5   |               |
| 13    | afiliar    | 831     |               |
| 14    | amoldar    | 672     |               |
| 15    | ejecutar   | 611.33  | X             |
| 16    | pegar      | 597.75  | X             |
| 17    | detener    | 581.25  |               |
| 18    | vincular   | 401.5   | X             |
| 19    | almacenar  | 372.5   | X             |
| 20    | revelar    | 355.25  |               |
| 21    | descargar  | 315     | X             |
| 22    | propiciar  | 307.75  |               |
| 23    | cifrar     | 269.5   |               |
| 24    | dialogar   | 199.67  |               |
| 25    | aclarar    | 181     |               |

|     |           |        |     |
|-----|-----------|--------|-----|
| 26  | cesar     | 141.5  |     |
| 27  | cargar    | 140.58 |     |
| 28  | soportar  | 136.5  | X   |
| 29  | recercar  | 123.25 |     |
| ... | ...       | ...    | ... |
| 40  | reiniciar | 68.5   | X   |
| ... | ...       | ...    | ... |
| 51  | enlazar   | 18.75  | X   |

**Tabla 4:** Primeras posiciones en la tabla de resultados

A la vista de estos resultados, comentar la significación estadística no parece relevante, en el sentido de que no hace falta demostrar que el algoritmo ha tenido un efecto sobre la clasificación. El test de la chi cuadrada, por ejemplo, nos diría que la probabilidad de haber obtenido este resultado por casualidad es prácticamente nula. Para ello debemos pensar cuál es la probabilidad de obtener 11 de los 13 verbos de la lista inicial concentrados en una quinta parte (el primer 20%) del ranking. Adoptando una hipótesis nula según la cual distribución fuese aleatoria, cabría esperar que los neologismos se encontraran repartidos por toda la tabla de manera más o menos uniforme, por tanto, en un 20% del ranking deberíamos encontrar alrededor de un 20% de los verbos. Para calcular esto, procedemos tal como se expresa en la Ecuación 2: por cada parte ( $i$ ) de las cinco en que dividimos el ranking ( $n$ ), calculamos el cuadrado de la diferencia entre la cantidad de verbos que observamos en esta parte del ranking y la cantidad de verbos que esperaríamos encontrar si la hipótesis nula fuese cierta. Para la serie de valores 11;1;1;0;0 (interpretable como: encontramos 11 verbos en la primera parte del ranking, uno en la segunda, otro en la tercera y cero en el resto) obtenemos el resultado 34.308 que, con cuatro grados de libertad (definidos como la cantidad de partes del ranking menos uno), equivale a un valor  $p$  inferior a 0.001.

$$\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i} \quad (2)$$

Para entrar más en el detalle de los resultados, las Tablas 3 a 9 muestran algunos de los elementos que aparecen como colocativos de los verbos y acusan su neologicidad. En las tablas, la frecuencia aparece dividida en tres períodos (aunque, como vimos, en el cálculo los dos primeros se suman). Sólo las primeras dos, *actualizar* y *copiar*, son casos de verbos que estaban efectivamente en la lista inicial. Del resto, podemos apreciar que algunos de los errores cometidos son consecuencia de problemas de ambigüedad con otras lenguas. En el caso de *activar*, por ejemplo, tenemos la ambigüedad con el inglés *active therapy*. Este es un error frecuente porque en el corpus hay todavía una importante cantidad de texto en otras lenguas, en particular inglés, catalán y francés.

Entre los casos considerados erróneos encontramos también algunos dudosos. Por ejemplo, *activar*, que es uno de los verbos que no estaba en la lista inicial sino en la muestra aleatoria pero que, por casualidad, también se relaciona con la informática. El error tiene origen en la cadena *activar la casilla de verificación*. ¿Se puede decir que este es un caso de neología semántica? ¿Ha cambiado el sentido del verbo *activar* en este contexto? El hecho de que *activar* se combine con un argumento nuevo no significa que el sentido del verbo haya cambiado, y por esta razón es justo que conste como un falso positivo. Sin embargo, a la luz de estos ejemplos el criterio que reúne a los verbos de la lista inicial parece cada vez más arbitrario. El verbo *copiar* (*archivos, texto, imágenes...*) es un caso similar al de *activar*. En el caso de *pegar*, parece más claro que se trata de un uso metafórico. También parece serlo

*cortar*, que Renau (2009) clasifica como neologismo semántico. Pero ¿es *copiar* 'datos informáticos' también un neologismo? En caso afirmativo, ¿no habría considerar a *activar* de la misma manera? El caso de *quitar* también puede ser dudoso. El sistema considera que *quitar* es neológico por expresiones como *agregar y quitar programas* o *clic en el botón quitar*. Es el mismo caso de *abreviar*, que se origina en la expresión *tecla del método abreviado*, que ha hecho creer al sistema que *abreviar* tenía un nuevo sentido. Lo mismo pasa con *añadir*. ¿Por qué *copiar* sí y *añadir* no? Ninguno de los diccionarios consultados registra acepciones relacionadas con la informática en estos casos dudosos, lo cual puede significar o bien que no han detectado la neología semántica o bien que no los han considerado neologismos. Con estos pocos ejemplos se puede observar que los procesos de neología semántica requerirán siempre una toma de posición con respecto a los significados previamente existentes, los cuales deberán estar descritos con gran precisión para facilitar el análisis de los significados recién llegados.

| Actualizar |           |      |      |      |       |                                                                                                                                                         |
|------------|-----------|------|------|------|-------|---------------------------------------------------------------------------------------------------------------------------------------------------------|
| rango      | forma     | per1 | per2 | per3 | score | contextos                                                                                                                                               |
| 0          | linkgua   | 0    | 0    | 1352 | 1352  | <i>Es trabajo cotidiano de los diputados federales actualizar las leyes y las normas que permiten la convivencia pacífica de todos los mexicanos...</i> |
| 1          | botón     | 0    | 0    | 519  | 519   |                                                                                                                                                         |
| 2          | federales | 0    | 0    | 410  | 410   |                                                                                                                                                         |
| 3          | mantente  | 0    | 0    | 381  | 381   |                                                                                                                                                         |
| 4          | diputados | 0    | 0    | 246  | 246   | <i>Las bibliografías de los libros de Linkgua son actualizadas en...</i>                                                                                |
| 5          | arthur    | 0    | 0    | 168  | 168   |                                                                                                                                                         |
| 6          | clic      | 0    | 0    | 89   | 89    | <i>clic en el botón actualizar</i>                                                                                                                      |
| 7          | opción    | 0    | 2    | 223  | 74.33 |                                                                                                                                                         |
| 8          | estado    | 735  | 928  | 2093 | 12    |                                                                                                                                                         |

**Tabla 5:** Detalle del análisis de *actualizar*

| Copiar |              |      |      |      |        |                                               |
|--------|--------------|------|------|------|--------|-----------------------------------------------|
| rango  | forma        | per1 | per2 | per3 | score  | contextos                                     |
| 0      | botón        | 0    | 2    | 748  | 249.33 | <i>clic en el botón copiar</i>                |
| 1      | opción       | 0    | 0    | 242  | 242    |                                               |
| 2      | formato      | 0    | 0    | 213  | 213    |                                               |
| 3      | transmitido  | 0    | 0    | 181  | 181    | <i>copiado o transmitido sin autorización</i> |
| 4      | cortar       | 0    | 0    | 154  | 154    |                                               |
| 5      | clic         | 0    | 0    | 132  | 132    |                                               |
| 6      | desea        | 0    | 1    | 237  | 118.5  |                                               |
| 7      | autorización | 0    | 0    | 59   | 59     |                                               |

**Tabla 6:** Detalle del análisis de *copiar*

| Activar |                |      |      |      |        |                                                                                                                    |
|---------|----------------|------|------|------|--------|--------------------------------------------------------------------------------------------------------------------|
| rango   | forma          | per1 | per2 | per3 | score  | contextos                                                                                                          |
| 1       | directory      | 0    | 0    | 3069 | 3069   | <i>active directory</i><br><i>active la casilla de verificación</i><br><i>highly active antiretroviral therapy</i> |
| 2       | casilla        | 0    | 0    | 1369 | 1369   |                                                                                                                    |
| 3       | ficha          | 0    | 0    | 1070 | 1070   |                                                                                                                    |
| 4       | opción         | 0    | 27   | 4243 | 748.75 |                                                                                                                    |
| 5       | antiretroviral | 0    | 0    | 475  | 475    |                                                                                                                    |
| 6       | therapy        | 0    | 0    | 227  | 227    |                                                                                                                    |
| 7       | verificación   | 0    | 0    | 144  | 144    |                                                                                                                    |
| 8       | server         | 0    | 2    | 310  | 103.33 |                                                                                                                    |
| 9       | pages          | 0    | 1    | 115  | 57.5   |                                                                                                                    |
| 10      | desactive      | 0    | 2    | 159  | 53     |                                                                                                                    |
| 11      | botón          | 0    | 1    | 80   | 40     |                                                                                                                    |
| 12      | receptores     | 9    | 11   | 233  | 29.66  |                                                                                                                    |
| 13      | siguen         | 0    | 2    | 26   | 8.66   |                                                                                                                    |

**Tabla 7:** Detalle del análisis de *activar*

| Quitar |           |      |      |      |        |                                                                     |
|--------|-----------|------|------|------|--------|---------------------------------------------------------------------|
| rango  | forma     | per1 | per2 | per3 | score  | contextos                                                           |
| 1      | botón     | 0    | 0    | 702  | 702    | <i>clic en el botón quitar</i><br><i>agregar o quitar programas</i> |
| 2      | programas | 0    | 3    | 1543 | 385.75 |                                                                     |
| 3      | clic      | 0    | 0    | 113  | 113    |                                                                     |
| 4      | opción    | 0    | 0    | 70   | 70     |                                                                     |
| 5      | remuevan  | 3    | 0    | 164  | 41     |                                                                     |
| 6      | sirva     | 0    | 0    | 12   | 12     |                                                                     |
| 7      | habelle   | 0    | 0    | 8    | 8      |                                                                     |
| 8      | permitió  | 0    | 2    | 23   | 7.66   |                                                                     |
| 9      | pidió     | 35   | 53   | 156  | 5.66   |                                                                     |

**Tabla 8:** Detalle del análisis de *quitar*

| Abreviar |              |      |      |      |       |                                                                                                               |
|----------|--------------|------|------|------|-------|---------------------------------------------------------------------------------------------------------------|
| rango    | forma        | per1 | per2 | per3 | score | contextos                                                                                                     |
| 0        | teclado      | 0    | 0    | 937  | 937   | <i>métodos abreviados de teclado</i><br><i>tecla de método abreviado</i><br><i>teclas de método abreviado</i> |
| 1        | modificación | 0    | 0    | 378  | 378   |                                                                                                               |
| 2        | teclas       | 0    | 0    | 174  | 174   |                                                                                                               |
| 3        | modelos      | 0    | 0    | 106  | 106   |                                                                                                               |
| 4        | tecla        | 0    | 0    | 94   | 94    |                                                                                                               |

**Tabla 9:** Detalle del análisis de *abreviar*

| Añadir |              |      |      |      |        |                                                                                                               |
|--------|--------------|------|------|------|--------|---------------------------------------------------------------------------------------------------------------|
| rango  | forma        | per1 | per2 | per3 | score  | contextos                                                                                                     |
| 0      | clase        | 0    | 0    | 672  | 672    | añadir una copia en disquete<br>añadir notas de clase para<br>el botón añadir<br>retirar del fuego y añadir a |
| 1      | botón        | 0    | 0    | 418  | 418    |                                                                                                               |
| 2      | copia        | 2    | 0    | 328  | 109.33 |                                                                                                               |
| 3      | disquete     | 0    | 0    | 58   | 58     |                                                                                                               |
| 4      | retirar      | 1    | 0    | 89   | 44.5   |                                                                                                               |
| 5      | volviendo    | 0    | 1    | 22   | 11     |                                                                                                               |
| 6      | terminó      | 1    | 0    | 19   | 9.5    |                                                                                                               |
| 7      | onzas        | 0    | 2    | 26   | 8.66   |                                                                                                               |
| 8      | recomiendan  | 3    | 0    | 33   | 8.25   |                                                                                                               |
| 9      | levantándose | 0    | 1    | 16   | 8      |                                                                                                               |
| 10     | repente      | 1    | 1    | 24   | 8      |                                                                                                               |
| 11     | marquesa     | 0    | 0    | 7    | 7      |                                                                                                               |
| 12     | medida       | 43   | 58   | 218  | 5.75   |                                                                                                               |
| 13     | sacando      | 3    | 0    | 21   | 5.25   |                                                                                                               |

**Tabla 10:** Detalle del análisis de añadir

## 5. Conclusiones y trabajo futuro

En este resumen del seminario se ha presentado una introducción a la neología semántica desde un enfoque cuantitativo y también los resultados preliminares de un experimento de extracción de neologismos semánticos basado en estadísticas de coocurrencia. Uno de los aspectos que resaltan de este trabajo es la utilización de los listados de engramas del corpus de libros digitalizados por Google, una recopilación que se extiende desde los años de Gutenberg hasta el presente. Gracias a este corpus, ahora podemos tener una idea bastante aproximada de cuándo entró cada palabra en la historia de la lengua, cuándo se dejó de usar y hasta cómo cambió la manera de usarla. El presente trabajo ha pretendido presentar una propuesta concreta en la medición del cambio semántico a través del estudio de las palabras con las que una unidad analizada se combina en una ventana de contexto limitada a cinco palabras. La metodología con la que procedemos en el análisis es original, ya que si bien continúa las líneas exploradas por otros autores, su diferencia fundamental radica en que se trata de una idea conceptual y computacionalmente más simple. Los resultados obtenidos por el momento son cuanto menos interesantes, considerando que estamos en los primeros pasos de esta investigación.

La conclusión más importante que se desprende del presente trabajo sobre la neología semántica es que se confirman las ideas vertidas por Bastuji (1974) en cuanto a la gran dificultad para establecer una distinción clara entre aquello que es una unidad léxica y aquello que es una expresión referencial. La unidad léxica pertenece al sistema de la lengua, mientras que la referencia es únicamente una propiedad del discurso. En lo que resulta una paradoja, la unidad léxica es constante en tanto signo lingüístico pero es también variable, ya que su sentido se ve afectado por el uso que se hace de ella en el discurso. No podemos dejar de relacionar estas ideas con las que presentó simultáneamente Eco (1975), coincidencia no del todo sorprendente teniendo en cuenta que ambos siguen una línea de pensamiento ya presente en Benveniste (1974), entre otros. En el *Tratado de semiótica general*, Eco aboga por la necesidad de distinguir, tanto en lingüística como en semiótica, entre un estudio de la significación, que podemos relacionar con el estudio de la lengua, la competencia o los códigos, y un estudio de la

comunicación o de la producción de signos, que podemos entender como el estudio del habla, de la actuación o el análisis del discurso. Si bien se los puede estudiar por separado, no son ámbitos independientes, ya que existe entre ambos una tensión fundamental. Por un lado, puede haber creatividad en el uso de la lengua, pero esa creatividad está sujeta a las reglas de la lengua (en términos de Eco, del *código*). Dicho esto, también es posible que la creatividad del hablante llegue a subvertir las reglas cuando el mensaje instaura su propio código. Si observamos este proceso en su dimensión temporal, como hemos hecho a lo largo de este trabajo, comprobamos que un uso creativo del lenguaje puede convertirse en una unidad léxica para la siguiente generación de hablantes sin que nadie llegue a ser consciente del cambio.

Como una conclusión secundaria de este trabajo, el estudio de la creación de nuevas unidades léxicas suscita también la reflexión sobre cómo se mantiene el equilibrio en el vocabulario de la lengua. Ya en los análisis exploratorios del corpus comprobamos que a pesar de la entrada en circulación de nuevas unidades léxicas, el tamaño del vocabulario castellano se mantiene relativamente estable en el tiempo, de modo que la cantidad de palabras nuevas guarda una relación con la cantidad de palabras (o sentidos) que se dejan de utilizar. Estos datos confirman la concepción de la lengua como un sistema autorregulado, capaz de encontrar su propio equilibrio y controlar, entre otras variables, el vocabulario de la comunidad de hablantes. Este orden, que no viene directamente impuesto por una institución y parece mantenerse de forma espontánea, es necesario para el buen funcionamiento de la comunicación, puesto que si se disparase la creación de nuevas unidades léxicas, la comunicación no sería posible. De esta manera, la lengua se nos presenta como un sistema aparentemente racional, y sin embargo no depende de la razón del hablante, sino que se trata de una propiedad emergente del conjunto de ellos. La lengua parece así controlada por una mano invisible, como en las teorías de la economía de libre mercado de Adam Smith. Naturalmente, las instituciones que imponen la norma lingüística tienen también su influencia. Los hablantes de una lengua aceptan en mayor o menor medida una especie de contrato social lingüístico, por el que delegan la toma de decisiones sobre lo que es correcto en cuestión de norma lingüística en una autoridad centralizada. Dicho esto, no existe institución que imponga obediencia automática en la población de hablantes. Es, en cualquier caso, en la tensión entre la norma lingüística y la evolución natural de la lengua donde se define la futura norma. Esa evolución no es sólo natural sino también histórica, y por ello no se puede soslayar la importancia de una variable extra-lingüística como las revoluciones tecnológicas en los medios de comunicación. Deberíamos calibrar el efecto de las tecnologías en el sistema de la lengua, incluso determinar si hablamos de efecto o de transformación. Contrariamente a los medios de comunicación de hace unas décadas, que favorecerían la difusión de información desde puntos centralizados, medios como Internet permiten a los usuarios comunicarse entre ellos saltándose las instituciones centralizadas y su función reguladora. En contrapartida, las tecnologías facilitan el conocimiento de la norma lingüística, y el hablante tiene acceso a diversas fuentes de conocimiento y puede conocer el uso correcto de la lengua, ya sea utilizando las fuentes autorizadas o bien utilizando motores de búsqueda de Internet como corpus para juzgar la corrección de una determinada expresión, a modo de consulta de la voz popular.

Para finalizar, me gustaría comentar algunas de las posibilidades de trabajo futuro de esta investigación, entre ellas la de incluir mayor información lingüística en el experimento. El corpus con el que hemos trabajado en este seminario consiste sólo en formas, de modo que para afinar el análisis será necesario proceder con su lematización, etiquetado morfosintáctico, *chunking* para la detección de unidades sintagmáticas y análisis de dependencias para estudiar la relación entre predicados y argumentos. Otra dirección de trabajo es repetir el experimento en otro corpus. Como ya se mencionó, la elección de este corpus está motivada por el gran tamaño que tiene, pero es posible que este mismo

experimento se pueda repetir con un corpus de menor tamaño, como los archivos de un periódico. Como trabajo a medio plazo, hay un proyecto bastante más ambicioso que es hacer una especie de historia del ciclo de vida de las palabras incorporando el análisis de coocurrencias. En este proyecto analizaremos cada palabra para ver cuáles son sus patrones típicos de combinación. Por ejemplo, si es un verbo, estudiaremos cuáles son sus argumentos, designados como clases de objetos. Para poder llevar a cabo un proyecto así será necesario desarrollar una taxonomía del castellano (que está de hecho en proceso actualmente). Una vez que dispongamos de estos patrones de combinación, luego podremos separarlos por épocas, distinguiendo entre patrones estables (consolidados), patrones nuevos y patrones que ya no se utilizan.

**Comentario Mercè Lorente:** Sabes que existen ya taxonomías del castellano, como WordNet.

**Re:** Sí, pero para un proyecto de estas características sería necesario desarrollar una taxonomía con una cobertura mucho mayor a la que tienen las versiones de WordNet en castellano que circulan actualmente.

**Comentario Lluís De Yzaguirre:** Una posibilidad de trabajo futuro puede ser la utilización de las marcas temáticas que contienen los diccionarios para obtener unidades que tienen un sentido especializado además de uno general y de esa manera ver si el sistema es capaz de detectar esos usos en el corpus.

**Re:** Es buena idea, y la utilizaré en el futuro para extraer las palabras para analizar, como los verbos relacionados con la informática de este experimento. En general, es mejor que las fuentes lexicográficas no formen parte del algoritmo (para que éste esté totalmente basado en corpus), pero si se trata de utilizar fuentes lexicográficas para evaluar la estrategia, entonces no veo inconveniente.

**Comentario Muntsa Padró:** Posiblemente no cambie los resultados del trabajo pero creo que en lugar de frecuencia absoluta deberías utilizar frecuencias relativas al año.

**Comentario Jorge Vivaldi:** Yo también creo que la frecuencia tiene que ser relativa, las diferencias de tamaño lo justifican. Y con respecto a la inclusión de información lingüística, eso es algo muy complicado ya que si intentas lematizar y etiquetar morfosintácticamente listados de enigramas encontrarás que no tienes suficiente contexto para desambiguar.

**Re:** Lo de la frecuencia absoluta fue para simplificar esta explicación, y también para facilitar el cálculo, ya que no creo que las diferencias de tamaño del corpus en los años de la muestra sean tan grandes como para tener un efecto importante en el resultado. En cuanto a la lematización y el etiquetado morfosintáctico de los enigramas, sin duda es un tema suficientemente complejo e interesante como para dedicarle un artículo.

## Referencias

- Baayen, H., Renouf, A. (1996). Chronocling the Times: Productive Lexical Innovations in an English Newspaper . Language, Vol. 72, Nº. 1 (Mar., 1996), pp. 69-96.
- Bastuji, J. (1974). Aspects de la néologie sémantique. Langages. Vol. 8 Nº. 36, pp. 6-19.
- Benveniste, É. (1974). Problemas de lingüística general. México, D.F. Siglo XXI.

- Blank, A. (1999). Why do new meanings occur? A cognitive typology of the motivations of lexical change. En Blank, A. y Koch, P. (eds.). *Historical Semantics and Cognition*. Berlín: Mouton de Gruyter, pp. 61–89.
- Bloomfield, L. (1933). *Language*. New York: Allen & Unwin.
- Boussidan, A., Ploux, S. (2011). Using Topic Salience and Connotational Drifts to Detect Candidates to Semantic Change. *Bréal, M. (1904), Essai de sémantique*. Paris: Hachette.
- Cabré, M. Teresa (2009). La classificació dels neologismes: una tasca complexa. En *Les paraules noves: criteris per detectar i mesurar els neologismes*. Vic, Barcelona: Eumo; Universitat Pompeu Fabra, pp. 11–37.
- Cabré, M.T. & Estopà, R. (2009). *Les paraules noves. Criteris per detectar i mesurar els neologismes*. Vic/Barcelona: Eumo Editorial/Universitat Pompeu Fabra, pp. 11–37.
- Cabré, M.T., Janssen, M., Nazar, R., Vivaldi, J. (2011). Detección semiautomática de neologismos: los neologismos gramaticales. Trabajo presentado en el II Congreso Internacional de Neología de las Lenguas Románicas (CINEO 2011).
- Cabré, T.; Nazar, R. (en prensa). Towards a new approach to the study of neology. Experiments on formal, syntagmatic and semantic neology. *Neologica. Revue internationale de néologie*.
- Cannon, G. (1988). *Historical change and English word-formation*. New York: Peter Lang.
- Collier, A. (1998). Identifying diachronic change in semantic relations. En *Explorations in corpus linguistics* A. Renouf (ed.). Rodopi: Amsterdam, pp. 259–268.
- Cook, P., Stevenson, S. (2010). Automatically identifying changes in the semantic orientation of words. *Proceedings of the 7th International Conference on Language Resources and Evaluation*. Valletta, Malta, pp. 28–34.
- Cook, P.; Hirst, G. (2011). Automatic identification of words with novel but infrequent senses. *Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation*, Singapore.
- Diccionario Clave de Uso Del Español Actual. Diccionarios SM.
- Diccionario de la Lengua Española. 22. Real Academia Española.
- Diccionario del Español Usual en México, El Colegio de México.
- Diccionario de Uso del Español de América y España. Vox.
- Diccionario de Uso del Español María Moliner. 2. Alfaguara.
- Eco, U. (1975/1991). *Tratado de semiótica general*. Barcelona: Lumen.
- Estopà, R. (2009). La neologia especialitzada, rept constant per al mediador lingüístic. En: *Aspectes de terminologia, neologia i traducció* Coromina, E. y Mestres, J.M. (ed.) . Barcelona: Institut d'Estudis Catalans, 2010, pp. 15–39 .
- Feliu, J.; Martínez, L. ; Salazar, H. R. ; Tadeo, B. (2009). Els neologismes formats per resemantització. En *Les paraules noves: criteris per detectar i mesurar els neologismes*. Vic, Barcelona: Eumo, pp. 91–110.
- Freixa, J. Luna, R., Suárez, M. (2010). La neología semántica en las Antenas Neológicas: descripción de algunos fenómenos relevantes. *Actes del I Congrés Internacional de Neologia de les Llengües Romàniques*. M. Teresa Cabré, Ona Domènech, Rosa Estopà, Judit Freixa i Mercè Lorente (eds.). Documenta Universitaria, Barcelona, pp. 835–854.
- Gross, G. (1994). *Classes d'objets et description des verbes, *Langages*, N° 115, pp. 15–30.*
- Groult, M., Louis, P. & Roger, J. (1988). *Transfert de vocabulaire dans les sciences*. Paris: Éditions du Centre National de la Recherche Scientifique.
- Guilbert, L. (1975). *La créativité lexicale*. Paris: Librairie Larousse.
- Gulordava, K., Baroni, M. (2011). A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus. *Proceedings of the GEMS 2011 Workshop on Geometrical Models of Natural Language Semantics*, Edinburgh, Scotland, pp. 67–71.
- Harris, Z. (1954/1985). *Distributional Structure*. En J.J. Katz (ed.) *The philosophy of linguistics*. New York: Oxford University Press, pp. 26–47.
- Hanks, P. (2004). The Syntagmatics of Metaphor and Idiom. *International Journal of Lexicography*, Vol. 17, N° 3, pp. 245–274.
- Heaps, H. S. (1978). *Information retrieval: computational and theoretical aspects*. New York: Academic Press.
- Holz, F., Teresniak S (2010). Towards Automatic Detection and Tracking of Topic Change. Gelbukh, A (ed.) *Proceedings of CICLing 2010, I Conference on Intelligent Text Processing and Computational Linguistics*, LNCS 6008. Springer LNCS, 327–339.
- Issac, F., Benhariz Ouenniche, S. (2008 / 2010). Pour une veille néologique a partir du web: l'outil telanaute, *Actes del I Congrés Internacional de Neologia de les Llengües Romàniques*. M. Teresa Cabré, Ona Domènech, Rosa Estopà, Judit Freixa i Mercè Lorente (eds.). Documenta Universitaria, Barcelona, pp. 1163–1172.
- Janssen, M. (2005). *Open Source Lexical Information Network*. Third International Workshop on Generative Approaches to the Lexicon, Geneva, Switzerland, pp. 79–106.
- Janssen, M. (2009). Detección de Neologismos: una perspectiva computacional. *Debate Terminológico 5*. París: RITERM (Red Iberoamericana de Terminología), pp. 68–75.

- Janssen, M. (2010). NeoTrack: Un analyseur de néologismes en ligne. Actes del I Congrés Internacional de Neologia de les Llengües Romàniques. M. Teresa Cabré, Ona Domènech, Rosa Estopà, Judit Freixa i Mercè Lorente (eds.). Documenta Universitaria, Barcelona, pp. 1173–1186.
- Lara, L. F. (1997). Teoría del diccionario monolingüe. El Colegio de México.
- Lorente, M. (2011). Neología y gramática. Trabajo presentado en el II Congreso Internacional de Neología de las Lenguas Románicas (CINEO 2011).
- Makri, J. (2009). La création lexicale en espagnol péninsulaire contemporain: Étude néologique, typologie des procédés et réflexions. Tesis doctoral. Montpellier: Universitat Paul-Valéry.
- Martínez de Sousa, José (2005). ¿El Diccionario panhispánico de dudas cumple con su deber?. <http://www.martinezdesousa.net/az-dpd.pdf> [Con acceso Diciembre 2011].
- Mejri, S. (2006). La reconnaissance automatique des néologismes de sens. En Blampain, D.; Thoiron, P.; Van Campenhout, M. (éd). Mots, termes et contextes, Actes des septièmes Journées scientifiques du réseau LTT. Paris: Editions des Archives Contemporaines et Agence universitaire de la francophonie, pp. 545–557.
- Mejri, S. (2010). Néologie et traitement automatique. 2010 Actes del I Congrés Internacional de Neologia de les Llengües Romàniques. M. Teresa Cabré, Ona Domènech, Rosa Estopà, Judit Freixa i Mercè Lorente (eds.). Documenta Universitaria, Barcelona, pp. 99–116.
- Meyer, I. & Mackintosh, K. (2000). "L'Étirement" du sens terminologique: Aperçu du phénomène de la détermination. En H. Béjoint & P. Thoiron (Eds.), Le Sens en Terminologie. Lyon: Presses Universitaires de Lyon, pp. 198–217.
- Michel J. B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Google Books Team, Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. A. & Aiden, E. L. (2010). Quantitative analysis of culture using millions of digitized books science. Science, Vol. 331, N° 6014, pp. 176–182.
- Nazar, R. (2010). A Quantitative Approach to Concept Analysis. Tesis doctoral, Universidad Pompeu Fabra, Barcelona.
- Nazar, R. (2011). Estudio diacrónico de la terminología especializada utilizando métodos cuantitativos: Ejemplos de aplicación a un corpus de artículos de lingüística aplicada. Revista Signos, Vol. 44, N° 75, pp. 48–67.
- Nazar, R.; Azarian, J. (2010). The Living Lexicon: methodology to set up synchronic dictionaries. Proceedings of the XIV Euralex International Congress (Leeuwarden, 6-10 July 2010). Ljouwert: Fryske Akademy. pp. 527–538.
- Nazar, R. & Vidal, V. (2010). Aproximación cuantitativa a la neología. Actes del I Congrés Internacional de Neologia de les Llengües Romàniques. M. Teresa Cabré, Ona Domènech, Rosa Estopà, Judit Freixa i Mercè Lorente (eds.). Documenta Universitaria, Barcelona, pp. 865–878.
- Renau, I. (2009). Estudio lexicográfico de verbos del español en relación con los usos de se y con vistas a la confección de un diccionario para extranjeros. Actas del XXIV Encuentro Internacional de la Asociación de Jóvenes Lingüistas, Universitat Autònoma de Barcelona, Bellaterra (Barcelona), pp. 1–12.
- Renau, I., Nazar, R (2011). Análisis cuantitativo del uso real de los verbos pronominales estrictos del castellano utilizando un corpus diacrónico (Google Books). Actas del III Congreso Internacional de Lingüística de Corpus, AELINCO, Valencia, pp. 287–298.
- Renouf, A. (2010). Identification automatique de la néologie lexicologique et sémantique: questions soulevées par notre méthode. Actes del I Congrés Internacional de Neologia de les Llengües Romàniques. M. Teresa Cabré, Ona Domènech, Rosa Estopà, Judit Freixa i Mercè Lorente (eds.). Documenta Universitaria, Barcelona, pp. 129–142.
- Rohrdantz, C., Haulti, A. Mayer, T., Butt, M., Keim, D., Plank, F. (2011) Towards Tracking Semantic Change by Visual Analytics. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, Portland, Oregon, pp. 305–310.
- Sablaryrolles, J.F. (1997). Néologismes: Une typologie des typologies. Cahier du CIEL, pp. 11–48.
- Sagi, E., Kaufmann, S., Clark, B. (2009). Semantic density analysis: Comparing Word Meaning Across Time and Phonetic Space. Proceedings of the EACL 2009 . Workshop on GEMS: Geometrical Models of Natural Language Semantics. Athens, Greece, pp. 104–111.
- Sánchez Manzanares , M.C. (2009). Procedimientos trópicos en la neología semántica: sistematicidad y creatividad. Revista de Investigación Lingüística, Universidad de Murcia, N° 12, pp. 123–146.
- Schierz, A. (2007). Monitoring knowledge: A text-based approach. Terminology, Vol. 13, N° 2, pp. 125–154.
- Steffens, M., Baiwir , E. (2009). Neología semántica y polisemia: el caso del truncamiento. Revista de Investigación Lingüística, Universidad de Murcia , N° 12, pp. 175–184.
- Traugott, E. Dasher, R. (2002). Regularity in Semantic Change. Cambridge University Press.
- Ullmann, S. (1957). Principles of Semantics, Oxford: Blackwell.
- 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). Barcelona: Observatori de Neologia, IULA, pp. 85–108.