Detección de Neologismos: una perspectiva computacional

Maarten Janssen[1]

Resumo: Nos últimos vinte anos, tem sido construído um número elevado de ferramentas novas para a detecção automática de neologismos. No entanto, como este artigo pretende demonstrar, o funcionamento básico das ferramentas mais recentes é ainda quase o mesmo que o de ferramentas mais antigas. A razão principal para este falta de mudança é que, no fundo, existem apenas três maneiras diferentes para a detecção automática: (1) utilizar uma lista independente de palavras conhecidas (*lista de exclusão*), (2) utilizar padrões linguísticos que caracterizam os neologismos, (3) contar as ocorrências das palavras, comparando um texto recente com um corpus de textos mais antigos. Este artigo descreve essas três abordagens, com as suas vantagens e problemas. Particularmente, explica-se porque as ferramentas recentes utilizam a exclusão lexicográfica quase que da mesma maneira que as primeiras ferramentas. Também é introduzido um conjunto de termos para a detecção automática de neologismos a partir da perspectiva computacional.

Palavras chave: Neologismo Ortográfico, Detecção Automática, Ocorrência Neológica, Neologia

Resumen: En los últimos veinte años se ha construido un número elevado de nuevas herramientas para la búsqueda automática de neologismos. Pero, como este artículo pretende demostrar, el funcionamiento de base de las herramientas más recientes es todavía casi el mismo que el de las herramientas más antiguas. El motivo principal para esta falta de cambio es que en el fondo hay solamente tres maneras diferentes para la detección automática: (1) utilizar una lista independiente de palabras conocidas (lista de exclusión), (2) utilizar patrones lingüísticos que caracterizan los neologismos, (3) contar las ocurrencias de las palabras, comparando un texto reciente con un corpus de textos de fechas anteriores. Este artículo presenta estos tres abordajes, señalando las ventajas y problemas de todos ellos. Particularmente, se explica por qué las herramientas recientes utilizan la exclusión lexicográfica, que es casi la misma manera de operar que las primeras herramientas. También se introduce un número de términos para la detección de neologismos desde un punto de vista computacional.

Palabras clave: Neologismo Ortográfico, Detección Automática, Ocurrencia Neològica, Neología

Abstract: Over the last twenty years, a significant number of automatic neologism tools have been developed. This article tries to show that the recent tools work basically in the same way as the older tools. This is largely because there are only three fundamentally different ways of detecting neologisms: (1) by ways of exclusion based of a list of known words (for instance a dictionary); (2) by looking for patterns in the text that are characteristic for neologisms; (3) by using statistical analysis of texts, typically the comparison of distributional behaviour of words between new texts and older texts. This article gives an overview of these computational methods, with a view on their respective advantages and problems. It also introduces some terminology for the automatic detection of neologisms from a computational perspective. Moreover it shows why the newest tools are based on lexicographic exclusion in almost the same way as the first computational neologism detection tools were.

Keywords: Orthographic Neologism, Automatic Detection, Neologistic Occurrence, Neology

Cómo citar este artículo: JANSSEN, MAARTEN. Detección de neologismos: una perspectiva computacional. *Debate Terminológico.* Ago. 2009, No. 05; pp. 68-75

Artículo recibido: Mayo 2009. **Aprobado:** Junio 2009

[1] IULA - Universitat Pompeu Fabra. Maarten.Janssen@upf.edu

Introducción

La búsqueda y el análisis de neologismos tienen una tradición amplia, especialmente en los países de habla inglesa. Por razones evidentes, el trabajo sistemático en neología se ha beneficiado en los últimos veinte años de los ordenadores, empezando por el proyecto "Aviator" (Renouf 1993). La neología originalmente era solamente una práctica lexicográfica, cuya finalidad era recopilar palabras nuevas que entran en la lengua para la creación de diccionarios de neologismos y para la actualización de diccionarios de lengua existentes. Pero ahora, la neología se ha transformado en una disciplina extensa, que tiene un número de finalidades diversas, entre las cuales se encuentra el perfeccionamiento de las herramientas para el procesamiento del lenguaje, la regularización de préstamos y el estudio de la productividad morfológica.

Actualmente, los ordenadores no solamente son mucho más potentes que en los años ochenta, también son más comunes, lo que implica que hay más gente capaz de desarrollar programas útiles para el trabajo en neología. Y debido a la popularización de los ordenadores, los textos en soporte digital son ahora tan corrientes que la confección de un corpus para la extracción de neologismos ha dejado de ser un trabajo complejo, como lo era anteriormente.

Si bien el avance en lingüística computacional ha sido enorme en las últimas dos décadas, el software para la detección de neologismos no ha cambiado significativamente. Desde la perspectiva del usuario, las diferencias entre Aviator y herramientas más recientes como Telenaute (Issac 2008) son probablemente bastante grandes, sin embargo, la funcionalidad computacional es casi la misma. Existen dos motivos principales de esta falta de cambio. En primer lugar, no hay muchas maneras fundamentalmente diferentes para detectar neologismos de manera automática. Y en segundo lugar, aunque la búsqueda de neología haya cambiado mucho, los objetivos de las herramientas para la detección automática siguen siendo los mismos.

Este artículo tiene dos objetivos: por un lado, intenta dar una perspectiva global de las maneras diferentes en las cuales los ordenadores pueden ser utilizados en la búsqueda de neologismos. Por otro lado, intenta reanalizar la noción de neologismo desde una perspectiva computacional. El trabajo que presentamos consta de tres partes: las secciones 2 y 4 proporcionan una descripción de los métodos (semi)-automáticos para la detección de neologismos. La sección 3 ofrece algunas nociones de neologicidad desde la perspectiva computacional. Finalmente, la sección 5 discute cuáles de los métodos automáticos son los más adecuados para los diferentes objetivos de la neología.

1. Abordajes computacionales

La detección automática de neologismos es una tentativa para reconocer todos los neologismos de un conjunto de textos nuevos (*corpus de estudio*). En algunos casos se hace utilizando una comparación entre el corpus de estudio y un conjunto de textos más antiguos (*corpus de referencia*).

Existen básicamente tres abordajes distintos para la detección automática: (1) utilizar una lista independiente de palabras conocidas (*lista de exclusión*), (2) utilizar patrones lingüísticos que caracterizan los neologismos, (3) contar las ocurrencias de las palabras del corpus de estudio. Esta sección esboza estos tres abordajes brevemente. Esta descripción enfoca los métodos semiautomáticos, que tienen el objetivo de detectar los *candidatos a neologismo*: palabras que probablemente son neologismos, pero para los cuales es necesaria aún la intervención humana para decidir si lo son o no. Para convertir estos métodos en completamente automáticos, se necesitan herramientas de filtraje adicionales para automatizar el proceso de decisión.

1.1 Listas de exclusión

La idea más simple y obvia para detectar palabras nuevas es contrastar una lista de palabras conocidas (*lista de exclusión*). En este caso, los candidatos a neologismo son las palabras que se encuentran dentro del corpus de estudio y que no forman parte de la lista de exclusión.

Esta es la idea que hace funcionar la mayoría de las herramientas para la detección de neologismos, utilizada en el proyecto Aviator y aun aplicada sin cambios fundamentales en Cénit (Roche & Bowker 1999), Sextan (Vivaldi 2000), NeoTrack (Janssen 2005a), Telenaute (Issac 2008) y Buscaneo (Cabré &

Estopà 2009). No es solamente un abordaje fiable, sino también un método fácil de implementar. La funcionalidad básica puede ser reducida a una línea de comandos UNIX:

(1) cat corpus.txt | tokenize | sort | uniq | join –v 1 – listaexclusion.txt

Esta línea comienza con el corpus de estudio (*corpus.txt*) y hace cuatro pasos consecutivos: colocar cada palabra del texto en una línea distinta (*tokenize*), alfabetizar la lista de palabras (*sort*), retirar las duplicaciones (*uniq*), y eliminar todas las palabras que constan en la lista de exclusión (*listaexclusion.txt*). La lista de palabras que queda contiene los candidatos a neologismo.

Hay dos tipos de fuente para la lista de exclusión: la más genérica es utilizar la lista de todas las palabras del corpus de referencia. Pero la más común es utilizar un diccionario desplegado (incluyendo las formas flexionadas). Es posible utilizar la lista de exclusión como una lista fija, pero también se puede utilizar un método más dinámico, como en NeoTrack: el software está integrado en una base de datos léxica llamada OSLIN (Janssen 2005b), que no solamente funciona como lista de exclusión, sino también para agregar las palabras encontradas en el proceso de detección de neologismos.

El hecho de que la mayoría de los detectores de neologismos utilizan la misma estrategia para detectar los neologismos (aunque implementada de maneras diferentes) no significa que sean iguales: hay diferencias no solamente en el origen de la lista de exclusión (que afecta a los resultados significativamente), sino también en la manera de tratar los candidatos. Todos los sistemas tienen un método para rechazar los nombres propios, comúnmente basado en el hecho de que los nombres propios se escriben con mayúscula. Cénit (Roche & Bowker 1999) utiliza un analizador morfológico para rechazar las formas flexionadas (y derivadas). Sextan (Vivaldi 2000) utiliza un etiquetador morfológico con el mismo objeto (ver también 4.1). Finalmente, el diseño y la funcionalidad de la interfaz son bastante diferentes en cada una de las herramientas.

1.2 Patrones léxico-sintácticos

La única tentativa de detectar neologismos sin corpus de referencia ni lista de exclusión es utilizar un método que actualmente es conocido como extracción con *patrones léxico-sintácticos*. La idea es que un neologismo puede ser reconocido observando las palabras en su contexto. Por ejemplo, la frase *que se llama* en (2) es un indicación de que la palabra *UMD* o *universal media disc* es un neologismo.

(2) Lo nuevo de la PSP es el formato de los juegos, según dicen es un tipo de minidisc de alta capacidad *que se llama UMD* "universal media disc" y que…

En este ejemplo, la construcción *"que se llama X"* es el patrón léxico-sintáctico, y los candidatos a neologismos son todas las palabras que se encuentran en el corpus de estudio en la posición X de este patrón, tal como otras palabras en posiciones determinadas respecto de otros patrones como, por ejemplo, las palabras en itálicas o las palabras entre paréntesis. Para una descripción de los patrones polacos para los neologismos ver Paryzek (2008). La idea de utilizar este método de Chlebda (1991) es anterior a la obra más conocida de Hearst (1992), que propone utilizar patrones léxico-sintácticos para la detección de relaciones hiperonímicas. La extracción con patrones se funda en la idea de que los neologismos son palabras poco conocidas por los lectores y, por lo tanto, son introducidas de una manera especial. No todos los candidatos que cumplen con los requisitos de los patrones propuestos son neologismos, también habrá palabras terminológicas no neológicas y otros tipos de ruido.

1.3 Análisis estadístico

El último método para la detección de neologismos se basa en cuantificar las ocurrencias de palabras en el corpus de estudio, normalmente en comparación con el corpus de referencia. Hay muchas variantes posibles para cuantificar palabras, pero se puede distinguir fundamentalmente cuatro aproximaciones distintas.

La primera aproximación no utiliza ningún corpus de referencia. La idea, normalmente atribuida a Baayen & Renouf (1996; 1998) y utilizada, por ejemplo, en NeoloSearch (Janicijevic & Walker 1997) es que los neologismos se encuentran en los *hapax legomena* del corpus de estudio: las palabras que ocurran exactamente una sola vez. Entonces, los candidatos a neologismo son estas palabras que se encuentran una vez en el corpus de estudio. Es un método basado en la aparición de palabras, pero no

un abordaje estadístico.

La segunda aproximación propone que los neologismos del corpus de estudio son todas las palabras de este corpus que tienen frecuencia cero en el corpus de referencia. Esta era la idea original de Aviator (Renouf 1993). Aun siendo un método, en principio, de frecuencia de aparición de palabras, no es distinto de utilizar una lista de exclusión basada en el corpus de referencia.

El tercer abordaje consiste en comparar la frecuencia de todas las palabras en los dos corpus. Cuanto mayor sea la frecuencia de la palabra en el corpus de estudio en comparación con su frecuencia en el corpus de referencia, tanto más probable será que se trate de un neologismo. Por ejemplo, una palabra que se utiliza diez veces en los textos de fechas anteriores (corpus de referencia), pero mil veces en los textos nuevos (corpus de estudio), probablemente es una palabra que ha cambiado su uso, y entonces probablemente es un neologismo semántico. Una variante de este método es dividir el corpus de referencia en porciones temporales, por ejemplo, en periodos de un año y hacer una gráfica del progreso del uso de la palabra durante el tiempo (Nazar & Vidal 2008). En este caso, las palabras que exhiben un crecimiento exponencial en su frecuencia de uso son los candidatos a neologismo.

El cuarto abordaje no consiste en contar la frecuencia de las palabras en sí misma, sino la frecuencia de las palabras que ocurren en su contexto. La idea es que cuando para una palabra dada cambian las palabras con las cuales se combina, eso es indicativo de un cambio del uso (o significado) de esta palabra. Este método es el más complejo de los cuatro y en este momento no hay sistemas operacionales basados en él.

1.4 Conclusión

Resumiendo, hay tres maneras diferentes de enfocar la detección automática de (candidatos a) neologismos: el basado en listas de exclusión, el de los patrones léxico-sintácticos y el análisis estadístico. Los patrones son utilizados solamente de manera marginal, y aunque hay muchos proyectos para la utilización de métodos estadísticos, la mayoría de las herramientas existentes están basadas en métodos de exclusión. No solamente porque son los más fáciles de implementar, sino también porque es el único método que puede ser utilizado con textos pequeños como ediciones individuales de diarios (ver 5.1). Antes de retomar la discusión sobre cuál de los métodos es el más adecuado para cada tipo de búsqueda de neologismos en la sección 5, las dos secciones siguientes introducen algunas particularidades de la búsqueda automática de neologismos.

2. Nociones computacionales de neología

La terminología del dominio de neología es bastante estable desde los años ochenta: no todos los neólogos utilizan todas las distinciones, pero las clases de neologismos (formales, semánticos y pragmáticos, préstamos, formaciones libres, etc.) están bien definidas y poco discutidas. Sin embargo, todas estas nociones son conceptos abordados desde la perspectiva lingüística. En una aproximación computacional de la neología, en cambio, hay algunos conceptos relevantes que han sido poco tratados. Esta sección introduce algunas nociones desde esta perspectiva que son relevantes para el resto de esta ponencia.

2.1 Neologismos ortográficos

Se dice comúnmente que las herramientas para la detección automática de neologismos producen listas de neologismos formales. Sin embargo, el resultado de las búsquedas automáticas son entidades de una naturaleza diferente, que podemos denominar *neologismos ortográficos*. No todos los neologismos formales son neologismos ortográficos, ni todos los ortográficos son formales.

El procesamiento automático del lenguaje opera sobre las entidades que se encuentran en los textos: cadenas de caracteres (*strings*). Los candidatos a neologismo verificados en todos los métodos descritos en la sección anterior son por lo tanto *cadenas neológicas*. En el tratamiento manual de los candidatos, estas cadenas están normalmente relacionadas con las entradas léxicas correspondientes.

Un *neologismo ortográfico* es una entidad léxica que en una o más de sus formas flexionadas tiene una cadena neológica. Por ejemplo, el verbo ingles *fax* (mandar por fax) es (o era) un neologismo

ortográfico, pero aunque la forma de citación es homógrafa con el nombre *fax* (telefax), la forma del pasado (*faxed*) es una cadena neológica.

Como los neologismos ortográficos, los neologismos formales son unidades léxicas y no cadenas de caracteres. Sin embargo, no todas las cadenas neológicas son un neologismo formal. Por ejemplo, en el caso de una reforma de la ortografía, las nuevas formas para las palabras no son consideradas palabras nuevas: el hecho de que la palabra *factual* (factual) en portugués (europeo) ahora se deba escribir como *fatual* no cambia la palabra ni su significado – la única cosa que cambia es la ortografía. Entonces, el neologismo ortográfico *fatual* no se considera un neologismo formal.

En la otra dirección, cuando una palabra no es un neologismo ortográfico, en algunas ocasiones puede considerarse un neologismo formal. En principio, todos los homógrafos (nuevos) son neologismos formales, pero la categoría más pertinente es la de los homógrafos de clases gramaticales diferentes. Por ejemplo, el nombre inglés *build* es una palabra nueva para indicar una versión de un software (como en la frase "*the latest build of Microsoft Office"*) y entonces es un neologismo formal. Sin embargo, la forma singular (*build*) tal como la forma plural (*builds*) son homógrafas con formas verbales del verbo *build* (contruir), lo que significa que ninguna de las formas de *build* es una cadena neológica y entonces el nombre *build* no es un neologismo ortográfico.

2.2 Neologismos de producción

El término *neologismo* para indicar palabras nuevas es bastante común, por ejemplo, en los diccionarios. Pero normalmente, no *todas* las palabras adicionadas en el diccionario son nombradas neologismos, sino solamente aquellas palabras que *se sienten* nuevas. Es por eso que Rey (1975) argumentaba que la etiqueta *neologismo* en los diccionarios no tiene un sentido fijo, sino que no es más que el sentimiento subjetivo del editor. Los neologismos de esta concepción se llaman *neologismos psicológicos* (Cabré 2000). Una clase de neologismos relacionada son las palabras para conceptos nuevos, por ejemplo, las palabras para innovaciones tecnológicas como *SMS*, *flexicuridad,* o *motor híbrido*. Se llaman *neologismos conceptuales* y la mayoría de los neologismos conceptuales, si no todos, son neologismos psicológicos también. En los dos casos, neologismos no son sólo palabras nuevas, sino palabras con un nivel elevado de "innovación".

Desde la perspectiva automática, por su parte, los neologismos son sólo cadenas de caracteres que son nuevas o se utilizan de una manera nueva. En los textos digitales, no hay significados ni aspectos psicológicos: para eso, se necesita la interpretación (humana) del texto. Así pues, con la excepción potencial de la detección utilizando patrones (sección 2.2), los métodos automáticos sólo tienen en cuenta la utilización de las palabras y verifican si esta utilización es nueva y, si es nueva, entonces los proponen como candidatos a neologismos. Los neologismos desde este punto de vista (*neologismos de producción)*, no son siempre considerados neológicos por todos los usuarios.

2.3 Ocurrencias neológicas

Tradicionalmente, en la búsqueda de neologismos, los *ocasionalismos* no cuentan como neologismos (Teubert 1998). Ocasionalismos son palabras (nuevas) que no son realmente elementos (nuevos) de la lengua, sino palabras que se utilizan una vez metafóricamente, como broma o ironía, o sólo se utilizan dentro de un contexto muy específico, por un solo autor. Los neologismos, por otro lado, son entidades léxicas nuevas o nuevos significados de la lengua; algunos ya se han integrado en el uso de la lengua y, por lo tanto, tendrían que aparecer en las actualizaciones de los diccionarios de lengua general. Es ésta la noción de "neologismo" que utilizan los lexicógrafos tradicionalmente.

Las búsquedas (semi)-automáticas de neologismos, por su lado, se concentran especialmente en las primeras utilizaciones de palabras nuevas, antes de que se incorporen completamente a la lengua. Para confeccionar la lista de candidatos a neologismo, no importa si son unidades nuevas que se incorporarán al acervo de la lengua o si son ocasionalismos. Para evitar confusión, sería mejor llamar a los productos de la búsqueda automática *ocurrencias neológicas*: ocurrencias de palabras en el corpus de estudio que cumplen los criterios de neologicidad. En este texto, sin embargo, cuando no haya indicaciones contrarias, utilizaremos la palabra "neologismo" ampliamente para referirnos a las ocurrencias neológicas.

3. Métodos computacionales complementarios

Los métodos computacionales descritos en la sección 2 se pueden complementar con diversas técnicas computacionales. Las tres técnicas más relevantes para la detección de neologismos son discutidas en esta sección: el etiquetaje morfosintáctico, la extracción de términos y el filtraje.

3.1 Etiquetaje morfosintáctico

Casi todas las herramientas de procesamiento de lenguaje se inician con el etiquetaje morfosintáctico: la atribución automática de la categoría morfosintáctica a todas las palabras del corpus. Es evidente que el etiquetaje podría ayudar en la búsqueda de neologismos: no solamente daría más información sobre los candidatos, sino también podría dar la posibilidad de detectar neologismos formales que no son cadenas neológicas (ver 3.1).

El etiquetaje automático ha obtenido un nivel muy elevado, y las mejores herramientas tienen una tasa de error de sólo un 2%. Sin embargo, con el tratamiento de los neologismos, los etiquetadores son mucho menos fiables. Esta diferencia se basa en el funcionamiento de los etiquetadores: la primera etapa es verificar cuáles son las categorías posibles para cada palabra del corpus a etiquetar. Solamente para las palabras que pueden tener más de una categoría (como, por ejemplo, *toca* que puede ser una forma verbal o un sustantivo) el etiquetador utiliza el contexto para decidir cuál de las posibilidades es la más probable. Las cadenas neológicas son por definición palabras que no tienen una entrada en el léxico, y mucho menos entradas múltiples. Cuando son confrontados con un neologismo sintáctico, los humanos normalmente deciden a partir del contexto cuál es la categoría de la palabra, pero ninguno de los etiquetadores existentes puede hacer, por el momento, eso. Para una palabra sin entrada léxica, algunos etiquetadores recurren al análisis morfológico, otros atribuyen una categoría por defecto o dejan de atribuir una categoría. Esos resultados del etiquetador son los menos fiables, lo que quiere decir que para los neologismos la tasa de error es muy superior al 2%.

3.2 Extracción de términos

Una restricción considerable de los métodos computacionales descritos en la sección 2 es que en principio solamente funciona con unidades léxicas de una palabra entre blancos (unidades monoléxicas), y no detectan neologismos sintagmáticos que consisten en más de una palabra (unidades poliléxicas), como, por ejemplo, *motor híbrido*. La causa la encontramos en que la primera etapa de los métodos descritos es la *tokenización*: dividir el corpus en unidades monoléxicas, y tratarlos después como unidades primitivas.

Para resolver esta restricción, sería posible extraer todos los términos o expresiones de más de una palabra del corpus, y en vez de la lista de palabras, utilizar la lista de palabras y términos como entrada para los diversos métodos. La extracción de unidades polilexemáticas se realiza de manera estadística: las palabras que se utilizan en conjunto (una al lado de la otra) dentro del corpus con una frecuencia más elevada de la que se explicaría por casualidad son candidatos a unidades polilexemáticas.

La combinación de unidades polilexemáticas y métodos de exclusión es problemática en dos sentidos. Primero, hay muchas más expresiones en una lengua que palabras, y no existe ni es fácil de hacer una lista completa de expresiones en el mismo sentido que un diccionario es una lista completa de palabras. Segundo, muchos neologismos sintagmáticos formales no son cadenas neológicas, no solamente por las razones descritas en la sección 3.1, sino también por el motivo siguiente: cuando las partes de la expresión son suficientemente frecuentes, es muy probable que se encuentren a veces una al lado de la otra sin formar una expresión. Por ejemplo, el neologismo sintagmático inglés *world music* se encuentra a veces como secuencia accidental en textos sin conformar una unidad, como en la frase siguiente:

(3) For most people in the world music has become an indispensable part of their lives.

Con estas dificultades, en principio la única manera de detectar neologismos sintagmáticos es la comparación de la frecuencia entre el corpus de estudio y el corpus de referencia, como se ha escrito en la sección 2.3.

4. Métodos y objetivos

Como ya se ha indicado en la introducción, hoy en día la búsqueda de neologismos tiene intereses nuevos. Esta sección intenta describir brevemente cuál de los métodos automáticos es más adecuado para dos objetivos distintos.

4.1 Palabra del día

Una manera muy común de buscar neologismos es explorar todos los neologismos que surgen en el diario (en línea) de un día concreto. Una edición de un diario contiene un promedio de 50.000 palabras, por lo tanto un tratamiento manual sería muy costoso. Con un método automático de exclusión (2.1), el trabajo disminuye bastante: una edición produce normalmente menos de 500 candidatos. Visto que los resultados son también guardados en una base de datos, las herramientas de exclusión son muy útiles para este tipo de búsqueda, incluso cuando no detecta neologismos gramaticales, semánticos ni sintagmáticos.

Si la edición de un diario tiene un tamaño adecuado para ser tratado con los métodos de exclusión, para un procesamiento estadístico un corpus de estudio de 50.000 palabras es considerado muy pequeño. Los métodos estadísticos se basan en números muy superiores, y un corpus de este tamaño no otorga resultados fiables. Los métodos basados en la frecuencia de las palabras son más apropiados para un corpus de estudio más amplio, como, por ejemplo, un año completo de un diario. Un año de ediciones digitales típicamente tiene más de un millón de palabras que ya es un tamaño más adecuado para el tratamiento estadístico. Teniendo en cuenta que la búsqueda de neologismos semánticos hoy es solamente posible con métodos estadísticos, no puede realizarse periódico a periódico, sino al menos con las ediciones de un periódico correspondientes a un año.

4.2 Productividad léxica

Un objetivo más reciente para la búsqueda de neologismos es el estudio de la productividad relativa de los diversos procesos de formación de palabras. Cuestiones típicas de este tipo de búsqueda son, por ejemplo: ¿cuáles son los sufijos más productivos en una lengua? ¿hay más neologismos morfológicos o más préstamos? ¿cuál es la lengua más productiva entre los préstamos? etc.

Para este objetivo, como ya observó Baayen & Renouf (1996), no es realmente necesario tener un conjunto completo de neologismos, y tampoco es necesario tener un conjunto completamente limpio (pueden quedar candidatos falsos). Lo más importante es que el conjunto de neologismos utilizado para el estudio sea representativo de los neologismos de la lengua. Baayen & Renouf argumentan que la productividad relativa de los sufijos productivos entre los *hapax legomena* es la misma que la productividad entre los neologismos.

Sin embargo, no hay una garantía de que la proporción de los neologismos morfológicos y los préstamos sea la misma entre los *hápax* que entre los neologismos que no lo son. Por ejemplo, si sabemos que los préstamos tienden a establecerse más rápidamente en la lengua que los neologismos morfológicos, podríamos hipotetizar que es probable que haya relativamente pocos préstamos entre los *hápax* en comparación con los neologismos morfológicos. Así pues, sería más fiable estudiar la productividad relativa entre un conjunto fijo de neologismos sistemáticamente detectados. La manera más eficaz para establecer un tal conjunto es con las herramientas basadas en la exclusión.

5. Conclusión

Este artículo muestra que hay tres métodos fundamentales para la detección automática de neologismos, y que hay varias razones para que la utilización de una lista de exclusión sea desde siempre el más utilizado: no solamente porque es el más fácil de implementar, sino también porque es una manera transparente y funcional para hacer búsquedas de neologismos, que lleva un conjunto restringido y bien definido de candidatos a neologismo. Sin embargo, es un método que no permite detectar neologismos gramaticales, sintagmáticos y semánticos. Para detectar estos tipos de neologismos se tiene que utilizar métodos estadísticos.

También se muestra en este artículo que los resultados de la búsqueda automática de neologismos

no son las palabras típicamente indicadas como neologismos, sino *ocurrencias neológicas* y *neologismos ortográficos de producción*.

Uno de los temas fundamentales para la búsqueda automática de neologismos se deja sin discusión en este trabajo: las dos cuestiones importantes para cada tipo de herramienta de detección (de neologismos) (semi) automática son: cuántos de los neologismos en el corpus de estudio son recuperados por la herramienta (*cobertura*), y cuántos de los candidatos recuperados son de hecho neologismos (*precisión/ruido*). Al fin y al cabo, la utilidad de las herramientas basadas en patrones y métodos estadísticos para la detección de neologismos, especialmente para la detección de neologismos gramaticales, sintagmáticos y semánticos, depende de su cobertura y de su precisión.

Referencias bibliográficas

Baayen, Harald & Antoinette Renouf. 1996. Chronicling the Times: Productive lexical innovations in an English newspaper. *Language*, vol. 72: pp. 69-96.

Cabré, Teresa. 1992. *La Terminología: Teoría, metodología, aplicaciones*. Barcelona: Editorial Antártida.

Cabré, Teresa. & Estopà, Rosa 2009. "Trabajar en neología con un entorno integrado en línea: la estación de trabajo OBNEO*". Revista de Investigación Lingüística,* 12, pp. 17-38.

Chlebda, W. 1991. *Elementy frazematyki. Wprowadzenie do frazeologii nadawcy*. Opole: WSP.

Issac, Fabrice. *Telanaute : un outil de veille lexicale. En: Actas del I Congreso Internacional de Neología en las lenguas románicas*, Barcelona.

Janicijevic, T & Walker, D. 1997. *NeoloSearch: Automatic Detection of Neologisms in French Internet Documents*. ACH-ALLC'97, Kingston, Canada

Janssen, Maarten. 2005a. *NeoTrack: semi-automatic neologism detection*. Papel presentado en la conferencia XXI APL, Lisbon.

Janssen, Maarten. 2005b. *Open Source Lexical Information Network*. Third International Workshop on Generative Approaches to the Lexicon, Geneva, Switzerland.

Knowles, E. & Elliot, J. 1997. *The Oxford Dictionary of New Words.* Oxford: Oxford University Press.

Nazar, Rogelio & Vidal, Vanessa. 2008. "Aproximación cuantitativa a la neología" *En: Actas del I Congreso Internacional de Neología en las lenguas románicas*, Barcelona.

Paryzek, P. 2008. "Comparison of selected methods for the retrieval of neologisms". *Investigationes Linguisicae*, vol. XVI: pp. 163-181.

Renouf, Antoinette. 1993. "A Word in Time: first findings from dynamic corpus investigation" *In:* Aarts, de Haan and Oostdijk (eds.) *English Language Corpora: Design, Analysis and Exploitation*. Amsterdam: Rodopi.

Renouf, Antoinette & Baayen, R. Harald. 1998. "Aviating among the Hapax Legomena: morphological grammaticalisation in current British Newspaper English". *En:* Renouf (ed.) *Explorations in Corpus Linguistics*. Amsterdam: Rodopi.

Roche, Sorcha & Lynne Bowker. 1999. Cenit: Système de Détection Semi-Automatique des Néologismes. *Terminologies Nouvelles,* vol. 20.

Teubert, W. 1998. "Korpus und Neologie." *En:* Teubert (ed.) *Neologie und Korpus. Beitrage zur Neologismenlexikographie*. Tübingen: Gunter Narr Verlag.

Vivaldi, Jordi. 2000. SEXTAN: prototip d'un sistema d'extracció de neologisms. *In*: Cabré, Freixa & Solé (eds.) *La Neologia en el tombant de segle.* Barcelona: IULA. pp. 165-177.