ANÁLISIS LINGÜÍSTICO DE LA DETECCIÓN AUTOMÁTICA DE NEOLOGISMOS LÉXICOS

Judit Freixa y Elisabet Solé
Universitat Pompeu Fabra

Resumen

Los profesionales de la traducción deben dar respuesta a las necesidades neológicas que se les presentan en los textos meta. Las bases de datos neológicas construidas a partir de la detección automática de palabras nuevas en corpus textuales (especializados o no) se convierten en uno de los recursos más eficaces que la vertiente aplicada de la neología léxica ofrece a los traductores.

El análisis lingüístico de los resultados obtenidos a partir de sistemas de detección automática (en el marco del Observatori de Neologia de la Universitat Pompeu Fabra) permite examinar sus ventajas frente a la detección manual, e intentar plantear soluciones para solventar sus limitaciones.

Palabras clave: neología, neologismo, detección automática, trabajo de traducción.

Abstract

Translators must give answer to the neology needs of the goal texts. Neology databases built thanks to the automatic detection of new words in textual corpora (specialized or not) are now one of the most effective resources that the applied side of lexical neology offers to translators.

The linguistic analysis of the results obtained from systems of automatic detection (within the framework of the Observatori de Neologia at the Universitat Pompeu Fabra) allows examining their advantages over manual detection, and tries to give solutions to define their limitations.

Keywords: neology, neologism, automatic detection, translation work.

1. Introducción

A menudo los traductores deben dar respuesta a necesidades léxicas neológicas, especialmente denominativas pero también expresivas, que les plantean los textos. La neología léxica, disciplina de la lingüística aplicada que estudia las palabras nuevas, ofrece a la traducción aspectos teóricos y aplicados que facilitan la resolución de dichas necesidades (Freixa et al. 1998).[1]

1. Coherentemente, en los planes de estudios de Traducción e Interpretación se enseña a los futuros traductores dicha materia de forma más o menos explícita (o encubierta) en las asignaturas de terminología o afines.

La detección de los neologismos léxicos en corpus textuales y su compilación en bancos de datos neológicos, en diccionarios electrónicos, en las memorias de traducción, son aplicaciones de los estudios neológicos que se convierten en recursos disponibles para los traductores, junto a los estudios más propiamente lingüísticos de cuáles son los mecanismos (con sus condiciones y restricciones, su productividad, etc.) que permiten la formación de nuevos vocablos en las lenguas. Si, además, a ello le sumamos la posibilidad de abordar los neologismos en corpus textuales en más de una lengua (aunque no sean necesariamente paralelos, pero sí textualmente similares), los beneficios para la traducción de las aplicaciones resultantes es aún mayor.

La detección automática de neologismos léxicos, aplicada de forma sistemática a corpus textuales, es una realidad aún reciente en el ámbito de los estudios neológicos (Cabré et al. 1995, 2004); no obstante, ha proporcionado ya buenos resultados en determinadas aplicaciones. En este artículo nos proponemos llevar a cabo un análisis de las ventajas y limitaciones de la actual detección automática de neologismos léxicos mediante criterios lexicográficos, y apuntar algunas estrategias para superar las limitaciones actuales de este proceso.

Para ello vamos a partir de la experiencia acumulada en el Observatori de Neologia[2] (OBNEO) del Institut Universitari de Lingüística Aplicada de la Universitat Pompeu Fabra, que trabaja en la detección de neologismos léxicos desde 1988 y que desde el año 1998 empezó a automatizar dicha detección.

En este grupo de investigación, los textos de base para la detección de neologismos son periódicos de amplia difusión (de temática general) y, desde el año 2000, también textos orales procedentes de los medios de comunicación y textos escritos no especializados de publicaciones periódicas que no pasan por servicios de revisión lingüística, por tanto, más vinculados a la neología espontánea.[3]

2. Concepto de neologismo

Desde sus inicios, el Observatori de Neologia ha considerado como uno de sus objetivos prioritarios, aunque no exclusivo, la actualización de las obras lexicográficas de referencia, tanto para el catalán como para el castellano, y la elaboración de compendios neológicos de distinta índole. En concordancia con estos fines, los investigadores del Observatori parten de una concepción lexicográfica del concepto de neologismo, y no de una concepción exclusivamente temporal o psicolingüística.

Consecuentemente, se considera *neologismo* toda unidad léxica documentada en un *corpus textual de vaciado* que no consta, o bien aparece con distinta forma

2. Para mayor información sobre el Observatori de Neologia, consultar la página http://iula.upf.edu/obneo.

3. Puede verse una muestra de los neologismos del Banco de Datos del Observatori (BOBNEO) en http://bobneo/iula.upf.edu.

o sentido, en un *corpus lexicográfico de exclusión* determinado. Así pues, se ha descartado la posibilidad que el corpus de exclusión sea otro corpus textual y se ha definido como corpus de contraste un conjunto de obras lexicográficas de referencia en las distintas lenguas de trabajo, formado tanto por diccionarios descriptivos como por diccionarios prescriptivos o normativos de la lengua general.

3. Detección y extracción de neologismos

El proceso de detección manual de neologismos sigue la cadena siguiente: el neólogo lee el texto de vaciado y marca aquellas unidades que considera que podrían no encontrarse en las fuentes lexicográficas de referencia. Tras la comprobación en estas fuentes, el neólogo descarta aquellas unidades ya contenidas y elabora una ficha para cada una de las unidades nuevas no documentadas. Esta ficha, que automáticamente se convierte en un registro de una base de datos, contiene algunas informaciones que provienen del texto del que ha sido extraído el neologismo (forma, categoría gramatical, contexto, aspectos tipográficos, fecha y fuente del texto, etc.) y otras que no provienen de la fuente de vaciado sino del análisis del neólogo (tipo de neologismo, etc.).

Desde los inicios del proyecto se trabajó en la idea de automatizar esta cadena de trabajo y finalmente se creó una herramienta informática de extracción automática denominada Sextan (Vivaldi 2000; Observatori de Neologia 2004b). Con esta herramienta, los estudios sobre neología léxica se enriquecen gracias a la introducción de las nuevas tecnologías en el tratamiento del lenguaje.

El proceso de detección automática es paralelo al proceso que hemos visto para el vaciado manual. No obstante, para llevarlo a cabo es necesario que el corpus textual a partir del que se hará el vaciado se encuentre en soporte electrónico y haya sido procesado. Es necesario también disponer de un diccionario o diccionarios en soporte electrónico que actúen como filtro para presentar los candidatos a neologismos y, finalmente, disponer de una interfaz que permita la validación de los candidatos a neologismos por parte del neólogo.

Con todo ello se consigue que cualquier unidad léxica no documentada en los diccionarios que actúan de filtro sea considerada un candidato a neologismo. No obstante, es el neólogo quien debe decidir si se trata realmente de un verdadero neologismo, ya que el programa de detección ofrece cierta cantidad de ruido, es decir, de falsos neologismos.

Este ruido está relacionado con varios tipos de factores, entre los que cabe destacar los siguientes:

a) En primer lugar, el sistema ofrece como candidatos a neologismos las unidades que aparecen en las citas en otras lenguas que a veces encontramos en textos de prensa. Aunque sí interesa recopilar los préstamos de otras

lenguas usadas en la lengua de trabajo, no hay que confundir estos neologismos con las unidades que aparecen usadas directamente en la lengua de origen, dentro de una cita.

b) El sistema también ofrece como candidatos a neologismos las unidades que no reconoce debido a algún error en las fuentes de vaciado; los errores más comunes son de acentuación, pero otras veces el error se produce porque faltan o sobran letras o sílabas, o porque falta un espacio y en consecuencia dos palabras aparecen unidas como si se tratara de una nueva unidad. Son errores que frecuentemente la detección manual no percibe, pero que el programa de extracción detecta en todos los casos:

A pesar del resultado de los **analisis**, las familias de los dos jóvenes insistieron ayer en que sus hijos 'fueron engañados'. [EP 02/03/21][4]
Un amor **frustado** de juventud, algún estreno desdichado, como Delirio del amor hostil. [EP 02/03/21]
Enumeró el aislamiento de España respecto a la cultura internacional; la contradicción entre el carácter conservador, religioso y afín a la aristocracia del arquitecto frente a las **corrrientes** bohemias y laicas de las vanguardias, y también el hecho de que se necesitaba esta corriente moderna expresionista para revindicar su figura. [EP 02/03/21]
Todos los grupos, menos el PP, expresaron su 'indignación' por los 'insultos intolerables' proferidos por el periodista de la televisión **públicacontra** la diputada socialista Leyre Pajín, de quien López Castillo dijo en Guadalajara Dosmil: 'Exhibe algunos atributos muy respetables, pero insuficientes para compensar su cacumen o inteligencia'. [EP 02/03/21]
Somos muchos cargos públicos, concejales de pequeños **municios**. [EP 02/03/21]
El príncipe heredero Abdalá bin Abdelaziz, de Arabia Saudí, presentó su propuesta a Israel de **reconomiento** de los países árabes a cambio de su retirada de los territorios ocupados en 1967. [EP 02/03/21]

Tabla 1. Falsos neologismos: errores de los textos de vaciado.

c) En otros casos, el ruido proviene de errores en el lemario o en el procesamiento del texto; así, por ejemplo, el sistema descarta la mayoría de nombres propios dentro de texto, porque los identifica como tales, pero los presenta como candidatos a neologismos a inicio de frase, ya que la mayúscula inicial le resulta ambigua, o cuando la letra mayúscula no sirve de guía:

4. En los ejemplos, LV se refiere al periódico *La Vanguardia*, y EP al periódico *El País*.

Alejandra Steffani, de 47 años, es cuadripléjica, apenas puede mover el 30% de su mano izquierda y está postrada en la cama permanentemente. [EP 02/03/21]
La obra Misterium **anatomicum** concentra todo el sentido de la muestra, lo femenino y lo masculino, eros y tanatos. [EP 02/03/21]

Tabla 2. Falsos neologismos debidos a errores del procesamiento del texto.

Además de descartar estos falsos neologismos, la validación del neólogo en este punto del proceso de detección puede también descartar candidatos que no van a ser considerados neologismos por la existencia de criterios restrictivos impuestos al criterio de detección de partida, de acuerdo con las finalidades y objetivos de cada trabajo. Por ejemplo: pueden no interesar unidades muy predictibles con un bajo interés neológico, como adverbios en *–mente*; superlativos, diminutivos, aumentativos, etc. de palabras ya documentadas en el corpus de exclusión siempre que no estén lexicalizados; neologismos que contengan el prefijo *ex-* absolutamente predictible; etc. En la práctica, estas unidades que la herramienta de detección nos ofrece como candidatos a neologismos son verdaderos neologismos desde el punto de vista del criterio lexicográfico, aunque sean neologismos descartables por su poco interés o por las razones concretas de cada proyecto y, por lo tanto, el neólogo puede desestimarlos por criterios de exclusión complementarios.

4. Ventajas y limitaciones del vaciado automático

La cadena automatizada de vaciado ofrece ventajas evidentes sobre el vaciado manual, aunque también limitaciones importantes que no ofrece el vaciado manual. Vamos a ocuparnos en primer lugar de las ventajas, y a continuación de las limitaciones.

4.1. Ventajas del vaciado automático sobre el vaciado manual

Como es obvio, la *velocidad* es la primera ventaja del vaciado automático frente al vaciado manual: no solamente porque se obtiene un mayor número de candidatos en menor tiempo, sino también porque las fichas de vaciado aparecen ya prácticamente elaboradas puesto que el sistema recupera el contexto, marca el neologismo dentro del contexto y cumplimenta la mayor parte de los campos de información.

El proceso automatizado resulta también más exhaustivo; esta *exhaustividad* se debe, por una parte, al hecho de que el número de textos susceptibles de ser vaciados es mayor (lo cual provoca una mayor relevancia y fiabilidad de los resultados) y, por otra, al hecho de que el sistema detecta sin error los neologismos poco llamativos (neologismos muy predictibles y formados a partir de reglas muy productivas) que, a veces, el neólogo no advierte cuando realiza el vaciado manual.

Por otro lado, el vaciado automático presenta mayor *objetividad*, ya que permite realizar una fotografía de los mecanismos de vitalidad lingüística de la lengua más cercana a la realidad, aún cuando siga siendo una fotografía parcial condicionada por el corpus textual de partida; y también presenta mayor *sistematicidad*, porque el vaciado automático conlleva la posibilidad de estudiar la frecuencia de uso de las unidades.

La *minimización de la introducción de errores* por parte del neólogo es también una ventaja importante, puesto que su intervención en la implementación de la información es menor.

Y, finalmente, no podemos dejar de mencionar que el vaciado automático permite, frente al vaciado manual, una mayor facilidad de *reutilización* de los resultados en la cadena de procesamiento de corpus textuales, puesto que puede, por ejemplo y si se considera conveniente, alimentar los filtros lexicográficos de exclusión.

4.2. Limitaciones del vaciado automático

En la actualidad, la herramienta de detección automática de neologismos puede filtrar y proponer como neológicas aquellas unidades léxicas monolexemáticas (entre blanco y blanco) no contenidas en los diccionarios que actúan de corpus de exclusión; es decir, identifica como candidatos solamente los neologismos monolexemáticos que no coinciden con formas idénticas del corpus de exclusión, ya estén formados por procesos formales de creación de palabras o se incorporen como préstamos de otras lenguas.

El sistema, pues, no identifica como candidatos el resto de neologismos:

a) los neologismos polilexemáticos, es decir, formados por más de una palabra, como los neologismos formados por sintagmación y/o composición que presenten un blanco entre sus constituyentes;
b) los neologismos semánticos, es decir, creados por la atribución de un significado nuevo a palabras ya contenidas en los diccionarios, puesto que el lemario sólo está etiquetado morfológicamente y no contiene información semántica;
c) los neologismos sintácticos, es decir, los usos sintácticos nuevos de formas contenidas en el diccionario o lemario de exclusión.

5. Evaluación del vaciado automático según el proceso de formación del neologismo

Una vez descritas las ventajas y limitaciones del vaciado automático frente al vaciado manual, evaluemos sumariamente los resultados que nos ofrece la detección

automática de los neologismos léxicos según el proceso de formación que los ha generado. Veremos como, mientras para algunos de ellos la detección automática ofrece resultados muy destacables, para otros son menores o nulos.

5.1. Neologismos formados por derivación y composición culta

La detección automática de este tipo de neologismos tiene un elevado grado de rentabilidad, sistematicidad y exhaustividad, puesto que al tratarse de unidades monolexemáticas el sistema de detección automático las reconoce como candidatos a neologismos sin ningún tipo de problema.

Además, y puesto que tanto la prefijación y la sufijación como la formación culta a menudo proporcionan neologismos de lengua muy predictibles, la ventaja frente al vaciado manual es considerable porque la detección tiene lugar con total objetividad, dado que no interviene la idea subjetiva de novedad por parte del neólogo.

La canción que quizá me guste más del disco es "Noche de ronda", porque aunque es muy conocida, Babu Silvetti, mi *arreglista* habitual, ha hecho una versión muy especial. [LV 13/09/98]
Es consciente el autor premiado de que la creciente *mercantilización* de la literatura ha llevado a establecer la vida de un escritor en la duración de sus libros en la mesa de novedades de las tiendas. [EP 03/03/99]
Pero de repente, la originalidad de los partidos verdes no es ya hablar del medio ambiente o de la *bio-energía*, sino preparar un modelo alternativo de desarrollo, basado en las energías blandas y el crecimiento lento. [EP 18/03/93]

Tabla 3. Neologismos detectados formados por derivación y composición culta.

Tan sólo presentan problemas para la detección automática aquellos casos de neologismos por prefijación o composición culta en los que el prefijo o formante culto no está adjuntado directamente a la base: en este caso el sistema no detectará dichas unidades y se convertirán en silencio. Se trata, sin embargo, de un porcentaje muy bajo en este tipo de neologismos formales. Por ejemplo:

Dejemos que los israelíes partidarios de Paz Ahora, los grupos ortodoxos *pro paz*, el movimiento religioso blando se presenten voluntarios para turnarse en habitar el barrio judío de Hebrón y para dirigir las oraciones de los judíos. [EP 17/01/97]
En todo caso, Javier Solana no era de los que más se destacaban como *anti OTAN*. [LV 09/12/95]

Tabla 4. Neologismos silenciados formados por prefijación.

5.2. Neologismos formados por composición y sintagmación

A diferencia de los anteriores, los neologismos formados por composición y sintagmación presentan mayores problemas para la detección automática puesto que acostumbran a ser unidades polilexemáticas. No obstante, los neologismos por composición aglutinada, como los siguientes, sí que se detectan:

Dos policías muertos por un *camión-bomba.* [LV 19/11/90]
El Teatro del Arte madrileño con cómicos que simultanean danza, canciones y animación callejera, y otras actividades dispersas como *cuentacuentos* y un largo etcétera. [EP 14/08/98]

Tabla 5. Neologismos detectados formados por composición.

En cambio, los neologismos constituidos por varios lexemas que no están formalmente aglutinados, sino que presentan un espacio en blanco entre sus componentes, no pueden detectarse automáticamente:

Unos 15 detenidos en Gaza, relacionados con el *camión bomba.* [LV 23/03/95]
Si ingreso en una *cuenta vivienda* una cantidad que supera el 30% de la base imponible, ¿la diferencia entre lo ingresado y el 30% lo puedo desgravar el año que viene al comprar la vivienda? [EP 14/12/97]
Está hablando de las *autopistas de la información*, claro, un concepto tan de moda como indefinido. [EP 15/05/94]

Tabla 6. Neologismos silenciados formados por sintagmación.

A pesar de todo, si alguno de los lexemas que intervienen en el compuesto no aglutinado o en el neologismo por sintagmación es a su vez neológico, su detección automática será posible y el neólogo podrá recuperar la unidad completa a partir de la información contenida en el contexto seleccionado. Así, por ejemplo, en el siguiente contexto la herramienta detectará el neologismo *insertor*, y el neólogo recopilará el sintagma *insertor laboral*.

En la Roca del Vallès las vacantes son estas: una de administrativo, dos de inspector de policía, una de ingeniero técnico, una de asistente social, una de *insertor laboral*, una de educador social, una de trabajadora [...]. [LV 25/06/00]

Tabla 7. Neologismos silenciados recuperables.

5.3. Neologismos formados por truncación

Los neologismos formados por cualquiera de los procesos de truncación (abreviación, siglación o acronimia) se detectan automáticamente de forma sistemática si no coinciden formalmente con ningún lema del diccionario.

Los belgas son conscientes de que aún quedan muchos problemas por resolver: la adaptación de los ciudadanos para evitar posibles traumas a partir de la fecha señalada y la adaptación de las contabilidades de las pequeñas y medianas empresas (*pymes*). [LV 03/07/01]
Y como parece ser que los tiempos no están para romanticismos, pues ahí tenemos la *prota* repartiendo mamparas y haciendo saltar paredes con explosivos. [EP 21/07/96]
En primero y segundo curso disponemos de asignaturas específicas de ofimática y *turismática*, en conexión con Internet, y la aplicación de los programas Fidelo y Nuevo de gestión informática hotelera y de agencias. [LV 20/03/96]

Tabla 8. Neologismos detectados formados por truncación.

En cambio, no pueden detectarse aquellos neologismos que una vez truncados coinciden formalmente con alguna entrada del diccionario, como en los ejemplos siguientes, en los que coincide con un adjetivo y con un formante culto:

Las *eléctricas* siguen asumiendo su papel de dinamizadoras del mercado bursátil. [LV 19/01/94]
Los guerrilleros pretendían un golpe propagandístico contra los *narcos*, pero les han salido las bombas por la culata. [EP 14/10/90]

Tabla 9. Neologismos silenciados formados por truncación.

5.4. Neologismos sintácticos

Los sistemas actuales no pueden detectar automáticamente los usos sintácticos nuevos de palabras o formas contenidas en el diccionario que implican cambios de categorías gramaticales mayores (nombre, adjetivo, verbo y adverbio), ya se trate, como ilustran los ejemplos siguientes, de adjetivos que se nominalizan, de nombres que se adjetivizan, o de formas verbales no personales (infinitivos, participios de presente, participios de pasado) que se nominalizan. Ni tampoco es posible, a partir de la herramienta disponible, la detección de los cambios de subcategorización; por ejemplo, usos transitivos de verbos intransitivos, usos intransitivos de verbos transitivos, etc.

Telefónica, los constructores y los *inmobiliarios*, con su excelente posición cogieron el relevo alcista. [LV 24/02/98]
Le encanta a este Barça hurtar la *esférica*, tocar a los flancos, profundizar hasta la línea para servir el pase de la muerte y ajusticiar con el ariete o la llegada de los volantes. [EP 06/11/95]
El papa pide a Europa *interrogarse* sobre sus responsabilidades en la tragedia de Sarajevo. [EP 14/04/97]
Luego Java, en Indonesia, donde el Merapi derrama lava desde el 8 de julio y el Papandayan lleva meses *tosiendo* gas y fango. [EP 14/08/98]
Iván de la Peña *se reivindica*. El jugador cántabro cree que puede jugar en cualquier sistema. [LV 15/02/97]
Media Valencia *se enlata* conmigo todas las mañanas a las ocho y a las tres en el tranvía para ir a la rutina. [EP 09/06/97]

Tabla 10. Neologismos sintácticos silenciados.

5.5. Neologismos semánticos

Actualmente, el proceso de detección automática de neologismos no permite distinguir las palabras formadas por mecanismos semánticos, puesto que el contraste se hace en relación a una forma sin etiquetaje semántico. Así, neologismos como los siguientes sólo pueden detectarse manualmente:

La mayoría de estos establecimientos ofrecen especialidades como el *biberón* (café con leche condensada), que hace décadas que se consume en los bares de nuestros mercados... [LV 17/01/97]
A cambio de mantener esa *horquilla* en el IRPF, el Gobierno del PP se comprometía a ello. [EP 17/05/97]
El instructor resolvió en tal sentido al estimar razonables las imputaciones que las acusaciones atribuyen a Polanco en relación con un supuesto *maquillaje* contable. [LV 23/07/97]

Tabla 11. Neologismos semánticos silenciados.

No obstante, dentro de la diversidad de mecanismos semánticos que crean nuevas unidades, pueden detectarse automáticamente algunos casos porque son formas no contenidas en el lemario:

a) nombres propios utilizados como nombres comunes, es decir, ya sin la mayúscula inicial, y que no coinciden con ninguna otra forma del lemario;
b) neologismos semánticos de unidades léxicas que son ellas mismas candidatos a neologismos porque no coinciden con ninguna entrada del lemario.

Los ejemplos siguientes se detectarían por la primera o por la segunda razón:

El técnico, con una aguda afonía y jugando con dos *chupa-chups* en la mano, respondió: "Si había un día en el que podíamos perder la motivación después de las victorias de nuestros rivales, era hoy (por ayer). [EP 14/04/97]
El nudo de la Trinidad, por ejemplo, no será un *scalextric* y contará con una amplia plaza central. [LV 16/03/89]
Juan Miguel Quintana era piloto de rallies; ahora sólo conduce su bólido sobre la pista de un *scalextric*. [EP 22/04/94]

Tabla 12. Neologismos semánticos detectados.

5.6. Préstamos

En cuanto a los préstamos, cualquier préstamo monolexemático no contenido en el lemario que funciona de filtro de exclusión será detectado como candidato a neologismo sin ningún tipo de problema. Del mismo modo, también se detectarán

los préstamos polilexemáticos, aunque la herramienta solamente detecte una parte del neologismo y sea el neólogo quien deba delimitar la unidad correcta.

Así, en los ejemplos siguientes vemos como el neólogo ha recopilado respectivamente *kale borroka* y *feng shui*, aunque el programa no había detectado estas dos unidades pero sí las cuatro unidades: *kale*, *borroka*, *feng* y *shui*.

Reinicke fue el concejal de HB que condenó, el 18 de agosto, por vez primera en nombre de la coalición, una acción de la *kale borroka* (violencia callejera). [EP 21/09/98]
Zhang Hongyán también pretende protegerse de las calamidades recurriendo a las reglas de la geomancia tradicional, o *feng shui* (literalmente viento y agua). [EP 01/09/99]

Tabla 13. Préstamos neológicos detectados.

6. Conclusiones

Si hacemos una evaluación cuantitativa global del total de neologismos detectados en las fuentes de vaciado tratadas (ya sea mediante la detección automática, ya sea mediante la manual), podemos afirmar que actualmente en el Observatori de Neologia se detecta automáticamente alrededor del 85 % de los neologismos documentados en el corpus de vaciado. Pero es obvio que la piedra de toque no radica en la cantidad de nuevas palabras detectadas, sino en cómo se distribuye esa cantidad según el proceso de creación que ha generado el neologismo.

Como hemos podido observar en el apartado anterior, la detección automática actual de neologismos mediante criterios lexicográficos da como resultado un dibujo sesgado de la realidad neológica, puesto que, a pesar de ofrecer muy buenos resultados para determinados procesos de formación de neologismos, para otros las limitaciones son mayores que las ventajas y los resultados son o bien porcentajes muy bajos o incluso resultados nulos.

Así pues, dadas las limitaciones actuales de la detección automática hay que elegir entre ocuparse solamente de la neología formal monolexemática con muy buenos resultados o bien complementar la extracción automática con la extracción manual por parte de neólogos. Y mientras, indagar aquellas posibilidades que podrían representar una mejora en la detección automática de neologismos.

Creemos que pueden identificarse mecanismos específicos para cada proceso de formación que han de permitir un refinamiento de la detección automática. Si dejamos a un lado aquellos procesos en los que la detección automática ya puede considerarse óptima (derivación, composición culta, truncación y préstamo), habría que buscar mecanismos específicos para la detección de los neologismos polilexemáticos (sintagmación y composición patrimonial no aglutinada), de los neologismos sintácticos y de los semánticos.

Sin ninguna duda, para la detección de neologismos polilexemáticos podrían adaptarse las estrategias desarrolladas para la detección automática de terminología

(Estopà et al. 1998). Estas estrategias pueden ser de base lingüística, basadas en el análisis de patrones morfosintácticos —formantes y patrones estructurales—, léxicos especializados de referencia, etc.; de base estadística, teniendo en cuenta la frecuencia de aparición de determinadas coocurrencias; o de base mixta, lingüística y estadística.

La detección de neologismos semánticos y sintácticos es la que, en estos momentos, se encuentra en un estadio menos avanzado, pero a medida que se vaya desarrollando el marcaje morfológico, sintáctico y semántico de los corpus textuales irá en aumento la posibilidad de introducir mejoras en la detección automática de estos tipos de neologismos.

La idea de complementar el criterio lexicográfico para la detección de unidades con pistas del texto, de carácter lingüístico, metalingüístico o tipográfico, que puedan relacionarse con la idea de *novedad* podría suponer también un avance que solamente pretendemos apuntar.

No queremos cerrar esta aproximación a la detección automática de neologismos léxicos sin recordar que, a pesar de sus limitaciones actuales, las ventajas del vaciado automático frente al vaciado manual son muy relevantes, por todo lo expuesto anteriormente, pero especialmente porque el 75 % de los neologismos de lengua general son mayoritariamente neologismos léxicos formales monolexemáticos, es decir, aquellos que las herramientas de detección automática detectan con un éxito muy elevado.

Bibliografía

Cabré, M. Teresa y Lluís DeYzaguirre (1995). Stratégie pour la détection semiautomatique des néologismes de presse. *Technolectes et dictionnaires* 2, 89-100. [También en M. Teresa Cabré, Judit Freixa y Elisabet Solé (eds.) (2002), 107-114.]

Cabré, M. Teresa, Judit Freixa y Elisabet Solé (eds.) (2002). *Lèxic i neologia*. Barcelona: Observatori de Neologia, Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra. (Sèrie Monografies, 5).

Cabré, M. Teresa, Meritxell Domènech, Rosa Estopà, Judit Freixa y Elisabet Solé (2004). La lexicografia i la identificació automatitzada de neologia lèxica. En *De Lexicografia. Actes del I Symposium Internacional de Lexicografia (Barcelona: 16-18 de maig de 2002)*. Paz Battaner y Janet DeCesaris (eds.), 287-294. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra. (Sèrie Activitats, 15).

Estopà, Rosa, Jorge Vivaldi y M. Teresa Cabré (1998). *Sistemes d'extracció automàtica de (candidats a) termes: estat de la qüestió*. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra. (Papers de l'IULA. Sèrie Informes, 2).

Freixa, Judit, Elisabet Solé y M. Teresa Cabré (1998). Interès de la neologia en el marc de la traducció: l'Observatori de Neologia. En *Actes del III Congrés Internacional*

sobre Traducció. 28-30 de març de 1996. Pilar Orero (ed.), 637-648. Bellaterra: Departament de Traducció, Facultat de Traducció i Interpretació, Universitat Autònoma de Barcelona.

Observatori de Neologia (2004a). *Llengua catalana i neologia*. Coordinación de Judit Freixa y Elisabet Solé. Barcelona: Editorial Meteora. (Cronos, 2).

Observatori de Neologia (2004b). *Metodología del trabajo en neología: criterios, materiales y procesos*. Barcelona: Observatori de Neologia, Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra. (Papers de l'IULA. Sèrie Monografies, 9). <http://www.iula.upf.es/04mon009.htm>

Vivaldi, Jorge (2000). Sextan: Prototip d'un sistema d'extracció de neologismes. En *La neologia en el tombant de segle*. M. Teresa Cabré, Judit Freixa y Elisabet Solé (eds.), 165-173. Barcelona: Observatori de Neologia, Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra. (Sèrie Monografies, 5).