instituto de ingeniería del conocimiento

OBSERVATORIO de la lengua española y las culturas hispánicas en los Estados Unidos

BÚSQUEDA DE ANGLICISMOS EN EL ESPAÑOL ESTADOUNIDENSE A TRAVÉS DE TWITTER

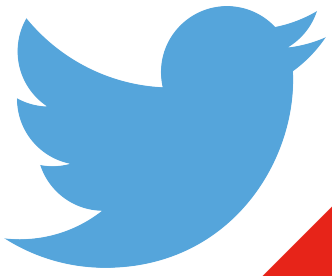
Antonio Moreno Sandoval
Instituto de Ingeniería del Conocimiento – Universidad Autónoma de Madrid

Francisco Moreno Fernández
Instituto Cervantes en la Universidad de Harvard – Universidad de Alcalá

instituto de ingeniería del conocimiento

OBSERVATORIO de la lengua española y las culturas hispánicas en los Estados Unidos

BÚSQUEDA DE ANGLICISMOS EN EL ESPAÑOL ESTADOUNIDENSE A TRAVÉS DE TWITTER



El proyecto “Búsqueda de anglicismos en el español estadounidense a través de Twitter” es una iniciativa de investigación conjunta, entre el Observatorio de la lengua española y las culturas hispánicas en los Estados Unidos del Instituto Cervantes en la Universidad de Harvard y el Instituto de Ingeniería del Conocimiento (IIC), que combina el conocimiento experto de los lexicógrafos y de los lingüistas computacionales en una nueva metodología para descubrir préstamos empleados en las redes sociales.

Antonio Moreno Sandoval
Instituto de Ingeniería del Conocimiento – Universidad Autónoma de Madrid

Francisco Moreno Fernández
Instituto Cervantes en la Universidad de Harvard – Universidad de Alcalá

El origen de esta investigación está en la redacción de un *Diccionario de anglicismos del español estadounidense* (DAEE). La participación del IIC se concreta en la aportación metodológica a un problema lexicográfico bien conocido: ¿cómo identificar neologismos y préstamos que se están produciendo en una comunidad lingüística concreta donde se vive con intensidad el contacto entre dos lenguas? El método habitual es recoger datos de fuentes escritas (periódicos, documentos oficiales, etc.), lo que contribuye a obtener una información "estable" que facilite su análisis. Sin embargo, la escritura convencional no refleja el proceso de préstamo en toda su complejidad, por lo que resulta imprescindible atender a los datos de la lengua hablada.

La observación de la lengua hablada permite registrar el "cambio lingüístico en marcha", es decir, las innovaciones y modificaciones léxicas que se están produciendo en la actualidad en los hablantes de español estadounidense. Como es sabido, estos procesos de cambio en gran medida son una pugna entre diversas alternativas innovadoras, de entre las cuales, finalmente, se impondrá una variante (la palabra o préstamo) por ser más aceptada que las otras. Asimismo, la difusión social de los préstamos puede producirse tanto desde arriba (de forma consciente y desde grupos sociales de prestigio) como desde abajo (desde grupos sociales populares). En este último caso, el cambio se desarrolla primero en el habla espontánea, en los niveles más informales, y se asocia inconscientemente a una estrategia de identidad colectiva (en este caso, la comunidad hispana). El grado de consolidación de un préstamo dependerá del alcance de su difusión dentro de la comunidad.

La observación del habla espontánea, sin embargo, no es fácil de realizar, por las dimensiones sociales y geográficas de la población hispana de los Estados Unidos y porque la recogida de grabaciones orales exige abordar un proceso manual lento y complejo. Este proceso de transcripción se tiene que realizar manualmente, por su complejidad (ruidos, pronunciación descuidada, solapamientos, etc.). Esta metodología, por tanto, no es viable para un estudio exhaustivo de la aparición de nuevos préstamos en tiempo real. La propuesta metodológica de este proyecto es observar la innovación lingüística que se está produciendo en una red social tan activa como Twitter. Nos interesan las innovaciones léxicas en marcha que se puedan observar de manera masiva, general y rápida. Los mensajes de Twitter proporcionan una fuente de datos inmejorable para este fin por su inmediatez, por su tamaño y por la disponibilidad en formato electrónico.

Efectivamente, la lengua en las redes sociales es un tipo de discurso que no se puede interpretar estrictamente en términos de lengua hablada o lengua escrita. De acuerdo con la sociolingüística de la globalización, las redes sociales son un medio a través del cual se revelan algunas características fundamentales de las sociedades contemporáneas: movilidad, nuevas identidades, interacciones a diferentes escalas, comunicación instantánea, entre otras. Desde el punto de vista lexicográfico, esta compleja realidad obliga a prestar atención a una nueva dimensión del préstamo como proceso, donde los usos muestran la espontaneidad de la lengua hablada, pero al mismo tiempo exhiben la fijación de la lengua escrita. En otras palabras, las redes sociales nos muestran usos lingüísticos difíciles de encontrar en otras fuentes y que son muy apropiados para la detección de neologismos.

Entre las redes sociales en la actualidad, Twitter destaca por ser una de las más difundidas internacionalmente. Este servicio de "microblogs" ofrece características reseñables desde el punto de vista lingüístico: la primera es la limitación en el número de caracteres, lo que obliga a una simplificación del contenido y a una flexibilización de la forma. Esto afecta negativamente al estilo sintáctico o discursivo, pero no al léxico. Por otra parte, la espontaneidad e inmediatez de los mensajes los sitúa muy próximos al discurso oral, en el que nacen muchos de los préstamos léxicos.

La aportación del IIC al proyecto se ha centrado en una nueva metodología de búsqueda de anglicismos producidos en Twitter para seleccionar candidatos y validarlos con ejemplos del corpus. El proceso se articula en tres fases:

1. Compilación del corpus de mensajes, siguiendo unas pautas sociolingüísticas
2. Procedimiento de "limpieza" de los mensajes
3. Extracción de candidatos a anglicismo

FASE 1. Compilación del corpus de mensajes

Esta metodología implica el uso de herramientas informáticas que ayudan a filtrar una gran cantidad de datos hasta proporcionar una lista de candidatos léxicos que los lexicógrafos pueden analizar exhaustivamente en contexto para decidir su inclusión en el diccionario.

El corpus de Twitter fue recopilado de entre todos los tuits emitidos desde los Estados Unidos por usuarios que tienen identificado en su perfil que hablan español, que son hispanos o proceden de un país hispano, recogidos en dos momentos diferentes: entre agosto-diciembre de 2014 y entre enero-febrero de 2016. En total, se superan los 850.000 mensajes y más de 15 millones de palabras, con cerca de 175.000 palabras diferentes.

FASE 2. Limpieza del corpus de mensajes

Efectivamente, al ser recopilados de manera exhaustiva por los parámetros de lengua o lugar declarados en el perfil, el conjunto de tuits recolectados presenta mucho "ruido". Debido a esto, la estrategia básica ha sido eliminar todo tuit que esté solo en inglés y conservar los que contienen una mayoría de palabras en español. Hay que tener en cuenta que muchos anglicismos pueden conservar la ortografía original inglesa. Estos han sido los procedimientos de limpieza de datos:

- Eliminación de los mensajes que no están en español mediante un programa que identifica el idioma predominante en el mensaje. No olvidemos que la mayoría de los hispanos estadounidenses son bilingües y escriben en ambas lenguas. Con esto, se redujo el corpus a menos de 400.000 tuits.

- Eliminación de los tuits repetidos, ya que no añaden nada nuevo y falsean los recuentos de frecuencia.

- Eliminación de fragmentos dentro de los tuits que no añaden información a la búsqueda de anglicismos, como los emoticonos o la identificación de usuarios (@) y URLs. También se han eliminado fragmentos que no contienen caracteres en inglés o español; es decir, todas aquellas cadenas de caracteres con elementos que no son propiamente palabras (signos de puntuación, números, caracteres gráficos).

El resultado final del proceso de limpieza ha producido una lista de más 172.000 palabras diferentes. Aquí se incluye cualquier cadena de caracteres separada por blancos, desde palabras (reales, inventadas y mal escritas) y acrónimos, hasta expresiones con hashtags (#), exclamaciones y vocalizaciones.

FASE 3. Extracción de candidatos a anglicismo

Naturalmente, la inmensa mayoría de la lista no son anglicismos, pero no es realista seleccionar a mano los 172.000 términos para localizar los posibles candidatos. Por ello, hemos aplicado una serie de filtros automáticos para reducir el número de palabras que han de analizarse manualmente:

- **Análisis morfológico automático:** tomamos la lista y la pasamos por el analizador GRAMPAL, que está basado en un gran lexicón de más de 50.000 lemas. A todas aquellas palabras que el analizador no reconoce, se les asigna la etiqueta de "palabra desconocida". De esta manera, nos quedamos solo con las palabras desconocidas por el diccionario y descartamos las que ya sabemos que existen en español.

- **Filtro de descarte nombres propios, onomatopeyas, etc.** Asumimos que una parte sustancial de las palabras desconocidas serán anglicismos, pero también nombres propios e incorrecciones. Este filtro ha reducido la lista a más de la mitad: 95.000 palabras diferentes. De nuevo, examinar esta lista a mano es inviable, porque incluye innumerables faltas de ortografía o expresiones típicas de la escritura en Twitter. Por tanto, aplicamos un nuevo filtro que descarte nombres propios, repeticiones de letras, onomatopeyas, siglas, risas, letras aisladas, diminutivos en ito o –azo y otros elementos similares.

- **Extracción de palabras con sufijos productivos en español:** la lista seguía siendo muy extensa, así que se optó por hacer búsquedas agrupadas por terminaciones. Verbos que acaben en ear, - izar; sustantivos en er, -eo, -eador, -ería, -ación; adjetivos en ente, -oso, -ivo. Este último filtro generó una lista de 5.684 palabras flexionadas diferentes, que ya se verificó a mano, eliminando variantes, faltas de ortografía y nombres propios que se habían escapado anteriormente. Esta lista limpia contenía 3.876 palabras, potenciales anglicismos. Sin embargo, había muchas palabras ya reconocibles como anglicismos. Finalmente, los especialistas del Observatorio del Instituto Cervantes en Harvard seleccionaron 578 candidatos para contrastar los ejemplos en contexto y confirmar si eran o no auténticos anglicismos.

La siguiente tabla resume el proceso de selección que ha llevado de los 172.000 candidatos iniciales a menos de 600:

Fase	Subproceso	Resultado	Tamaño
Compilación del corpus de Twitter 01	- - - -	Corpus en bruto (tuits emitidos por hispanos en EE. UU.)	850.000 mensajes y más de 15 M de palabras
Limpieza del corpus 02	Selección de tuits mayoritariamente en español	Corpus en español	392.000 mensajes
	Limpieza de ruido propio de Twitter: RT, @, http, emoticonos	Lista de palabras	172.000 palabras (tokens) diferentes
Extracción de candidatos a anglicismo 03	Extracción de palabras desconocidas por etiquetador morfológico	Lista de palabras	95.000 palabras diferentes
	Eliminación de ruido y extracción de candidatos por terminaciones productivas		5.684 palabras diferentes
	Limpieza de faltas de ortografía y regularización de variantes		3.876 palabras diferentes
	Selección manual de candidatos		578 palabras diferentes

El proceso culminó con la identificación en forma de lista de todos los casos correspondientes a los 578 potenciales anglicismos con sus respectivos contextos. La información aportada en la lista final se disponía del siguiente modo:

Término: *Premiación* Texto del tuit	Nombre usuario	Descripción de perfil	Lugar	Lengua	Geolocalización
La ceremonia de ***premiación*** de @MassChallenge estará disponible en live-stream	ConsulmexBoston	Consulado General de México en Boston	Boston, Massachusetts	es	42.3136695,-71.0887545

El fichero con la lista final recoge toda la información útil para el lexicógrafo y consta de más de 89.000 ejemplos repartidos entre los 578 usos candidatos a ser identificados como anglicismos. Este material es de gran interés lingüístico tanto por lo que se confirma como por lo que se descubre. De este modo, estos mensajes podían respaldar el uso de anglicismos detectados en otras fuentes o completar la información previamente disponible. Por ejemplo, el uso plural de formas de las que solo se tenía constancia escrita en singular: cámpuses, cartunes, nuyorricans, utílitis… Análogamente, los ejemplos han servido para documentar el uso de la flexión verbal de formas prestadas: *baquéate, performeas, vacuneó…*

Desde una perspectiva académica, los mensajes de Twitter son un excelente medio para observar cómo se está produciendo la adaptación gráfica de los préstamos, fenómeno de gran variabilidad y que genera dudas tanto entre los hablantes como entre los lexicógrafos. Así ocurre con las variantes ortográficas más populares en Twitter de *brifin, bróder, cherman, cul, díler o mui.*

Sin embargo, la aportación de los mensajes de Twitter se hace especialmente relevante para el descubrimiento de nuevas formas o voces no registradas con anterioridad o que están surgiendo en el momento de realizar el trabajo lexicográfico. En este sentido, la metodología aquí explicada ha permitido la detección de más de 500 voces, supuestos anglicismos susceptibles de aumentar el lemario del DAEE.

Son formas de algún modo interesantes y novedosas, pero que requieren un análisis lingüístico contextualizado para comprobar si efectivamente merecen su incorporación al diccionario. Tras completar diferentes análisis, el conjunto final de nuevos anglicismos aportados por este método constituye una cifra de varias decenas. En el artículo académico que ha de publicarse dando cuenta de todo este proceso de búsqueda e identificación, se analizan las siguientes voces: *bróder, chilin, coworker, fangirlear, favear, gossipeo y chin.*

Como conclusión, esta investigación hace una aplicación exhaustiva de las **tecnologías informáticas** y de la **lingüística de corpus** a los nuevos canales de comunicación social para encontrar ejemplos de anglicismos que están surgiendo y se están desarrollando en los Estados Unidos. Esta metodología aporta evidencias difícilmente disponibles en corpus escritos u orales, especialmente en el ámbito del descubrimiento de neologismos no registrados con anterioridad por los lexicógrafos.

Por otra parte, los datos recogidos ayudan a mejorar la información ya existente en otras fuentes lexicográficas, incorporando una cuantificación del uso de las nuevas voces, la observación de la adaptación gráfica de los préstamos o la evolución de las formas flexivas de verbos y sustantivos. Una parte esencial de esta metodología es proporcionar el contexto de uso con información sobre el perfil de usuario o los datos de geolocalización del envío del mensaje. Esta información servirá para analizar la dimensión geolectal del anglicismo.

En un próximo análisis proyectamos aplicar la misma metodología sobre el español utilizado en blogs y en textos de Facebook.

www.iic.uam.es

instituto
de ingeniería
del conocimiento

Informes del Observatorio / Observatorio Reports

OBSERVATORIO
de la lengua española y las culturas hispánicas en los Estados Unidos

C/ Francisco Tomás y Valiente, 11 EPS
Edificio B, 5ª planta
UAM Cantoblanco. 28049 Madrid
Tel.: (+34) 91 497 2323
Fax: (+34) 91 497 2334
iic@iic.uam.es

Harvard University
Faculty of Arts and Sciences
2 Arrow Street, 4th floor
Cambridge, MA 02138
info-observatory@fas.harvard.edu