EDICIONES Universidad Valladolid

MarcoPolo: una plataforma digital de análisis del discurso electoral en Twitter [1]

MarcoPolo: a digital platform for analyzing electoral discourse in Twitter

MANUEL ALCÁNTARA PLÁ
Wor(l)ds Lab, Facultad de Filosofía y Letras, Universidad Autónoma de Madrid
manuel.alcantara@uam.es
ORCID: http//orcid.org/0000-0002-2438-0293

ANA RUIZ SÁNCHEZ
Wor(l)ds Lab, Facultad de Filosofía y Letras, Universidad Autónoma de Madrid a.ruiz@uam.es
ORCID: http://orcid.org/0000-0001-6178-3334

MARISOL BENITO REY
Wor(l)ds Lab, Facultad de Filosofía y Letras, Universidad Autónoma de Madrid marisol.benito@uam.es
ORCID: http://orcid.org/0000-0002-2187-0094

VANESSA AMESSA GARCÍA
Wor(l)ds Lab, Facultad de Filosofía y Letras, Universidad Autónoma de Madrid
vanessa.amessa@gmail.com

ALEJANDRO MARTÍN JIMENO
Wor(l)ds Lab, Facultad de Filosofía y Letras, Universidad Autónoma de Madrid
jandro.martinj@gmail.com

Recibido: 15 de julio. Aceptado: 30 de octubre.
Cómo citar: Alcantara Plá, M., Ruiz Sánchez, A., Benito Rey, M., Amessa Garcia, V. & Martín Jimeno, A. (2018), "MarcoPolo: a digital platform for analyzing electoral discourse in Twitter", *Revista Estudios del Discurso Digital (REDD)*, 1: 7-29.
DOI: https://doi.org/10.24197/redd.1.2018.7-29

[1] Este trabajo se ha realizado en el marco del Proyecto de Investigación *Estrategias de encuadre y articulación del discurso político en 140 caracteres*, financiado por el Ministerio de Economía y Competitividad español en su programa de I+D (FFI2014-53958-P).

Resumen: Objetividad, validez y replicabilidad son requisitos básicos de la investigación científica. Los estudios de análisis del discurso digital no deben ser ajenos a ello, sobre todo en ámbitos con alto impacto social como es el discurso político. En este artículo presentamos la plataforma MarcoPolo, una herramienta implementada para cumplir con ellos respetando las restricciones de licencias del corpus y para hacer las investigaciones accesibles a un público más general. Permite navegar por el corpus a través de más de 10.000 gráficos que representan los usos léxicos en Twitter de los cinco partidos más votados en España y sus candidatos entre octubre de 2015 y junio de 2016. Se describe el diseño de la página, su uso y algunos ejemplos de su funcionalidad para los investigadores del discurso político digital.

Palabras clave: discurso político, marcos, CADS, campaña electoral, Twitter

Abstract: Objectivity, validity and replicability are basic requirements of scientific research. Studies into digital discourse analysis must not be unrelated to this, especially in areas with elevated social impact such as political discourse. In this article, the MarcoPolo platform is presented, a tool implemented to comply with these requirements while respecting the restrictions on corpora licenses and make research accessible to a general public. It allows users to browse the corpus through over 10,000 graphs representing the lexicon used on Twitter from the five most voted parties in Spain and their candidates between October 2015 and June 2016. The design of the page is described along with its usage and a selection of examples of its functionality for digital political discourse researchers.

Keywords: political discourse, frames, CADS, campaign, Twitter

INTRODUCCIÓN

MarcoPolo es una plataforma en línea para facilitar la navegación y el análisis del contenido del discurso político digital en España. Se encuentra dentro de la página de *Wor(l)ds Lab*, nuestro grupo de investigación, y su dirección es www.worldslab.eu/marcopolo. Los datos que maneja provienen de los mensajes publicados en Twitter entre octubre de 2015 y junio de 2016 por las cuentas oficiales de los cinco principales partidos de ámbito nacional (PP, PSOE, Podemos, Ciudadanos e Izquierda Unida[2]) y de sus correspondientes cinco líderes

[2] Izquierda Unida se presentó a las elecciones 2015 en una coalición de formaciones de izquierdas bajo el nombre de Unidad Popular en Común y a las elecciones 2016 en coalición con Podemos bajo el nombre de Unidos Podemos.

(Mariano Rajoy, Pedro Sánchez, Pablo Iglesias, Albert Rivera y Alberto Garzón). Determinadas circunstancias históricas han dotado a este corpus de una de sus características más interesantes: recopilado en un espacio de tiempo relativamente breve, en él se incluyen dos campañas electorales, la de diciembre de 2015 y la de junio de 2016.

MarcoPolo responde a dos retos. El primero es cómo analizar el discurso empleado en los tuits aprovechando que el medio digital nos permitía recuperar el corpus prácticamente íntegro. Esta condición de integridad ofrece atractivos evidentes para la investigación. Podemos trabajar con un corpus cuyo volumen de datos es suficiente para poder hacer una radiografía real del discurso político tanto global como por partidos. Superamos así la dificultad científica frecuente en estudios del discurso de tener que extraer conclusiones a partir de muestras poco representativas. Las frecuencias estadísticas, sin embargo, son insuficientes por sí solas para ser suficientemente validadas. Decidimos adoptar una metodología híbrida que combinara las posibilidades de la Lingüística de corpus para el tratamiento de los 117.010 *tuits* junto con el estudio cualitativo propio del Análisis Crítico del Discurso[3]. Para ello tuvimos que dotarnos de un sistema que facilitara la interpretación de los datos estadísticos y que permitiera a la vez incorporar el análisis cualitativo.

El segundo reto al que MarcoPolo da respuesta responde a cómo poner a disposición de la comunidad científica y del público en general no solo las conclusiones de nuestros estudios, sino también el material suficiente que permitiera validarlos.

El enfoque abierto que pretendíamos nos exigió creatividad para solventar legalmente las restricciones de la Ley de Protección de Datos y las de la propia red social. Twitter tiene unas normas estrictas sobre la redistribución de *tuits*[4]. En concreto, solo permiten compartir el contenido de 50.000 mensajes diarios y sin hacerlos en ningún caso públicos. Su recomendación para compartir grandes cantidades de mensajes, como sería nuestro corpus, es hacerlo a través del código que

[3] Véase Partington et al (2013) para una descripción detallada de la métodogía híbrida de análisis del discurso basada en corpus (Corpus Assisted Discourse Studies, CADS), y Alcántara-Plá & Ruiz-Sanchez (2017, 2018) para ejemplos de nuestra aplicación con Análisis Crítico del Discurso).

[4] Se pueden consultar en https://developer.twitter.com/en/developer-terms/more-on-restricted-use-cases

identifica a cada mensaje cuando se descarga a través de la *API*. Este código permitiría al receptor localizarlo en la plataforma para recuperar el contenido. Obviamente estos requisitos hacen imposible conseguir nuestros retos mostrando el corpus de forma directa. MarcoPolo es la solución que encontramos.

1. DISCURSO DIGITAL Y DISCURSO POLÍTICO

El análisis del discurso político digital ha ganado protagonismo en los últimos años por la importancia que las campañas electorales le han dado a este espacio. Los primeros usos exitosos de Internet para la comunicación política son de principios del siglo XXI, cuando se crearos los primeros blogs (Tuñez & Sixto, 2010). La campaña de las presidenciales estadounidenses de Barack Obama en el año 2008 supuso la constatación de que el ámbito digital ya no podría volver a obviarse en política.

La comunicación digital no es una adaptación de los mensajes tradicionales, sino que ha supuesto toda una transformación de las estrategias políticas y de campaña (Gibson & Römmele, 2008; Conway et al, 2015). Esto es así hasta el punto de que se habla de la *democracia electrónica* o de la *democracia digital* (Dader, 2003) para hacer referencia a las nuevas formas de comunicación y acción políticas que buscan facilitar la participación de la ciudadanía en los procesos políticos (Coleman, 2013).

La comunicación política en España en la red social Twitter cobra fuerza en la campaña de las elecciones municipales de 2011. Desde entonces, ha sido objeto de estudio en diversos proyectos (p.ej. Aragón et al., 2013; Mancera & Pano, 2013; Zamora & Zurutuza, 2014; Padilla, 2015; Zugasti & Sabés, 2015).

2. UN CORPUS DE TEXTOS DIGITALES

La hipótesis de partida del proyecto de investigación en el que se enmarca este trabajo era que la fuerte restricción provocada por el formato de 140 caracteres -que imponía Twitter entonces[5]- podría

[5] En la actualidad Twitter ha ampliado el número de caracteres a 280.

generar una intensificación del discurso, en comparación con lo que sucede en el resto de formatos tradicionales en comunicación política, como son los discursos parlamentarios, programas electorales y los argumentarios. Si la hipótesis era cierta, deberían verse afectadas las estrategias discursivas de encuadre y articulación en el discurso político de esta red social. Tanto los corpus como la plataforma que se presentan aquí han sido los instrumentos de trabajo utilizados para comprobar esta hipótesis. El desarrollo actual de MarcoPolo permite una variedad de estudios sobre el corpus que supera los planteamientos iniciales del proyecto.

Metodológicamente y teniendo en cuenta que nuestra investigación se centraba en la selección léxica y en las estrategias de encuadre, adoptamos como base conceptos claves de la *Teoría de la articulación* (Howarth, 2005) y de la *Semántica de marcos* (Fillmore, 1982; Langacker, 1991; Huckin, 2002).

2.1. La fuente de los textos

Con la intención de trabajar desde un punto de vista empírico, utilizamos como objeto de estudio la suma total de *tuits* publicados en las diez cuentas oficiales antes citadas. Estas diez cuentas acumularon en el periodo comprendido entre octubre de 2015 y junio de 2016 unos 117.010 mensajes, que suman 2.120.371 palabras[6].

El corpus fue obtenido automáticamente con un sencillo *script* que registraba los *tuits* en el momento de su publicación gracias a la *Application Programming Interface* (API) ofrecida por Twitter. Lo implementamos en el lenguaje de programación *Python*[7] utilizando el módulo *Tweepy*[8] de Joshua Roesslein.

De esta forma, Twitter nos permitió descargar los mensajes en el formato estándar abierto *JavaScript Object Notation* (*JSON*) con toda la información etiquetada: texto del *tuit*, momento en que fue publicado,

[6] El concepto de *palabra* tiene aquí un sentido amplio y poco ortodoxo. Como explicamos más abajo, el corpus fue anotado morfológicamente y segmentado teniendo en cuenta la existencia de multipalabras. El recuento de palabras incluye también elementos característicos de los textos de esta red social que no existen en otros contextos como los *hashtags* o las direcciones web.

[7] https://www.python.org

[8] http://www.tweepy.org

datos del emisor, número de *retuits*, etc. MarcoPolo solo hace uso en su versión actual del texto de los mensajes y del nombre de la cuenta que los publica.

La decisión de descargar los *tuits* en el momento que se publicaban nos garantizó que no perdíamos información de lo comunicado a través de la red social independientemente de que los autores pudieran haber decidido borrarlo después. A cambio, asumimos el riesgo de perder algún mensaje en caso de que hubiera cortes puntuales en la red. Como demuestra la cantidad de *tuits* incluidos en el corpus, esta pérdida, inevitable con un *script* funcionando de continuo durante nueve meses, no es significativa.

Un sesgo más importante, aunque también inevitable, tiene que ver con la representatividad en el corpus. La siguiente tabla (1) muestra el número de mensajes publicados por cada cuenta. Las diferencias entre ellas son amplias. Entre los partidos políticos, Podemos es el que más publicó (26.210) y lo hizo el doble de veces que el Partido Popular, que fue el que menos (13.361). En cuanto a los candidatos, Albert Rivera (4.586) y Pedro Sánchez (4.513) fueron los más profusos mientras que Pablo Iglesias apenas llegó a publicar un tercio de esas veces (1.629).

Podemos	**n**	**IU**	**n**
@Ahorapodemos	26.210	@iunida	20.111
@Pablo_Iglesias	1.629	@agarzon	3.316
C's	**n**	**PP**	**n**
@CiudadanosCs	19.728	@PPopular	13.361
@Albert_Rivera	4.586	@marianorajoy	4.299
PSOE	**n**		
@PSOE	18.350		
@sanchezcastejon	4.513		

Tabla 1. Número de *tuits de* cada cuenta de octubre de 2015 a junio de 2016

Este desequilibrio se ve matizado en MarcoPolo por otra característica particular del corpus apreciable en la tabla: la cantidad de publicaciones de los partidos es casi inversamente proporcional a la de sus candidatos. Ya hemos señalado que Podemos es el partido que utilizó con mayor intensidad la red social y que su candidato es el que menos lo hizo. Entre ambos publicaron 27.839 *tuits*. Por otro lado, el Partido

Popular fue el que menos escribió, pero su candidato Mariano Rajoy es el segundo con más *tuits*. Entre ambos suman 17.660.

MarcoPolo muestra los subcorpus sumando cada partido con su candidato y no de manera independiente. Esta decisión se tomó para garantizar que había datos suficientes para el mayor número posible de palabras, pero también reduce las diferencias de tamaño aunque estas siguen siendo importantes en los casos más extremos. Como veremos en la sección dedicada al diseño de la plataforma, esta permite seleccionar los subcorpus que queremos que se visualicen, facilitando que nos centremos, si así lo deseamos, en el uso concreto de un partido/candidato. No obstante, habrá que tener siempre presentes las probabilidades de que un subcorpus tenga más peso que otro al calcular las estadísticas generales de un término.

Otro aspecto a tener en cuenta a la hora de interpretar los resultados es que el corpus incluye todos los mensajes que fueron difundidos a través de las cuentas estudiadas. No se diferencia, por lo tanto, si fueron realmente escritos desde esa cuenta o solo difundidos por ella. Cualquier mensaje transmitido desde las cuentas se considera parte de la campaña de comunicación de los partidos y candidatos que estas representan. Este dato es especialmente relevante en una red como Twitter, donde los *retuits* (mensajes de otros que otra cuenta distribuye entre sus seguidores) tienen gran protagonismo. Nuestra plataforma muestra la relación entre conceptos expresada a través de cada cuenta, sin diferenciar la autoría original, siendo fiel así a la exposición que tuvieron los seguidores de dichas cuentas.

Por otro lado, la autoría es difícil de determinar al tratarse de cuentas de organizaciones donde la comunicación se diseña y filtra a través de equipos profesionales. De nuevo, no tenemos manera de saber quién escribió realmente cada mensaje (aunque aparezca en una cuenta personal de un candidato), pero sí lo que recibieron sus seguidores.

2.2. El texto procesado

Hemos señalado que MarcoPolo solo analiza los textos de los mensajes, obviando los metadatos proporcionados por la *API* de Twitter. Explicaremos en la siguiente sección que los datos que se presentan en la plataforma no son los textos en sí, sino datos obtenidos estadísticamente a partir de ellos utilizando la herramienta de lingüística de corpus Sketch Engine (Kilgarriff, 2003).

Para poder obtenerlos, los textos deben pasar por varios procesados, siendo el primero su etiquetado morfológico. Lo llevamos a cabo con la librería de licencia pública Freeling (Padró y Stanilovsky, 2012) y es crucial en nuestra investigación por dos motivos. Por un lado, el etiquetado morfológico desambigua las palabras homógrafas evitando errores posteriores a la hora de contar la frecuencia de cada término. Por otro lado, veremos que tanto Sketch Engine como MarcoPolo utilizan la diferenciación entre clases de palabras para efectuar y mostrar las estadísticas.

Freeling segmenta el texto en palabras y después las etiqueta. El segmentador utiliza expresiones regulares adaptadas al español mientras que el etiquetador se basa en módulos que sirven para detectar diferentes elementos como son la puntuación, las fechas, las multipalabras, etc. Para analizar un término, lo busca en un diccionario de lemas al que se aplican reglas de sufijación. Cuando varias soluciones pueden ser correctas, el sistema calcula la probabilidad de cada una según su experiencia previa y el contexto del término analizado. Cuando la palabra no se encuentra en el diccionario, intenta el tipo de palabra más probable teniendo en cuenta su terminación.

El etiquetado morfológico es fundamental para obtener unos resultados tan ricos como los ofrecidos por nuestra plataforma. Sin embargo, la breve descripción del funcionamiento de Freeling en el párrafo anterior debe servir de advertencia en cuanto a los resultados. Nuestro corpus tiene características especiales que dificultan la automatización del etiquetado porque el español de Twitter sigue convenciones propias. A continuación ejemplificamos las tres más frecuentes.

La más distintiva es el uso de etiquetas o *hashtags*, términos que funcionan como enlaces entre mensajes que utilizan la misma etiqueta y que se pueden considerar metafóricamente como constructores de espacios de conversación (Alcántara-Plá, 2017). Se identifican formalmente porque comienzan con el símbolo # y porque no aceptan espacios. Esto último hace que se empleen a veces guiones bajos en su lugar o, lo que es más común, que aparezcan varias palabras juntas. En nuestro corpus encontramos ejemplos como *#PactoporEspaña* o *#LaEspañamoderada* que obviamente no podrían estar en el diccionario de Freeling.

También es frecuente la aparición de enlaces a páginas o elementos multimedia con el formato *URL* (“http://www….”). Aunque nosotros no

los hemos considerado palabras en la plataforma, sí que forman parte de las oraciones que encontramos en los mensajes y estorban inevitablemente al cálculo de probabilidades que tengan en cuenta el contexto. Si una palabra puede ser de una clase u otra dependiendo del tipo de palabra que le sigue, que esta sea una *URL* no le dará al sistema la información que espera.

Por último, la ortografía que se utiliza en las redes sociales a menudo no respeta la norma, sacrificada esta para lograr mayor expresividad o para reducir el número de caracteres necesarios para expresar el mensaje. Suelen omitirse las letras que no representan sonidos (como la h o la u que sigue a la g), se reducen las palabras a sus esqueletos consonánticos (*t spr dsps* por *te espero después*) o incluso se utilizan logogramas (*q* en lugar de *que*) (Alcántara-Plá, 2017, p. 91-101). Aunque se podría esperar una ortografía más cuidada de lo habitual por ser parte de la imagen de unos partidos políticos, lo cierto es que sí encontramos bastantes casos de estos fenómenos. Si buscamos, por ejemplo, *x* en MarcoPolo, observamos que aparece 237 veces y que lo hace frecuentemente después de verbos como *apostar*, *trabajar*, *luchar* o *impulsar*: se trata, por lo tanto, del logograma de la preposición *por*.

El caso de la *x* sirve pone de manifiesto tanto los problemas que un corpus así genera para un sistema como Freeling, entrenado sobre corpus de español normativo, como que estas dificultades no evitan que la plataforma acabe por darnos información interesante incluso para estas expresiones. No obstante, es importante tenerlo en cuenta a la hora de utilizarlo para cualquier estudio (uno sobre las preposiciones, por ejemplo, probablemente debería considerar esa *x* e incluir sus gráficos).

3. LA INFORMACIÓN EN MARCOPOLO

Hemos comentado que el tamaño del corpus de estudio hacía necesario el uso de métodos estadísticos de la lingüística de corpus para trabajar con una cantidad tan elevada de textos; también que la aproximación teórica del proyecto en que se inscribe MarcoPolo es la del estudio de los marcos semánticos transmitidos en los *tuits*. Distintas teorías, la mayoría próximas a la lingüística cognitiva, proponen el estudio de la relación entre dos palabras a través de su frecuencia de uso en un mismo contexto. Esta es la idea que hay detrás, entre otros, del *Cognitive Network Model* (Santanen, Briggs and de Vreede, 2000), del *Lexical Priming* (Hoey, 2005), del *Associative Network Model of*

Memory (Anderson, 1983) y, dentro del ámbito de la comunicación, del modelo de la *Network Agenda Setting* (Lang, 2000).

Sin hacer normalmente referencia a ninguna de estas teorías, el trabajo de la lingüística de corpus ha asumido esta relación dándole un papel fundamental a la repeticiones de colocaciones. Sketch Engine proporciona una herramienta valiosa, el Word Sketch (WS), para acercarnos a estas frecuencias. Un WS agrupa las colocaciones más relevantes de una palabra, clasificadas según las relaciones gramaticales (McCarthy et al, 2015). El grado de relación entre palabras se calcula con un algoritmo que toma como base la medida estadística *logDice*. Esta tiene la cualidad de no verse influida por el tamaño del corpus, evitando que el tamaño de este afecte al resultado final.

Cada WS muestra las palabras que se relacionan con el término buscado en diferentes relaciones gramaticales. Por ejemplo, si buscamos el WS de un nombre, nos indicará qué verbos suelen tenerlo como sujeto, cuáles como objeto directo, qué adjetivos le acompañan frecuentemente o qué nombres aparecen completando oraciones del tipo "un x es un …".

En el momento de escribir este artículo, Sketch Engine muestra los WS en dos versiones distintas, la oficial y una beta que se lanzará próximamente. Ambas están compuestas de tablas en las que aparecen las palabras con mayor relevancia para cada relación gramatical con el término buscado en un corpus concreto. La versión beta simplifica los datos al no presentar las puntuaciones numéricas ni las frecuencias. Es importante señalar que los resultados no muestran las palabras más frecuentes en términos absolutos, sino las más relevantes según la puntuación obtenida con la citada medición estadística *logDice*.

La plataforma MarcoPolo basa su presentación en estos WS gracias a que Sketch Engine incluye una *API* que permite descargar esa información en formato *JSON*. Igual que descargamos el corpus de mensajes desde Twitter, hicimos un proceso similar para almacenar los WS de todos los términos de nuestro corpus. Lo conseguimos en dos pasos. Primero, descargamos el listado del vocabulario presente en el corpus, ordenado de mayor a menor frecuencia. Segundo, pedimos a la *API* el WS de cada una de las palabras de ese listado.

En realidad, esto no significa que contemos con un WS para cada palabra. Por un lado, la palabra debe cumplir el requisito de tener un mínimo de datos necesario para que el *logDice* sea significativo. Muchas palabras apenas aparecen una o dos veces, lo que hace que el WS sea imposible e irrelevante. Por otro lado, le hemos pedido al sistema la

información de seis WS por cada palabra, correspondientes al corpus global y a cada uno de los cinco subcorpus de candidato/partido. Como hemos señalado más arriba, estos subcorpus tienen tamaños diferentes y además no coinciden en la importancia dada a cada tema, con lo que una palabra puede tener ejemplos suficientes en unos subcorpus (o en el corpus general) y no en otros.

4. EL DISEÑO DE LA PLATAFORMA

MarcoPolo es una página web que permite navegar los marcos del corpus de *tuits* políticos con al menos tres ventajas frente a Sketch Engine (y con una desventaja importante).

La primera ventaja está relacionada con el reto explicado anteriormente de querer facilitar a la comunidad científica y al público en general el acceso a los datos de nuestra investigación con las mínimas restricciones. Las limitaciones legales impuestas por Twitter no nos permiten compartir el corpus en sí. Esta es la desventaja más clara de nuestra propuesta: no es un acceso directo a los textos literales que se están estudiando. Sin embargo, estos tal y como los distribuye Twitter en *JSON* tampoco ofrecerían un acceso sencillo a los datos estadísticos que fundamentan nuestras investigaciones. MarcoPolo es la forma que hemos encontrado de solucionar este problema: mostramos todos los datos que tenemos en nuestra mano sin la necesidad de distribuir realmente el corpus ni de obligar a tener una cuenta en Sketch Engine (que sería inútil para nuestro objetivo por la ausencia del corpus).

La segunda es la visualización a través de gráficos de los datos de los seis corpus de manera simultánea, lo que facilita la comparación. Como se verá enseguida, los gráficos están diseñados de forma que las diferencias entre partidos sean lo más evidentes posible. La disposición de los círculos concéntricos y el tamaño de sus secciones las muestran visualmente.

Por último, MarcoPolo incluye descripciones realizadas por nuestro equipo para muchas de las palabras del corpus. Estos textos, además de su valor para explicar algunos términos, pueden servir de guía sobre cómo formular posibles conclusiones de los datos estadísticos que se muestran.

4. 1. La web

La página es de manejo sencillo, pero merece la pena detenerse en su funcionamiento para poder sacarle el máximo provecho. Tiene tres espacios claramente diferenciados: un menú superior, una columna en el lateral izquierdo y un espacio central para mostrar los datos. Desde el punto de vista técnico, está diseñada en HTML5, CSS3 y Javascript, con algunas partes en jQuery[9]. La librería de presentación que hemos utilizado es Bootstrap[10] y la librería encargada de los gráficos en anillos es ChartJS[11], ambas con licencia libre.

El menú superior permite acceder a la página del proyecto que genera este recurso, a una ayuda específica para su manejo, y al tipo de investigaciones posibles a través del listado de contribuciones científicas realizadas con él. En caso de que se haya buscado ya un término, en su extremo derecho aparecerá en un recuadro con el número de veces que se ha encontrado en el corpus global.

La columna lateral sirve para lanzar la búsqueda de información sobre los términos que quieran consultarse. Para ello, disponemos de una caja de búsquedas y de dos listados con todas las palabras del corpus. La caja muestra dinámicamente las palabras que empiezan con las letras que vayamos introduciendo. Esto es especialmente útil por las características señaladas del corpus, donde algunos términos pueden incluir peculiaridades ortográficas diferentes a la norma.

[9] https://jquery.com
[10] https://getbootstrap.com
[11] http://www.chartjs.org

Imagen 1. Columna lateral de MarcoPolo

Los que aparecen en los listados son los lemas, es decir, si buscamos un nombre común, la información que se nos dará se habrá obtenido del uso de ese nombre tanto en singular como en plural. Si buscamos un verbo, los datos tendrán en cuenta su uso en cualquiera de las formas de su conjugación. Por este motivo, puede sorprender que aparezcan en los listados palabras marcadas morfológicamente (por ejemplo, nombres comunes en plural). Esto se debe a que su uso en los *tuits* tenía alguna peculiaridad que lo distinguía de la forma *normal* de la palabra. Lo más común es la aparición del signo # al principio del término, la clave que utiliza Twitter para identificar *hashtags*.

Los listados permiten recorrer las palabras en dos órdenes diferentes. El superior las lista de mayor a menor frecuencia en el corpus. Están excluidas aquellas palabras sin contenido sustantivo, como son los artículos, las conjunciones y las preposiciones. Teniendo esto en cuenta, la primera palabra que aparece -y, por lo tanto, la más frecuente- es el verbo "poder".

El listado inferior muestra sólo las palabras que incluyen un breve análisis cualitativo elaborado por nuestro equipo. Como se ha señalado más arriba, el enfoque teórico detrás de esta investigación se centra en el análisis de marcos semánticos. Los textos destacan aquellas colocaciones que nos ayudan a interpretar esos marcos.

Por último en la columna, hay un selector que permite elegir en qué subcorpus de partidos políticos y candidatos queremos hacer la búsqueda. Solo esos aparecerán representados en los gráficos. De esta forma, se facilitan los trabajos de comparación entre partidos concretos. Por defecto aparecen todos seleccionados y el corpus global es el único que no puede eliminarse de los gráficos.

El espacio central es el reservado para los resultados de las búsquedas. Muestra los datos y está organizado en pestañas. En caso de que la palabra buscada tenga una descripción textual, esa será la primera pantalla que veremos. La extensión del texto dependerá del contenido de la palabra y del interés de las relaciones. En aquellos casos en que se haya realizado un estudio completo del término y que este haya sido publicado en alguna revista o libro, se ofrece la referencia de la publicación junto a un breve resumen y se remite al listado de contribuciones científicas.

El resto de pestañas contiene las diferentes relaciones gramaticales según la clase de palabra. En el caso, por ejemplo, de un nombre común, aparecen las pestañas de sus modificadores, los verbos que le acompañan en su contexto izquierdo y derecho, nombres que aparecen enumerados con él, sintagmas preposicionales típicos, etc.

4. 2. Los gráficos

Para cada marco, obtenemos una representación en forma de gráfico de anillos bidimensional y una tabla con los datos que se han utilizado para dibujarlo.

Los anillos concéntricos representan los diferentes corpus, distinguidos por un código de colores. El corpus global aparece en gris. Los subcorpus utilizan los colores representativos de cada partido: verde para IU, rojo para PSOE, morado para Podemos, azul para PP y naranja para C’s. El orden en el que aparecen depende del número de casos que tengan de la relación que estemos analizando. En el ejemplo de la imagen 2, con los modificadores del nombre *voto*, IU es el círculo exterior porque es el partido que más casos tiene de esta relación (122), seguido de PP (62), C’s (50), PSOE (45) y Podemos (20). La suma de cada fragmento de circunferencia de cada partido forma el círculo gris completo.

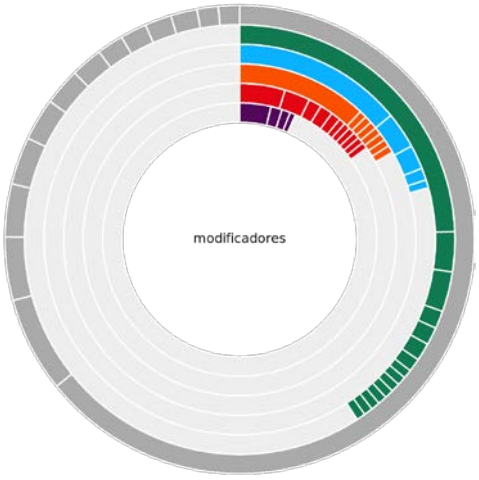

Imagen 2. Modificadores del nombre “voto”

Cada anillo está dividido en sectores que representan a las palabras que se han utilizado con el término buscado. Estas aparecen al pasar el curso por encima. El primer sector del anillo gris lo ocupa el término más frecuente en todo el corpus. Los sectores se ordenan de mayor a menor frecuencia según las manillas del reloj, y esa diferencia de frecuencias se ve reflejada en el tamaño de cada sector.

Si un subcorpus no tiene ejemplos de un marco determinado, MarcoPolo no lo mostrará aunque este partido aparezca en el selector de corpus.

Los sectores proporcionan además una utilidad secundaria: facilitar el análisis navegando por los marcos sin pasar por el buscador. Cada sector es un enlace a la página del término correspondiente, lo que nos permite establecer y contrastar relaciones.

Finalmente, debajo de cada gráfico aparecen los datos que aparecen en él. Hay una tabla por cada subcorpus, reconocibles por utilizar los mismos colores que los anillos, y una gris con los datos globales. Al igual que los gráficos, las tablas están ordenadas de mayor a menor frecuencia, empezando por el global.

Quienes estén acostumbrados a Sketch Engine, esperarán que los datos aparezcan ordenados según la puntuación que obtienen en cada algoritmo. Sin embargo, nuestra plataforma utiliza dos variables para elaborar sus gráficos y tablas. Las palabras que aparecen son seleccionadas por Sketch Engine por su alta puntuación en la medida *logDice*, pero después MarcoPolo las reordena según su frecuencia

absoluta en cada corpus. De esta forma, se facilita la comparación de los partidos puesto que no se presentan con un número relativo a cada corpus, sino la cantidad real de veces que un término se ha utilizado.

5. MARCOPOLO Y NUESTRAS INVESTIGACIONES

Los datos ofrecidos por MarcoPolo pueden utilizarse para múltiples estudios. En esta sección, presentamos brevemente y a modo de ejemplo algunos de los que hemos realizado en nuestro equipo sobre los corpus y cómo la plataforma nos ha ayudado a conseguir los retos planteados en la introducción de este artículo. Los acercamientos descritos de ninguna manera pretenden agotar todas las posibilidades de explotación de la plataforma.

Como hemos visto, las estadísticas dan información sobre los temas más frecuentes y los conceptos relacionados con una palabra determinada cuando es usada por diferentes políticos y partidos. Esto puede utilizarse para analizar el discurso desde diferentes perspectivas, como son el tratamiento de asuntos de interés (lo ejemplificaremos con la crisis de refugiados) o la representación de algunos grupos sociales (p.ej. las minorías sociales).

El acercamiento más evidente es el estudio pormenorizado de las principales frecuencias detectadas en el corpus, es decir, cuáles son los términos más repetidos en el discurso de Twitter en cuentas oficiales ligadas a partidos políticos durante el periodo analizado. A través de dichas frecuencias obtenemos los marcos más habituales según cada partido y líder. Como hemos señalado más arriba, el listado de las palabras ordenadas por las frecuencias aparece en la columna de la izquierda de MarcoPolo. En el ángulo superior derecho aparece la frecuencia total de la palabra que estemos investigando.

En términos generales, las palabras más frecuentes nos indican que el uso de Twitter durante las campañas electorales es casi exclusivamente el de actuar como altavoz de los eslóganes y como cámara de resonancia de los eventos offline. Predomina, por lo tanto, la función ecoica de la red social (Alcántara-Plá, en prensa). Los usos son sorprendentemente homogéneos entre los partidos.

Observamos, por ejemplo, que los nombres que hacen referencia a los usuarios de los candidatos están entre las primeras posiciones. Verbos dicendi como *decir* o *hablar* también son frecuentes, repitiendo lo que los políticos estaban comunicando en otros contextos (mítines, programas

en medios de comunicación, etc.). La repetición del adverbio *hoy* con verbos como *haber* y *ser* delata también la cantidad de mensajes que anuncian y retransmiten eventos externos a la red. El verbo *ver* se utilizó 3.016 veces invitando a visualizar entrevistas, debates y vídeos.

Una conclusión que obtenemos de ello es que no se ha aprovechado el potencial de esta red de microblogging como dinamizadora de la conversación con los ciudadanos, característica clave del éxito en campañas como la que llevó al poder a Barack Obama, referente en el uso político-electoral de Twitter (Chadwick et al, 2015).

Analizar con detalle el marco de encuadre de las frecuencias a través de los gráficos arroja otros resultados de gran interés. Por ejemplo, se aprecia que el marco más frecuente cuando se habla de *democracia* es un marco de vulnerabilidad, construido con la recurrencia de verbos como *regenerar, defender, respetar, secuestrar, recuperar, saltar, debilitar, despreciar, desestabilizar* y *sanear*. En un análisis más pormenorizado[12], tuvimos la ocasión de ver que los verbos que remiten a este marco fueron el 36,68% de los usados con *democracia* durante la campaña electoral de diciembre de 2015. En ese mismo estudio contrastamos estos resultados con el uso normal del término en el lenguaje digital, obtenido del corpus general esTenTen (Kilgarriff & Renau, 2013), y vimos que durante la campaña electoral la noción de vulnerabilidad se intensificaba claramente.

Además de las frecuencias, MarcoPolo permite búsquedas sobre temas de interés social específicos como pueden ser la discapacidad, la migración, la sanidad, las pensiones, Europa o el independentismo, por citar solo algunos ejemplos. Los datos muestran si el tema fue considerado relevante en el discurso político de los años 2015 y 2016, qué partidos y líderes los incluyeron en sus agendas, así como el tratamiento discursivo que le dieron.

Así hemos podido comprobar que algunos asuntos claves para la sociedad española apenas se citan. A modo de ejemplo, la palabra *hipoteca* solo se menciona en trece ocasiones, todas ellas por Podemos y su candidato, Pablo Iglesias. Además, nueve de esos casos son en el sentido metafórico de libertad frente a compromisos políticos poco transparentes, no tratando el problema de los afectados por las hipotecas.

[12] Ruiz-Sánchez y Alcántara-Plá (en prensa).

Estas ausencias en fuerte contraste con su relevancia social nos animaron a poner el foco en los silencios del corpus y lo hicimos centrándonos en la presencia de la crisis de los refugiados sirios durante la campaña de diciembre de 2015. Es preciso recordar que esa crisis, coincidente en el tiempo con la campaña, fue reconocida por Europa como su mayor crisis humanitaria desde la Segunda Guerra Mundial. Un informe de la Agencia de la ONU para los refugiados (UNHCR 2016) advirtió de que en el año 2015 hubo más desplazados forzosos que en dicha guerra. Como investigadores, quisimos comprobar si un tema crucial para el futuro de las sociedades europeas -que después se ha mostrado clave en las elecciones de otros países- estaba presente o no en el discurso en campaña[13].

La crisis de refugiados no fue tema en el discurso político en Twitter de la campaña. Solo 17 de los 16.306 *tuits* emitidos desde las diez cuentas durante la campaña electoral del 20-D se dedicaron a este fenómeno, el 0,1%. Si tenemos en cuenta que tres de los ejemplos contienen el término *refugio* con valor metafórico y sin relación con Siria, la cifra real de mensajes es aún inferior. A esto se suma que cinco son en realidad *retuits* del *hashtag #refugeeswelcome* o de las conclusiones -en inglés- de una cumbre de la Comisión Europea. En resumen, el discurso electoral propio relacionado específicamente con la crisis de refugiados sirios se reduce a 9 mensajes, apenas un 0,05% de lo publicado en campaña.

El análisis lingüístico de los nueve *tuits* que sí tratan la crisis abunda en esta falta de interés político. En ellos no se formula ninguna propuesta concreta sobre cómo se tiene previsto afrontarla más allá de formulaciones generales de alta excelencia moral del tipo del ejemplo (1).

(1) Vamos a ser referente en Europa de los derechos, la paz y la acogida de refugiados q huyen de la guerra (Podemos)

Las soluciones prácticas solo se formulan con sujetos impersonales que parecen diluir la capacidad actante del emisor y, por lo tanto, también cualquier responsabilidad.

[13] Alcántara-Plá y Ruíz-Sánchez (2018).

(2) Hay que garantizar los compromisos para que cualquier refugiado acceda a su derecho de solicitar (Podemos)

(3) Hay que abrir Europa a refugiados, inmigrantes y que cierre las puertas a Merkel y sus políticas (IU)

Usamos MarcoPolo para corroborar lo que habíamos encontrado en este estudio. La plataforma nos da el dato de que la palabra *refugiado* fue utilizada 631 veces en los nueve meses del corpus. Supone una proporción algo mayor que la de la campaña de diciembre, pero que tampoco supera el 0,5% de los *tuits* del corpus. La pestaña de los modificadores nos muestra que solo se escribió seis veces con el adjetivo *sirio*, y que las otras procedencias citadas son la somalí y la saharaui. Los verbos que más se utilizan con *refugiado* son *acoger, recibir* y *recoger* aunque también aparecen otros negativos como *devolver, deportar* y *bloquear*, así como *ahogar* y *huir*. También se menciona la necesidad de *reubicarlos* y *reasentarlos*.

Un detalle que llama la atención al observar estos datos es la ausencia de la agencialidad de los refugiados. Estos huyen, llegan y -trágicamente- se ahogan. Por lo demás, son movidos como si de objetos sin voluntad se tratara.

Podemos e IU enmarcan el término en un contexto más dramático y negativo. En la pestaña de los sintagmas preposicionales con *de* (encabezada lógicament por *la crisis de los refugiados*), encontramos términos con una polaridad muy negativa como *deportación, expulsión, traficante, criminalización* y *éxodo*.

La plataforma nos ayuda a corroborar la poca importancia dada a los refugiados y también nos facilita concretar el estudio de los marcos y la comparación entre partidos.

Un tercer ejemplo de investigación lo extraemos del análisis sobre la representación de las minorías en el discurso político en Twitter[14]. El estudio mostró que, durante la campaña del 2015, las minorías apenas aparecían en el discurso político y, por lo tanto, tampoco sus reivindicaciones específicas. Su presencia en el corpus es apenas perceptible. En los 16.306 *tuits* encontramos los siguientes números: mujer y género (446 *tuits*); migrantes y refugiados (258); minorías

[14] Alcántara-Plá y Ruiz Sánchez (2017).

religiosas, libertad religiosa y de conciencia (18), centrados estos últimos casi exclusivamente en la defensa de la laicidad (14); colectivo LGTB (10); como etnia solo se citaba a la etnia gitana (4); discapacidad funcional (3); y mundo rural (17). Estas cifras muestran un carácter excluyente de los discursos políticos en Twitter ya que los colectivos citados sí estaban presentes en la mayoría de los programas electorales (Ruiz-Sánchez y Alcántara-Plá, en prensa). También prueban la escasa permeabilidad del discurso político digital en relación con la presencia de las minorías sociales y, por lo tanto, de la diversidad.

La búsqueda de términos propios de estas temáticas en MarcoPolo delata que no fue un rasgo exclusivo de aquella campaña electoral. Si buscamos, por ejemplo, palabras con la raíz *gitan-*, solo encontramos *gitano* y *gitanofobia*, que aparecen solo 25 y 5 veces respectivamente; *LGTB* (y *LGTBI*) aparecen 72 veces; discapacidad está 66 veces; rural 147 veces; etc.

CONCLUSIONES

Investigar con fondos público exige un esfuerzo suplementario de los investigadores en relación con la transferencia y el posible impacto social de resultados de investigación. El análisis del discurso político debería ser una disciplina de alto impacto social. Muestra de ello es el interés de los medios de comunicación en este tipo de estudios frente a las contribuciones sobre otras temáticas científicas. Desde el punto de vista ético, esto exige minimizar al máximo posibles sesgos ideológicos en el diseño de la investigación que puedan desvirtuar los resultados o que faciliten la manipulación de la opinión pública. Una de las maneras de evitarlo es que diferentes actores puedan validar las búsquedas realizadas para la extracción de resultados. El sistema siempre es mejorable, pero en su condición de plataforma abierta, MarcoPolo constituye por sí mismo una respuesta a esta cuestión ética y un medio de transferencia de resultados de investigación eficaz y accesible tanto a la comunidad científica, como a expertos en comunicación política y al público en general interesados en discurso político.

Objetividad, validez y replicabilidad son requisitos básicos de la investigación. La validez exige entre otros factores que investigadores ajenos/as puedan replicar el experimento en cuestión en las mismas condiciones, confirmen las conclusiones, mejoren los parámetros y, sobre todo, puedan rebatir los resultados. Los estudios de análisis del discurso

digital no deben ser ajenos a estos criterios de calidad, sobre todo teniendo en cuenta las posibilidades de manejo y manipulación que nos da trabajar con documentos digitales.

Nuestra plataforma de análisis de discurso político en Twitter es una respuesta experimental a estos retos metodológicos. Siendo una base de datos con más de 10.000 gráficos que permite navegar por un corpus de 117.010 *tuits,* MarcoPolo ofrece alternativas a varias cuestiones claves en la actualidad.

En primer lugar, constituye en sí misma una propuesta experimental de resolución a la cuestión ética sobre protección de datos y compilación de corpus digitales habitual en este campo. En segundo lugar, amplifica el citado impacto social de una investigación financiada con fondos públicos al poner a disposición de otros equipos no solo las conclusiones del estudio, sino un volumen de datos navegable susceptible de ser investigado desde nuevas perspectivas. En tercer lugar, la descripción técnica y el uso de software libre facilitan tanto su replicabilidad como su optimización colaborativa.

Con todo ello, MarcoPolo supone un experimento innovador desde la perspectiva de la investigación abierta. Hemos logrado realizar un estudio sobre análisis del discurso digital que permite no sólo conocer cómo es el comportamiento de partidos y líderes, sino que cada ciudadano/a pueda extraer sus propias conclusiones informadas directamente de los datos.

BIBLIOGRAFÍA

Alcántara-Plá, M. (en prensa). La función de las redes sociales en las campañas electorales. En B. Gallardo Paúls. *Mutaciones discursivas en el siglo XXI: la política en los medios y las redes*. Valencia, Tirant lo Blanch.

Alcántara-Plá, M. (2017). *Palabras invasoras. El español de las nuevas tecnologías.* Madrid: Catarata.

Alcántara-Plá, M. & Ruiz-Sánchez, A. (2017). The Framing of Muslims in the Spanish Internet. *Lodz Pragmatics* 13-2, *Special Issue On The Pragmatics Of Othering: Stereotyping, Hate Speech And Legitimising Violence*.

Alcántara-Plá, M. & Ruiz-Sánchez, A. (2018). Not for Twitter: Migration as a silenced topic in 2015 Spain General Election. En M. Schröter, & Ch. Taylor. *Exploring silence and absence in discourse.* Palgrave.

Anderson, J. R. (1983). *The architecture of cognition*. Cambridge, MA: Harvard University Press.

Chadwick, A., Dennis, J., & Smith, A. P. (2015). Politics in the age of hybrid media: Power, systems, and media logics. *The Routledge Companion to Social Media and Politics*, 7–22.

Hoey, M. (2005). *Lexical priming a new theory of words and language*. London/New York: Routledge.

Kilgarriff, A., & Renau, I. (2013). esTenTen, a vast web corpus of Peninsular and American Spanish. *Procedia-Social and Behavioral Sciences*, 95, 12-19.

Lang, A. (2000). The limited capacity model of mediated message processing. *Journal of Communication*, 50(1), 46-71.

McCarthy, D., Kilgarriff, A., Jakubíček, M., & Reddy, S. (2015). Semantic Word Sketches. *Corpus Linguistics* (CL2015).

Partington, A., Duguid, A. & Taylor, C. (2013). *Patterns and Meanings in Discourse*. Amsterdam: John Benjamins.

Ruiz-Sánchez, A. & Alcántara-Plá, M. (en prensa). Us vs. Them: Polarisation and populist M. A. discourses in the online electoral campaign in Spain. En E. Hidalgo-Tenorio, Benítez-Castro, & F. De Cesare, *Unravelling Populist Discourse. A Methodological Synergy*. Routledge.

Ruiz-Sánchez, A. & Alcántara-Plá, M. (en prensa). ¿Quién es el pueblo? La exclusión de las minorías en la campaña electoral 2015 en España. En F. Sullet-Nylander, M. B. Christophe Premat & M. Roitman (eds.) *Political discourses at the extremes. Expressions of populism in the Romance Speaking Countries*. Stockholm: Stockholm University Press.

Santanen, E., Briggs, R., & De Vreede, G. J. (2000). The cognitive network model of creativity: A new causal model of creativity and a new brainstorming technique. *Proceedings of the 33rd Hawaii International Conference on System Sciences*.

UNHCR (2016). *Global Trends. Forced Displacement in 2015*. Disponible en http://www.unhcr.org/576408cd7 (11-07-2018).