Geocorpus del español de las redes sociales y cartografía automática

ANTONIO RUIZ TINOCO

Universidad Sofía

Resumen: En este trabajo se trata la metodología del procesamiento de los mensajes de la red social *Twitter,* llamados tuits, y su aplicación en el estudio de la variación lingüística del español. Para ello, se presentan algunos ejemplos de variación de dicha red de un corpus previamente preparado de algo más de 10 millones de tuits georreferenciados recogidos durante varios meses de los años 2014 y 2015. Para obtener los datos brutos hemos utilizado el *Streaming* API ver1.1 de *Twitter* y de modo automático han sido almacenados en una base de datos MySQL. Además de un texto de un máximo de 140 caracteres, la base de datos contiene las coordenadas exactas del lugar desde donde se emitió el mensaje, así como algunos datos de los perfiles públicos de los usuarios. Los datos proceden de todo el mundo sin ninguna restricción geográfica, tanto de zonas urbanas como rurales. Las coordenadas de gran precisión que se obtienen de *Twitter* nos permiten preparar atlas lingüísticos no solamente con las distribuciones por países, regiones o ciudades, sino incluso por barrios o zonas muy reducidas si fuera necesario.

Debido al alto nivel de ruido de los datos, es necesario procesarlos de antemano eliminando tuits que contienen textos en otras lenguas, son ilegibles o no son apropiados para nuestro estudio por cualquier otra razón. Para la visualización de los datos y el tratamiento de la cartografía utilizamos QGIS ver. 2.10, software de código abierto de Sistema de Información Geográfica (SIG), que tiene la ventaja de relacionar los datos de texto con los geográficos y crear distintos tipos de mapas que muestran la distribución geográfica de los fenómenos estudiados.

Los datos obtenidos de *Twitter* son sumamente útiles y complementan los datos obtenidos por los medios tradicionales de trabajo de campo permitiéndonos observar visual y cuantitativamente la distribución y frecuencia de los fenómenos estudiados, por lo que resulta una metodología muy apropiada para el estudio de la variación geolingüística.

Palabras clave: variación léxica; español; corpus; *Twitter*; cartografía.

1. Sobre la naturaleza de los datos de las redes sociales

Es necesario aclarar que la recogida de datos para el estudio de la variación lingüística se puede realizar de varias maneras y que la recogida a través de Internet, ya sea de *blogs*, prensa o redes sociales, no pueden sustituir a la investigación tradicional por medio de encuestas, ya que éstas tienen muchas ventajas gracias al contacto directo del investigador con los informantes, cuyos atributos se pueden determinar de antemano según el objetivo de la investigación. No obstante, al hablarse la lengua española en países tan lejanos unos de otros, la distancia física es un gran inconveniente para reunir datos sincrónicos.

Entre los estudios de variación lingüística del español que se han llevado a cabo hay que destacar el proyecto de variación léxica del español del mundo, VARILEX[1], dirigido por Hiroto Ueda, o de VARIGRAMA[2] de variación gramatical, dirigido por Toshihiro Takagaki, proyectos de los que formo parte como uno de los coordinadores. Aunque los datos de estos proyectos se han ido publicando según se iban obteniendo, el período de obtención de los datos ha durado varios años. Por otro lado, los datos digitalizados procedentes de Internet se pueden preprocesar en un tiempo relativamente corto, incluso en tiempo real. Esta característica los hace particularmente útiles, además de la posibilidad de obtener una enorme cantidad de datos. Los datos proceden no de docenas de informantes sino de miles o millones. Además, las palabras contenidas en los corpus formados con los tuits georreferenciados pueden sobrepasar fácilmente 100 millones en un tiempo relativamente corto de unos tres meses utilizando un entorno informático estándar.

También es necesario tener en cuenta que los datos obtenidos de los tuits son de naturaleza distinta a los obtenidos en el trabajo de campo tradicional, ya que en las encuestas se puede limitar tanto el tipo de informantes como los temas a estudiar dando como resultado una colección de datos preparada para un tipo de análisis determinado de antemano. Sin embargo, la información contenida en los tuits es muy variada no solamente desde el punto de vista geográfico, sino de temas, estilos, etc. Afortunadamente, algunos de estos inconvenientes se pueden superar parcialmente mediante el uso apropiado de técnicas estadísticas y de muestreo de la enorme cantidad de datos obtenidos. En cuanto a la naturaleza de los datos de los tuits, hay que considerar que no es una forma de comunicación equivalente a la escrita ni a la hablada, sino una nueva forma de comunicación con características compartidas, pero totalmente nueva, por lo cual el estudio del lenguaje de las redes sociales tiene un atractivo lingüístico añadido.

Aunque existen muchas redes sociales como *Facebook*, *Google Plus*, *MySpace*, etc., en este estudio se ha escogido *Twitter* no solo por la facilidad técnica de la obtención de los datos, sino también por abarcar toda la extensión de los países hispanohablantes y por su formato unificado, JSON, muy apropiado para el procesamiento posterior.

El grado de difusión de cada una de las redes sociales en cada país es diferente y es necesario tenerlo en cuenta en las conclusiones, por lo que frecuentemente será necesario añadir otros tipos de datos, especialmente los de tipo demográfico. En general se obtienen más datos en países con mayor población y difusión de *Twitter*, como España, México, Argentina o Venezuela, y son escasos en las zonas andinas, Cuba o Paraguay. Aunque no se tratan en este estudio, los datos procedentes de países como Estados Unidos o Canadá contienen una gran proporción de mensajes bilingües por lo que hay que tenerlo en cuenta para la interpretación.

Sin querer ser contradictorios, y aunque en general es una ventaja poder disponer de enormes cantidades de datos, no podemos olvidar su calidad, ya que tienen un nivel de ruido muy alto, como es la presencia de publicidad, mensajes ininteligibles, mezclas con otras lenguas, faltas de ortografía, etc. Evidentemente, las variantes urbanas se obtienen con más facilidad que las rurales o arcaicas de baja frecuencia, de las que no siempre se obtienen ejemplos suficientes en un tiempo limitado.

[1] HIROTO UEDA Y ANTONIO RUIZ TINOCO, "VARILEX, "Variación léxica del español en el mundo, Proyecto internacional de investigación léxica", Iberoamericana Vervuert, Madrid / Frankfurt, 2003, en RAÚL ÁVILA, JOSÉ ANTONIO SAMPER, HIROTO UEDA ET ALLES, *Pautas y pistas en el análisis del léxico hispano(americano)*, Iberoamericana Vervuert, Madrid / Frankfurt, 2003, págs. 141-278.

[2] TOSHIHIRO TAKAGAKI, HIROTO UEDA, MASAMI MIYAMOTO, NORITAKA FUKUSHIMA Y ANTONIO RUIZ TINOCO, *Encuesta sobre problemas sintácticos de la lengua española*. Informe de investigación para el Ministerio de Educación. 2008.

2. Recolección de los datos

Aunque los mensajes de *Twitter* se pueden leer directamente en el monitor, para recoger grandes cantidades de mensajes lo más adecuado es aprovechar su *Streaming* API[3]. Hay tantas formas de programar la obtención como tipos de lenguajes existentes de programación. En este caso, se han utilizado principalmente los lenguajes PHP y JavaScript debido sobre todo a la familiaridad con estos lenguajes en proyectos previos. Los datos se almacenan automáticamente en una base de datos relacional MySQL[4], de la que podemos extraer los datos que cumplan las condiciones necesarias para cada objetivo y finalmente son tratados con software de análisis espacial de código abierto como QGIS[5] para la preparación de varios tipos de mapas o atlas lingüísticos. Por este método se puede obtener el 1% de la totalidad de los tuits que *Twitter* facilita de forma gratuita, de los cuales están georreferenciados solamente alrededor del 3% según el país, ya que estos tuits proceden en su mayoría de teléfonos móviles con la función GPS activada. Para este estudio se ha usado un corpus de unos 100 millones de palabras, que se recogieron en un período aproximado de tres meses utilizando un servidor privado virtual (VPS) conectado al API ininterrumpidamente 24 horas diarias. Los detalles técnicos se pueden consultar en línea así como en varias obras publicadas como Christopher Ho and Bess Peri[6] y Matthew A. Russell[7]. El resultado de las búsquedas se obtiene en formato JSON, muy fácil de procesar y del que se puede encontrar la información técnica necesaria fácilmente.

2.1. ¿Qué información contiene un tuit?

Cada tuit contiene además del mensaje en sí limitado a 140 caracteres, otros datos de gran utilidad para el estudio de la variación lingüística, como son las coordenadas geográficas del lugar de procedencia. En la figura 1 se muestra la estructura de un tuit, como número de seguidores, veces que ha sido leído, veces que ha sido retuiteado, lengua, descripción del perfil del usuario, etc.

Fig. 1 Estructura parcial de un tuit

De esta manera se prepara el geocorpus, es decir, un corpus que contiene información espacial de los lugares de donde proceden. Una vez creado el corpus, se pueden hacer búsquedas a través del conocido software de código abierto PhpMyAdmin[8] o Adminer[9], ambos muy fáciles de instalar y usar. En la Fig. 2 se puede observar el resultado parcial de la búsqueda del término *poroto*. El resultado se

[3] https://dev.twitter.com/overview/documentation [10/01/2016]
[4] http://www.mysql.com [10/01/2016]
[5] http://qgis.org/es/site/ [10/01/2016]
[6] CHRISTOPHER HO AND BESS PERI, *Sams teach yourself Twitter API in 24 hours.* Pearson Education, Inc. 2011
[7] MATTHEW A. RUSSELL, Mining the Social Web: Data Mining Facebook, Twitter, Linkedln, Google+, GitHub, and More, Second Edition, O'Reilly Media, Inc. 2013
[8] https://www.phpmyadmin.net [10/01/2016]
[9] https://www.adminer.org [10/01/2016]

puede exportar a varios formatos como SQL, EXCEL o CVS para cualquier procesamiento posterior.

La base de datos contiene los mensajes originales sin reformar. Sin embargo, según el análisis que se quiera hacer, es necesario limpiarlos del inevitable ruido mencionado anteriormente.

Fig. 2 Resultado parcial de la búsqueda del término poroto.

tweet_text	created_at	geo_lat	geo_long
El pato criollo un poroto al lado mio media pelia eri	2015-10-15 20:22:59	-34.41014	-58.61285
Cuando me habla me molesta , y cuando no me habla también me molesta La gata flora un poroto al la...	2015-10-15 21:13:00	-34.89978	-57.98931
Vivir sola & comer 3 huevos fritos con Pan. Master chef un poroto.	2015-10-16 13:32:00	-34.45580	-58.57694
#Picasso un poroto, by Agustín Monzón	2015-10-16 15:03:19	-38.00000	-57.55000
I'm at El Palacio Del Poroto Con Riendas in Estación Central, Santiago Metropolitan Region https://t...	2015-10-17 13:51:34	-33.46003	-70.69598
@MiluDhondt @GuidoRosano Luisiti de cdp. Los dos un poroto.. Se comenta q a su casamiento lo hacen e...	2015-10-17 14:04:36	-32.90422	-60.90742
@agustinadevivo un poroto ese	2015-10-17 15:13:28	-34.63931	-58.37492
Valu Ramallo un poroto. Y con los amargos de la Peco niiiiii te... https://t.co/2TX3XOsB8P	2015-10-17 18:48:28	-32.95153	-60.64822
Con joni y poroto	2015-10-17 19:34:39	-33.01373	-58.55139
El coco basile un poroto al lado de mi prima jajajjaj	2015-10-18 17:34:08	-31.59192	-59.89131
@Camiilariadna como olvidar esa noche? Jajajaja los mineros un poroto,los quieroo	2015-10-18 19:14:52	-34.80018	-58.21428
Planté un poroto.. Y ahora se está convirtiendo en la primera planta de nuestro hogar..	2015-10-18 23:17:24	-33.43363	-70.65831
"@PrestaPrestico: El FBI un poroto al lado de una mina celosa." @MelinaFagliani nosotras somos mejor...	2015-10-19 03:30:54	-32.39281	-63.24969
Leche barbara un poroto jajaja	2015-10-19 04:05:22	-34.66517	-58.44575
"@NicolasssTm: El FBI un poroto al lado de una mina celosa." Jajajaja	2015-10-19 08:21:09	-32.97764	-60.65242
Bombón ! #Poroto @ Córdoba - Nueva Córdoba https://t.co/mzvyzp8rDm	2015-10-19 15:04:05	-31.42513	-64.18030
Y salió eso , #masterchef un poroto	2015-10-19 20:30:37	-34.65000	-58.58330
Me voy a morir de la tos que tengo, el pinguino de Toy Story un poroto al lado mio	2015-10-20 10:22:42	-41.14423	-71.26183
@AgustinDiaz88 la profe un poroto al lado tuyo	2015-10-20 12:20:59	-37.00225	-57.12177
Listo ya limpie todo Isaura la esclava un poroto al lado mío, ahora si , me fui :D	2015-10-20 13:30:29	-34.66793	-58.47340
Cupido un poroto	2015-10-20 15:29:33	-38.92254	-67.98016

Según las condiciones[10] de uso de *Twitter*, los datos no se pueden distribuir ni compartir libremente, lo cual puede dificultar la reproducción de los resultados por terceras partes. No obstante, hay varios métodos permitidos con algunas condiciones para compartir el mismo corpus como es ofrecer la lista de los códigos de identificación de los tuits y descargarlos por un método relativamente sencillo. Este método, aunque simple, puede ser otro obstáculo para la reproducción de los experimentos ya que para la descarga de los tuits basándose en los códigos de identificación es necesario un entorno informático similar al necesario para descargar todos los datos y metadatos de los tuits. Además, con el paso del tiempo, los usuarios pueden borrar los mensajes antiguos, por lo que el corpus en sí puede variar ligeramente. Afortunadamente, las bases de datos de *Twitter* así formadas pero protegidas mediante contraseñas parecen no contravenir las condiciones de uso. Pueden hacerse accesibles por Internet, por lo que resultan muy adecuadas para la investigación en equipo, incluso desde lugares geográficos diferentes.

3. Cartografiado de los tuits

Una de las características más importantes de los tuits georreferenciados es precisamente que contienen las coordenadas geográficas del lugar de procedencia con gran exactitud. Gracias a la tecnología SIG (Sistema de Información Geográfica, o GIS por sus siglas en inglés), es posible preparar atlas lingüísticos para mostrar los datos de múltiples formas. A continuación se muestran los datos de varios términos usando el software QGIS[11] de información geográfica de código abierto. Las técnicas SIG no solamente permiten

[10] https://twitter.com/tos?lang=es
[11] http://qgis.org/es/site/

el cartografiado digital de los datos con diferentes modelos de visualización, sino que se trata realmente de una base de datos que a su vez se puede complementar con otros tipos de datos en prácticamente cualquier formato. Para los datos geográficos hay varios tipos de formatos, como geojson, shapefiles, xml, kml, csv, etc. Además, casi todos los formatos utilizados en bases de datos, pueden utilizarse en los sistemas SIG, ya sea directamente o convirtiendo previamente los formatos. A continuación, se muestran varios ejemplos básicos de los mapas lingüísticos que se pueden preparar con una fracción de los tuits obtenidos, señalando algunas ventajas y desventajas de cada método.

3.1 Mapas de puntos

La Fig. 3 muestra un mapa de los puntos de procedencia de una muestra de 5.478.227 tuits. Como se puede apreciar, en esta muestra que no ha sido preprocesada, la enorme cantidad de puntos se reparte por casi toda la geografía mundial y no hay diferencia visual entre los países hispanohablantes y extensas zonas de Europa y de otras partes de África y Asia, como India, Japón, extensas zonas del Sudoeste Asiático, Australia, etc. Además, puede parecer que en Norteamérica hay más datos que los que realmente hay debido precisamente a la superposición de puntos que no refleja la realidad con exactitud. También se pueden apreciar puntos en el mar debido a que una mínima proporción de tuits contiene errores en los datos de sus coordenadas. Un mapa de este tipo muestra que, efectivamente, la lengua española es de características globales, pero no es el método más adecuado para ver los detalles de la variación lingüística. En cambio, si nos representamos un término como *bondi*, un tipo de autobús utilizado en Argentina o la expresión *un poroto al lado mío*, también de Argentina, mostrados en las figuras 4 y 5 respectivamente se puede apreciar su distribución con mayor claridad. Asimismo, se puede advertir también que, aunque el término *poroto* se usa en Chile y otros países cercanos, la expresión *un poroto al lado mío* es típica de Argentina, apareciendo también en Montevideo y zonas cercanas. No obstante, la acumulación de puntos en una zona reducida como es Río de la Plata dificulta la visualización d los datos.

Fig. 3 Mapa de puntos de 5 millones de tuits en español

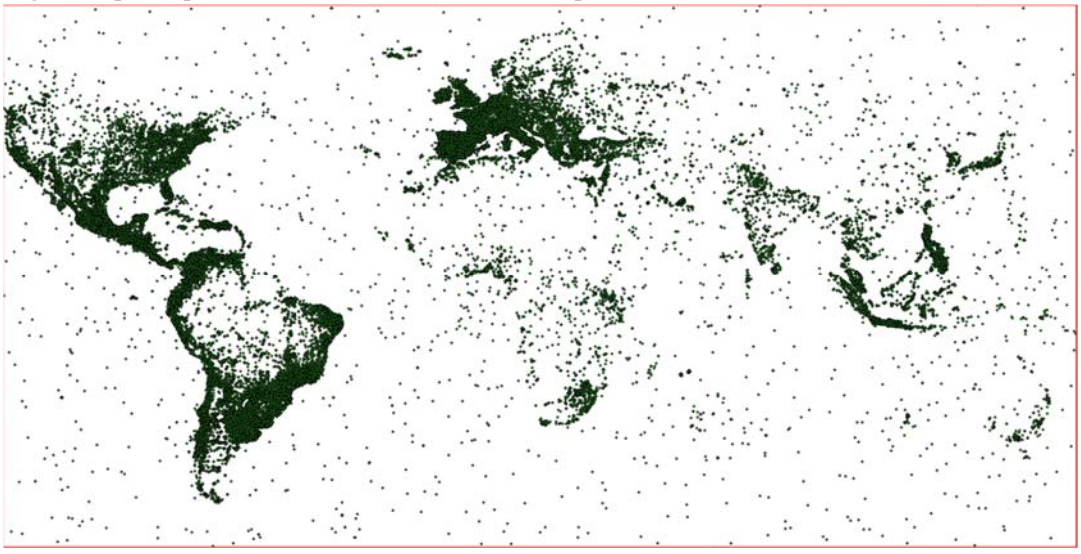


Fig. 4 bondi

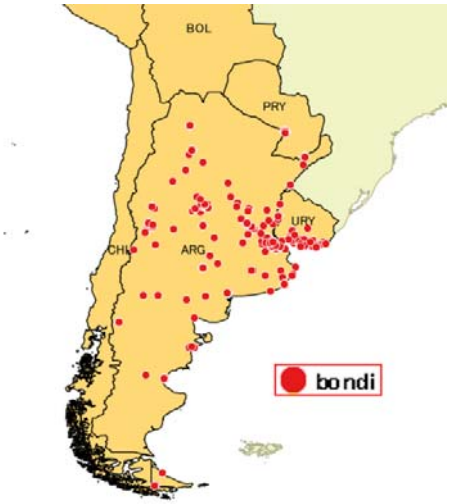


Fig. 5 un poroto al lado mío

El mapa de puntos resulta adecuado para constatar la presencia o no de un término o una expresión, tener una idea aproximada de la extensión del fenómeno o como paso previo a una investigación más detallada. Por ejemplo, las figuras 6 y 7 son una muestra pequeña de dos formas de diminutivo de *pueblo*: *pueblito,* muy frecuente en el español americano, aunque también se aprecia en España, y *pueblecito,* predominantemente peninsular.

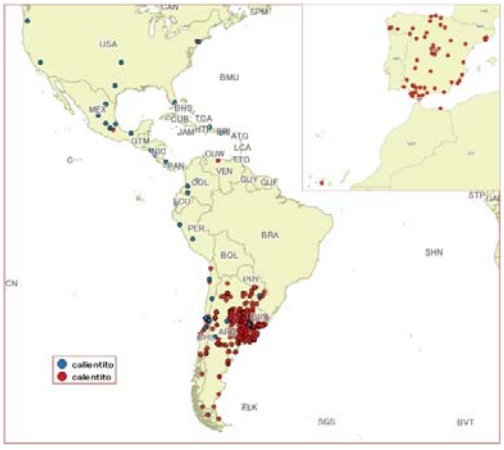
Fig. 6 pueblito

Fig. 7 pueblecito

En la figura 8 se puede apreciar con pocos ejemplos la distribución de *pueblín,* sobre todo por zonas septentrionales de la península. En la figura 9, con dos capas superpuestas, una para *calentito* y otra para *calientito*, se puede observar que la forma *calientito* no se refleja en los datos de España.

Fig. 8 pueblín

Fig. 9 calentito vs calientito



De esta manera, el mapa más sencillo compuesto solamente de puntos, puede ser muy útil para estudios previos por su facilidad de uso o para término o expresiones de baja frecuencia y constatar su aparición.

3.2 Mapas temáticos

Los mapas temáticos, también conocidos como *choropleths* en inglés, se basan en algún tipo de cuantificación de los atributos en una región determinada. Por ejemplo, en la figura 10 se han calculado las proporciones de uso de las formas *hiciste* e *hicistes.* Aunque la forma *hicistes* es de menor frecuencia, se aprecia que prácticamente en todas las zonas hispanohablantes coexiste llegando a casi un 20% del total en Puerto Rico y parte de Centroamérica. Los países en los que no se encontraron suficientes ejemplos para calcular la proporción van marcados con 0% . Este problema surge en los casos de baja frecuencia, que solamente se pueden resolver con más datos hasta llegar a cifras representativas que no se deban al azar.

Fig. 10 *hiciste* vs *hicistes*

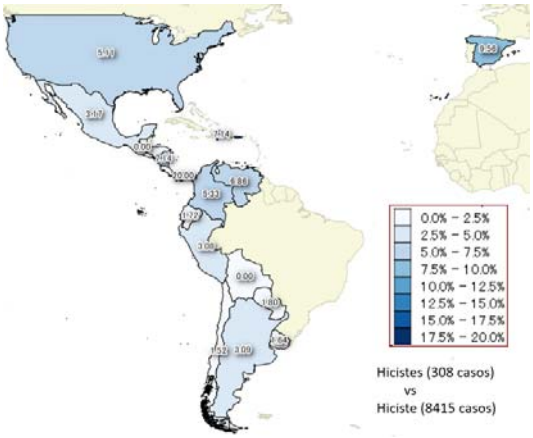
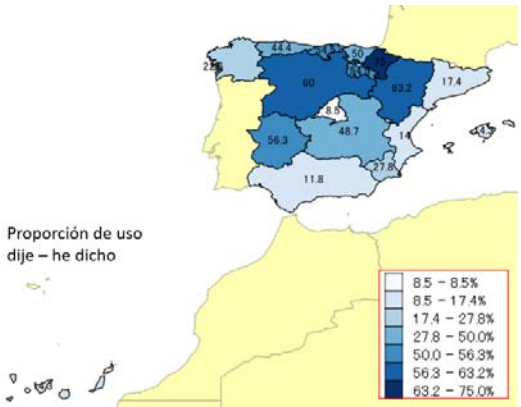


Fig. 11 *dije* vs *he dicho*



De la misma manera, la figura 11 muestra las proporciones de uso de *dije* en relación con *he dicho.* En la figura 10 se muestran los resultados de las ocurrencias por países, y en la figura 11 se muestran por comunidades autónomas de España. Aunque este tipo de mapa resulta algo más laborioso de preparar, no

ofrece ninguna dificultad técnica y son visualmente muy intuitivos. El principal problema que puede surgir es que no haya suficientes datos en alguna zona del mapa, como es el caso de Bolivia en la figura 10. De nuevo, el aumento del tamaño de la muestra de ejemplos puede resolver este problema. Además, se pueden utilizar varias escalas de valores, desde las lineales más simples, hasta otras más complejas como las llamadas "Natural Breaks (Jenks)", basadas en un tipo de análisis *cluster*. Hay que tener en cuenta que cada una de estas formas puede tener varias condiciones de uso y es preciso analizarlas por separado.

Ya sea por puntos o por medio de mapas temáticos, al cambiar los parámetros que se muestran en los mapas, podemos apreciar distintas particularidades del fenómeno en sí. Por ejemplo, la figura 12 muestra simplemente la cantidad de tuits en catalán en la península por comarcas. En el centro de la península se distingue Madrid por los números absolutos de tuits, posiblemente escritos por catalanes residentes. Sin embargo, si calculamos las proporciones de tuits escritos en catalán en relación con los escritos en castellano, obtenemos un mapa como el de la figura 13 en el que no se aprecia un valor destacado en Madrid. También apreciamos en la figura 12 que en la zona alrededor de la ciudad de Barcelona y zonas costeras hay un número alto de tuits, pero las proporciones respecto al castellano pueden ser incluso más altas en algunas comarcas del interior.

Fig. 12 Tuits en catalán

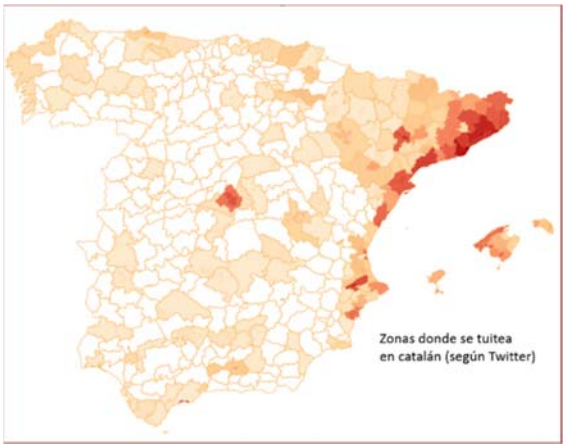


Fig. 13 Proporción de tuits CAT/ESP

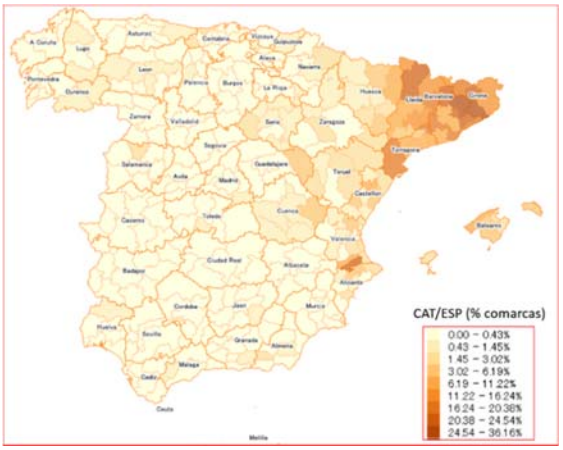
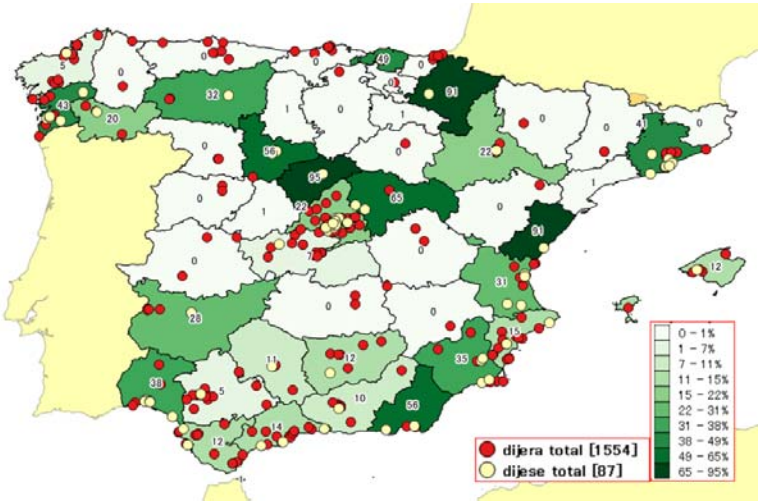


Fig. 14 Mapa combinado de puntos (2 capas) y temático de dijera vs dijese



En la figura 14 se muestran las ocurrencias de las formas *dijera* y *dijese* de forma combinada, usando un mapa temático y los puntos concretos de las ocurrencias superpuestos en dos capas de colores diferentes para cada una de las formas. Teóricamente se pueden superponer un número ilimitado de capas y como método de visualización puede ser adecuado para algunos fenómenos siempre que no se concentren demasiados datos en un espacio reducido.

3.3 Mapas de calor

Los mapas de calor, también conocidos como *heatmaps* en inglés, representan mediante una escala graduada de colores los distintos valores calculados para los atributos. Se suele emplear el azul para los valores bajos y el rojo para los valores altos, aunque se usan muchas otras combinaciones intuitivas.

Por ejemplo, en los mapas de las figuras 15 y 16 se muestran los mapas de calor de la palabra *friki.* Se puede observar, incluso teniendo en cuenta la población de dichas zonas, que la densidad de uso es bastante alta en Río de la Plata, aunque no en Santiago de Chile, y dentro de España se usa frecuentemente en Madrid, pero no tanto en Barcelona.

Fig. 15 friki (Río de la Plata)

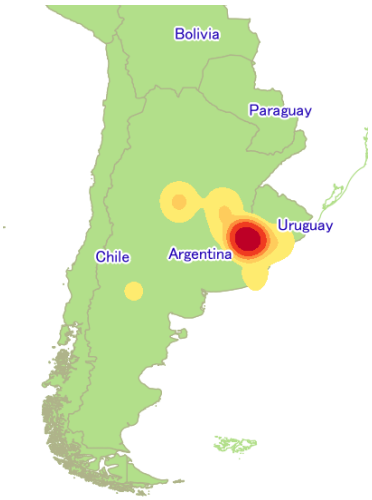
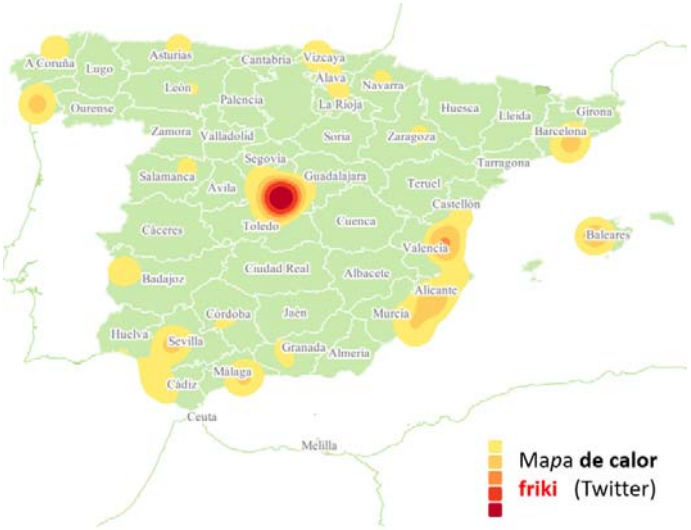


Fig. 16 friki (España)

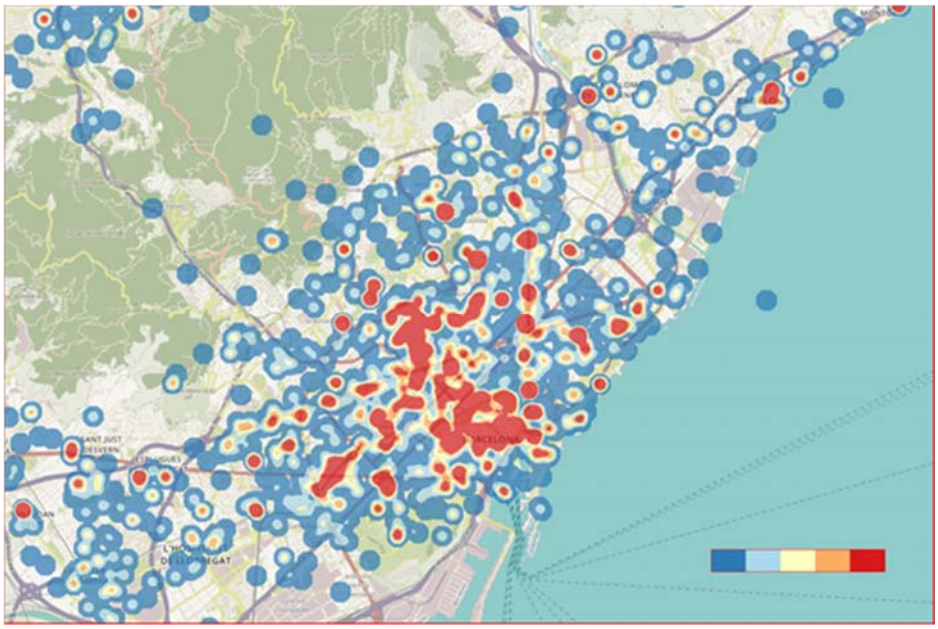


En la figura 17 el mapa de calor muestra la densidad de tuits en catalán en Barcelona y zonas de alrededor usando la escala típica de azul a rojo.

4. Conclusiones

Twitter es una valiosa fuente de datos de uso del español en todas sus variedades y zonas geográficas. Posee características del lenguaje escrito y hablado al mismo tiempo, además de otras formas de comunicación diferentes, como iconos o abreviaciones particulares.

Fig. 17 Densidad de tuits en catalán en Barcelona



El hecho en sí de ser usado por millones de hispanohablantes lo hace objeto de estudio lingüístico detallado, y especialmente resulta útil para investigar la variación lingüística sincrónica. Por otra parte, al conocer las coordenadas de cada uno de los tuits, los datos se pueden procesar con técnicas SIG y preparar mapas fácilmente. En la actualidad, este geocorpus va creciendo cada día y los análisis podrán ser más detallados.

Referencias bibliográficas.

Davies, M. , *Corpus del Español: 100 million words, 1200s-1900s*. http://www.corpusdelespanol.org, 2002-, [10/09/2015]

García Mouton, P. (ED.), *El español de América*. Consejo Superior de Investigaciones Científicas. Madrid. 2003.

Ho, Christopher And Peri, Bess, *Sams teach yourself Twitter API in 24 hours.* Pearson Education, Inc. 2011.

Oficina Nacional de Estadísticas. Anuario Estadístico de Cuba, 2011. Edición 2012. <http://www.one.cu/aec2011/esp/15_tabla_cuadro.htm> [10/09/2015]

Real Academia Española, *Diccionario panhispánico de dudas*. Madrid, Santillana. 2005.

Real Academia Española, Asociación de Academias de la Lengua Española, *Nueva Gramática de la Lengua Española: morfología y sintaxis*. Madrid, Espasa. 2010.

Real Academia Española, Banco de datos (CREA) [en línea]. *Corpus de referencia del español actual.* http://www.rae.es, [10/09/2015]

Ruiz Tinoco, A., "Twitter como corpus para estudios de geolingüística del español". En *Sophia Lingüística LX*, págs. 147-163, The Graduate School of Languages and Linguistics, Linguistic Institute for International Communication, Sophia University, Tokyo, 2013.

Russell, Matthew A., Mining the Social Web: Data Mining Facebook, Twitter, Linkedln, Google+, GitHub, and More, Second Edition, O'Reilly Media, Inc. 2013.

Takagaki, Toshihiro; Ueda, Hiroto; Miyamoto, Masami; Fukushima, Noritaka; Ruiz Tinoco, Antonio, *Encuesta sobre problemas sintácticos de la lengua española.* Ministerio de Educación, Cultura y Deporte de España. 2004

Vaquero de Ramírez, María, *El español de América II. Morfosintaxis y léxico.* Madrid. Arco/Libros, 1996.