# Creación de un corpus de noticias de gran tamaño en español para el análisis diacrónico y diatópico del uso del lenguaje

## *Creation of a large news corpus in Spanish for the diachronic and diatopic analysis of the use of language*

**Nombre Apellidos1**[1], **Nombre Apellidos2**[2]
[1]Universidad o lugar de trabajo
[2]Universidad o lugar de trabajo
Información de contacto

**Resumen:** Este artículo describe el proceso llevado a cabo para desarrollar un corpus de noticias periodísticas de gran tamaño en español. Todos los textos recopilados están ubicados tanto temporal como geográficamente. Esto lo convierte en un recurso de gran utilidad para trabajos en el ámbito de la lingüística, la sociología y el periodismo de datos, permitiendo tanto el estudio diacrónico y diatópico del uso del lenguaje como el seguimiento de la evolución de determinados eventos. El corpus se puede descargar libremente empleando el software que se ha desarrollado como parte de este trabajo. El artículo se completa con un análisis estadístico del corpus y con la presentación de dos casos de estudio que muestran su potencial a la hora de analizar sucesos.
**Palabras clave:** corpus, minería de texto, análisis diacrónico, análisis diatópico

**Abstract:** This article describes the process carried out to develop a large corpus of news stories in Spanish. The collected texts are located both temporally and geographically. This makes it a very useful resource to work with in the field of linguistics, sociology and data journalism, allowing the diachronic and diatopic study of the use of language and tracking the evolution of specific events. The corpus can be freely downloaded using the software developed as part of this work. The article includes a statistical analysis of the corpus and two case studies that show its potential for event analysis.
**Keywords:** corpus, text mining, diachronic analysis, diatopic analysis

### *1 Introducción*

Las noticias periodísticas cada vez se producen y consumen de manera más frecuente a través de Internet, favoreciendo la existencia de grandes volúmenes de este tipo de textos en formato digital. El análisis computacional de estos corpus de noticias puede llevar al descubrimiento de interesantes hallazgos desde un punto de vista tanto sociológico como lingüístico (Leetaru, 2011).

Este artículo presenta el desarrollo de un corpus de noticias en español de gran tamaño ubicadas tanto geográfica (lugar en el que se produjeron) como temporalmente (momento en el que tuvieron lugar). El objetivo de este corpus es el de servir de fuente para estudios en áreas como la lingüística, la sociología y el periodismo de datos, permitiendo un análisis diacrónico (evolución en el tiempo) y diatópico (evolución en el espacio) del uso del lenguaje. Además de describir las características del corpus y el proceso llevado a cabo para su obtención, este trabajo presenta ejemplos de casos de estudio centrados en su explotación.

La fuente de información empleada para la construcción de este corpus ha sido el periódico gratuito de información general *20 minutos*,[1] el cual ofrece en formato digital todas sus noticias desde enero de 2005 hasta la actualidad. Este periódico contiene secciones locales para todas y cada una de las provincias de España, lo que permite tener localizadas geográficamente cada una de las noticias. Pa-

[1] `https://www.20minutos.es`

ra el desarrollo de este corpus se han extraído todas las noticias publicadas en la sección local de cada una de las cincuenta provincias y las dos ciudades autónomas de Ceuta y Melilla, dando lugar a un corpus de casi dos millones de noticias publicadas a lo largo de los últimos trece años.

Además del contenido original de las noticias, se ha realizado un análisis lingüístico del texto incorporando información sobre lemas, entidades y palabras con contenido semántico, codificando toda esta información en formato JSON para facilitar su posterior procesamiento. Aunque el corpus no puede dejarse libremente disponible para descarga por los derechos de uso de *20 minutos*, se proporciona el software necesario para que cualquier investigador pueda replicar este corpus en su ordenador de manera sencilla.[2]

La gran ventaja que ofrece este corpus frente a otros existentes es la localización geográfica de los artículos, lo que permite realizar análisis diatópicos del lenguaje y estudiar los rasgos dialectales de distintas regiones. Hasta donde tenemos conocimiento, no existe en la actualidad un corpus de noticias con estas características libremente disponible.

El resto del artículo se estructura como sigue: la Sección 2 presenta diversos trabajos en el ámbito del desarrollo de corpus periodísticos; la Sección 3 describe el proceso llevado a cabo para la generación del corpus; en la Sección 4 se presentan distintas estadísticas extraídas del corpus y detalles sobre el uso del lenguaje a nivel geográfico y temporal; en la Sección 5 se ofrece un análisis cuantitativo de dos sucesos explotando la información contenida en el corpus; finalmente, la Sección 6 muestra las conclusiones y posibles trabajos futuros.

## 2 Trabajo relacionado

Existen diferentes estudios y proyectos que tratan sobre la construcción de corpus de textos periodísticos para su posterior análisis lingüístico. En esta sección nos vamos a centrar en revisar los trabajos realizados en idioma español y, por tanto, más afines con el corpus descrito en este artículo.

El primero de estos corpus es el recopilado en el proyecto Aracne[3], donde se presenta un estudio sobre la variación de la riqueza lingüística en la prensa española desde 1914 hasta 2014. Sobre este corpus de noticias se realizó un procesamiento lingüístico de los textos midiendo rasgos de variación léxica, densidad y complejidad de los textos. El resultado de este proyecto fue un corpus de 5.167 artículos y 1.921.566 de palabras. El corpus no está disponible para su estudio.

Otro trabajo en esta línea es el corpus *Spanish News Text* (Graff y Gallegos, 1995), compuesto de textos periodísticos extraídos de diferentes periódicos y agencias de noticias de hispanoamérica entre el año 1995 y 1996. El corpus cuenta con 170 millones de palabras y está disponible para los miembros del *Linguistic Data Consortium*[4] (LDC). Existe una versión posterior del corpus, el *Spanish Newswire Text, Volume 2* (Graff y Gallegos, 1999), que recoge noticias entre los años 1996 y 1998. Esta versión está disponible para el público en general previo pago.

El corpus *Timestamped JSI web*[5] está formado por artículos de noticias obtenidos de distintos servicios RSS a nivel mundial. Contiene textos desde el año 2014 hasta la actualidad en diferentes idiomas, entre ellos el español. El corpus está accesible para usuarios suscritos a la plataforma *Sketchengine* y sólo puede consultarse dentro de ésta, ofreciendo funcionalidades para análisis del lenguaje como la búsqueda de sinónimos, ejemplos de uso, frecuencia de palabras o identificación de neologismos.

Otro corpus relevante es el *Corpus del Español NOW (News on the Web)*[6] que contiene cerca de 5.700 millones de palabras obtenidos de periódicos digitales y revistas de 21 países de habla hispana, desde el año 2012 hasta la actualidad. Al igual que el corpus anterior, su acceso está limitado a la plataforma en la que se oferta, en la que se pueden realizar diferentes consultas a través de una interfaz.

Finalmente, *Molino Labs*[7] presenta un corpus formado a partir de artículos de prensa de España, Argentina y México, producidas entre los años 1997 y 2009. Cuenta con más de 1.700 millones de artículos y 660 millones de palabras. El corpus se puede consultar a través de una interfaz web, pero no

[2] https://github.com/analisis-20minutos/herramientas-analisis

[3] https://www.fundeu.es/aracne

[4] https://www.ldc.upenn.edu

[5] https://www.sketchengine.eu/jozef-stefan-institute-newsfeed-corpus

[6] https://www.corpusdelespanol.org

[7] http://www.molinolabs.com/corpus.html

está disponible para descarga. Esta interfaz ofrece, entre otras, la posibilidad de buscar palabras de una longitud fija que empiezan por determinados caracteres o localizar palabras que aparecen en compañía de otras.

Si bien algunos de estos trabajos permiten hacer un estudio diacrónico del corpus, a diferencia de nuestra propuesta ninguno de ellos ofrece la posibilidad de hacer un estudio de los textos a nivel diatópico. Adicionalmente, existe el problema del acceso a los documentos. La mayoría de las propuestas estudiadas sólo permiten analizar el corpus a través de una interfaz web con funciones limitadas, mientras que otras ofrecen su descarga solo a suscriptores o previo pago. En el caso de nuestro corpus, aunque por limitaciones de derechos de uso no se puede dejar libremente disponible para descarga, se ha publicado el software desarrollado para que cualquiera pueda descargarlo y replicarlo en su ordenador. Finalmente, otro de los puntos fuertes de nuestro corpus frente a otras propuestas es el volumen de noticias que presenta, la información de análisis lingüístico incluida y el contar con noticias actuales (ver Sección 4).

## *3 Creación del corpus*

En esta sección se describe todo el proceso de obtención del corpus. En primer lugar se describirá el proceso de descarga y limpieza de las noticias. A continuación se expondrá el proceso llevado a cabo para la eliminación de duplicados y casi duplicados. Finalmente, se describe el análisis lingüístico llevado a cabo sobre los documentos.

### 3.1 Obtención de noticias

La primera tarea llevada a cabo para la creación del corpus fue la obtención de las noticias en formato digital de la hemeroteca del diario *20 minutos*. Las noticias están accesibles a través de la sección *Archivo*[8] del diario donde se encuentran agrupadas por días. Se puede acceder a la página web de cada una de ellas pinchando en el correspondiente enlace, pero no existe una forma directa de descargar el contenido de la noticia.

Para poder obtener el texto limpio de las noticias se tuvo que desarrollar un programa para la extracción de datos de las páginas web (*web scraping*). El objetivo era obtener el título, resumen y cuerpo de las noticias, eliminando todas las etiquetas HTML y contenidos adicionales que se muestran en la página del diario (ej. menús, anuncios, noticias relacionadas e imágenes). Para dicha tarea se utilizó la herramienta *Scrapy*,[9] que permite extraer elementos de las páginas web mediante la definición de selectores CSS.

La principal dificultad de esta tarea fue la falta de homogeneidad en la estructura HTML de las páginas. Ésta variaba en función de los años y de algunas localizaciones, por lo que se tuvieron que realizar ajustes en la herramienta para contemplar estas variaciones. Se revisaron manualmente y de manera sistemática subconjuntos de noticias en las distintas localizaciones y años para comprobar que la obtención del contenido de todas ellas fuera correcto.

Como resultado de este proceso se obtuvo un volcado completo de las noticias del portal desde el 17 de enero de 2005, primer día del que se tiene registro, hasta el 5 de julio de 2018, momento en el que se finalizó la recolección de datos. Para cada noticia se almacenó en formato JSON, tal y como se puede ver en la Figura 1, su localidad (`province`), fecha en formato ISO 8601 (`date`), URL de la página de donde fue extraída (`url`), título (`title`), resumen (`lead`) y cuerpo (`body`). Para estos tres últimos campos, además del texto original (`raw_text`), se incluyó una serie de información adicional resultado del análisis lingüístico realizado, tal y como se describe en la Sección 3.3.

En total se obtuvieron 2.215.078 artículos de 52 localizaciones diferentes: las cincuenta provincias de España y las dos ciudades autónomas de Ceuta y Melilla.

### 3.2 Eliminación de duplicados

Dado que *20 minutos* cuenta con numerosas secciones locales, era de esperar que algunas noticias pudieran estar duplicadas o casi duplicadas entre distintas provincias. Por ejemplo, en ocasiones se usan noticias "plantilla" en las que sólo cambian los datos específicos correspondientes a cada localidad. Es el caso de eventos como "La fiesta del cine", donde lo único que varía de una provincia a otra es el número de asistentes. Otros ejemplos de noticias casi duplicadas son aquellas que presentan actualizaciones sobre una noticia anterior. Dentro de este grupo entran casos como el de las crecidas del Ebro, donde cada día se publican datos actualizados

[8] `https://www.20minutos.es/archivo`

[9] `https://scrapy.org`

```
{
  "province": "SEVILLA",
  "date": "2005-01-28T00:00:00",
  "url": "https://www.20minutos.es/noticia/1994/0/sevilla/bajo/cero/",
  "title": {
    "raw_text": "Y en Sevilla, bajo cero",
    "lemmatized_text": "y en sevilla bajo 0",
    "lemmatized_text_reduced": "sevilla",
    "persons": [], "locations": ["Sevilla"], "organizations": [], "dates": [],
    "numbers": ["0"], "others": []
  },
  "lead": {
    ...
  },
  "body": {
    ...
  }
}
```


Figura 1: Ejemplo de noticia del corpus en formato JSON. Los campos contenidos en `title` se encuentran también en `lead` y `body`. Se omiten aquí por motivos de espacio

utilizando el mismo cuerpo de noticia. Finalmente, también hay noticias de interés global que simplemente se duplican literalmente entre distintas localidades.

Para evitar textos repetidos que puedan afectar al análisis estadístico del corpus, se desarrolló un programa para la eliminación de duplicados y casi duplicados. El problema inicial de este proceso era la imposibilidad de cotejar cada noticia del corpus con todas las demás, ya que implicaba más de cuatro billones de comparaciones. Una de las posibilidades que se planteó para reducir este número fue la de comparar noticias sólo entre provincias cercanas. Sin embargo, tras un análisis realizado sobre un subconjunto de ellas, se comprobó que existían noticias duplicadas en localidades muy distantes. Lo que sí permitió este estudio preliminar fue identificar un patrón temporal en las noticias duplicadas: todas ellas se daban en el mismo día o en días muy próximos. Para dar margen suficiente, se decidió comparar las noticias en bloques de sesenta días estableciendo un solapamiento de seis días entre bloques consecutivos.

Para acelerar aún más el proceso de comparación entre pares de noticias, se utilizó una estructura de tipo *MinHash* (Broder, 1997) que permitía calcular la similitud de Jaccard (Leskovec, Rajaraman, y Ullman, 2014) entre dos textos en un tiempo lineal con respecto al tamaño del conjunto de documentos a comparar. En la implementación se empleó *Locality-Sensitive Hashing* (LSH) (Indyk y Motwani, 1998) para optimizar el proceso de comparación entre documentos. De esta manera se consiguió un algoritmo de complejidad $O(n \cdot \sqrt{n})$ para la búsqueda de duplicados, reduciendo la complejidad de tipo $O(n^3)$ que hubiera tenido de no haber aplicado estas optimizaciones.

Una vez establecido el procedimiento de comparación, era necesario determinar cuándo dos noticias se iban a considerar como duplicadas. Se probaron distintos porcentajes de solapamiento a nivel de palabra, comprobando el número de noticias que eran consideradas como duplicadas. Finalmente este umbral de solapamiento se estableció en un 70 %, ya que con este valor se alcanzaba estabilidad en el número de noticias eliminadas, y para valores menores se consideraba que podía llevar a la obtención de falsos duplicados. El número total de noticias que quedó como resultado de este proceso fue de 1.826.985, eliminando en total casi cuatrocientas mil noticias.

### 3.3 Procesamiento lingüístico

Con el objetivo de enriquecer el corpus con información lingüística que permitiera un análisis más profundo de los textos, se procedió a la extracción de distintas características de éste para incorporarlas a la estructura JSON desarrollada. Concretamente, se extrajeron los siguientes elementos de cada noticia para el título, cuerpo y resumen (ver ejemplo en la Figura 1):

- Texto lematizado, sin signos de puntua-

ción (`lemmatized_text`).

- Texto lematizado sólo de adjetivos, nombres, verbos y adverbios acabados en "mente" (`lemmatized_text_reduced`). El objetivo es mantener aquí sólo aquellos términos que tienen valor semántico y aportan significado al texto.
- Lista de entidades nombradas: organizaciones (`organizations`), personas (`persons`), lugares (`locations`), fechas (`dates`), números (`numbers`) y otras (`others`).

Para obtener esta información se utilizó la herramienta FreeLing (Padró y Stanilovsky, 2012). Después de realizar unas pruebas de rendimiento, se estimó que el procesado del corpus llevaría aproximadamente 108 horas de ejecución. Para reducir este tiempo se decidió paralelizar el análisis de las noticias mediante la librería *OpenMP*.[10] Con esta decisión se consiguió una reducción del 72 % en el tiempo de procesamiento del corpus, completando el proceso en 30 horas.

## *4 Análisis del corpus*

Sobre el corpus creado se han realizado una serie de análisis estadísticos para determinar la riqueza léxica de los textos, estudiar la evolución temporal del lenguaje y también su uso en distintas zonas geográficas. Los siguientes apartados de esta sección profundizan en cada uno de estos aspectos.

### 4.1 Estadísticas generales

Como se ha comentado anteriormente, el corpus final cuenta con un total de más de 1.8 millones de noticias (tras la eliminación de duplicados) y un tamaño cercano a los 20 GB después de incorporarle la información resultante del procesamiento con FreeLing. La Figura 2 muestra un mapa coroplético con el número de noticias publicadas en cada provincia para el total del corpus. Se ha creado una infografía interactiva completa, accesible en Internet, donde se puede ver el número exacto de noticias por provincia y su evolución a lo largo de los años.[11]

Las provincias con mayor número de noticias son Sevilla (176.903), seguida de Valencia (100.455) y Cantabria (92.268), mientras

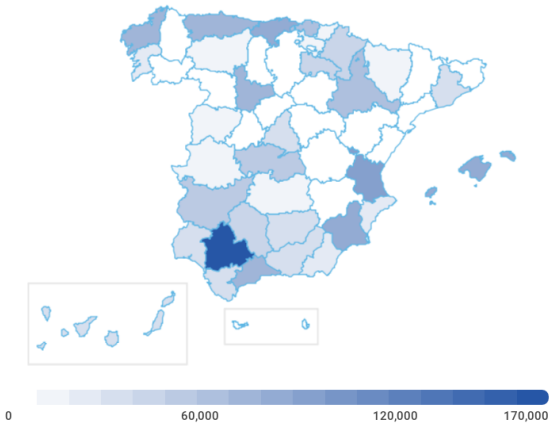


Figura 2: Número de noticias por provincia para el total de años analizado

que la que menos tiene es Ceuta (2.789). El lugar con mayor número de noticias en un único día es Murcia, el 12 de mayo de 2011, con 164 artículos (este punto se describe con más detalle en la Sección 5.1). La media de noticias por localización es de 35.134,33, mientras que su desviación estándar es de ±33.989,52, reflejando éste último dato una diferencia notable en el número de publicaciones entre distintas ediciones locales.

En cuanto a la distribución de noticias por años, los periodos de mayor actividad fueron 2011 (200.599), 2014 (199.721) y 2013 (199.334), mientras que los que menos publicaciones tuvieron fueron 2005 (26.845), 2009 (28.203) y 2006 (42.809). La media por años se sitúa en 130.498,93 noticias con una desviación estándar de ±73.869,20, situación análoga al análisis anterior por provincia.

Por lo que respecta al número de palabras del corpus, éste se compone de un total de 711.840.945 términos. Para medir la riqueza léxica del corpus analizamos el *Type-Token Ratio* (TTR) (Holmes, 1985), que representa el cociente entre el número de palabras diferentes del corpus y el número total de palabras. Dado que el TTR es dependiente de la longitud del texto, para minimizar el impacto de ésta se calculó su valor de manera individual para cada noticia y luego se promedió sobre el total. En la Figura 3 se puede ver la evolución del TTR a lo largo de los años, teniendo en cuenta tanto el vocabulario total (etiqueta *TTR* en el gráfico) como el que incluye únicamente palabras con valor semántico (gráfico *TTR reducido*). En la gráfica se observa una pérdida de riqueza léxi-

[10] `https://www.openmp.org`

[11] `http://cort.as/-C56M`, accedida en noviembre de 2018.

ca con el paso de los años, desde 2005 hasta 2011, estabilizándose posteriormente con un repunte de la riqueza en los últimos tres años estudiados.

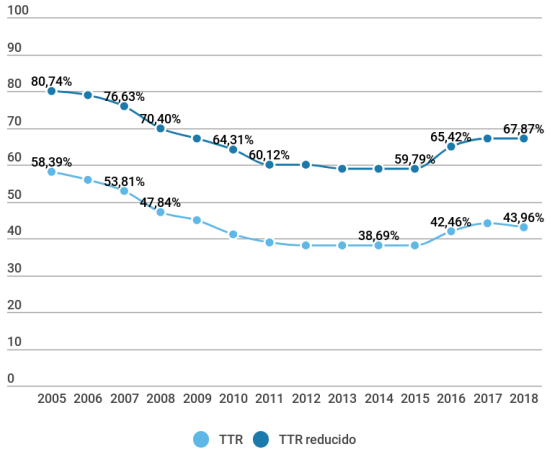


Figura 3: Evolución del TTR a lo largo de los años, tanto para el léxico completo (*TTR*) como para el subconjunto de los términos con valor semántico (*TTR reducido*)

Este cálculo se llevó a cabo también a nivel de localización geográfica.[12] El estudio revela que las provincias con mayor TTR son Alicante (69,08 %), Pontevedra (68,46 %) y Madrid (67,91 %). En la otra cara de la moneda, Toledo (60,72 %), Castellón (60,81 %) y Sevilla (61,45 %) presentan los índices más bajos de riqueza léxica.

## 4.2 Análisis diacrónico

El objetivo del estudio llevado a cabo en este apartado es averiguar, para cada año, cuáles han sido los temas más populares entre las publicaciones del periódico a partir del léxico empleado. Para la obtención de los términos más populares se tuvieron en cuenta solo las palabras con contenido semántico, tal y como fueron descritas en la Sección 3.3. Se ha realizado una infografía donde se pueden apreciar en forma de nube de palabras los cincuenta términos (lemas) más importantes y su frecuencia de aparición para cada anualidad.[13]

La Figura 4 muestra como ejemplo la nube de palabras correspondiente al año 2017. Hay términos esperables en esta lista de palabras más populares como "gobierno", "ayuntamiento", "PP" (muy por encima de otros partidos políticos) o "presidente". Llama la atención la frecuencia de uso de la palabra "persona", hecho que puede deberse a la neutralidad que presenta a nivel de género, siendo por ello muy utilizada por los periodistas para producir textos que contengan un lenguaje inclusivo y no sexista.

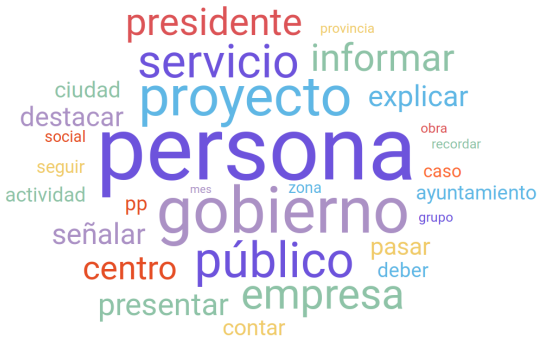


Figura 4: Nube de palabras con los cincuenta términos (lemas) más frecuentes en el año 2017

## 4.3 Análisis diatópico

De manera análoga al apartado anterior, se procedió a realizar un estudio de los términos más usados en las distintas localizaciones estudiadas, obteniendo los cincuenta términos más frecuentas en cada una de ellas. También se evaluaron las entidades más populares que habían sido extraídas por FreeLing, tal y como se describe en la Sección 3.3. Toda esta información está incluida en la infografía mencionada en el apartado anterior.

Como era de esperar, entre las palabras más repetidas para cada lugar está el propio nombre de la provincia y su gentilicio, junto con la comunidad autónoma a la que pertenece. Es destacable también la aparición de partidos políticos locales (como "BNG" en A Coruña, "PNV" en País Vasco y "Foro [Asturias]" en Asturias) y la preponderancia que tienen algunos grupos en determinadas regiones (en Cádiz tanto "PP" como "PSOE" son términos muy frecuentes) frente a otras en las que desaparecen (por ejemplo, en Barcelona no aparecen ninguno de estos dos entre los cincuenta términos más frecuentes).

En aquellas regiones en las que existen lenguas propias, se pueden encontrar entre los más frecuentes términos muy representativos de dichos lugares, como "generalitat" en Valencia, "xunta" en A Coruña, "euskadi" en

[12] Todos los detalles están disponibles en la siguiente infografía: `http://cort.as/-C56R`, accedida en noviembre de 2018.

[13] `http://cort.as/-C56W`, accedida en noviembre de 2018.

Vizcaya o “mossos” en Barcelona.

## *5 Casos de estudio*

En esta sección se muestran dos casos de ejemplo del tipo de información que se puede obtener del corpus desarrollado mediante el análisis de texto. Una de las aplicaciones interesantes de este corpus es su uso para el periodismo de datos (Gray, Chambers, y Bounegru, 2012), una especialidad del periodismo que refleja el creciente valor de los datos en la producción y distribución de información, cuyo objetivo es recabar gran cantidad de datos y hacer la información comprensible a la audiencia ayudándose de herramientas como las infografías, representaciones gráficas o aplicaciones interactivas.

En el primero de los casos de estudio se analiza el terremoto que tuvo lugar en Lorca el 12 de mayo de 2011, mientras que en el segundo estudio se aborda el tema de la independencia de Cataluña a lo largo de los dos últimos años.

### 5.1 Terremoto de Lorca

Tal y como se comentó en la Sección 4.1, la provincia que más noticias tuvo en un único día fue Murcia, el 12 de mayo de 2011, con 164 artículos. El suceso que provocó este aluvión de información fue el seísmo ocurrido el día anterior en la localidad de Lorca con una magnitud de 5,1 en la escala Richter, causando numerosos daños materiales, más de 300 heridos y 9 muertos. Para recuperar las noticias relacionadas con este evento, se localizaron todas aquellas que contenían el término “Lorca” junto con “terremoto” o “seísmo”.

Para analizar el suceso partiendo del corpus desarrollado, se ha creado una nueva infografía[14] para identificar los términos más relevantes asociados con esta noticia, los lugares en los que más se habló del tema (desde el punto de vista del número de noticias publicadas) y un estudio de la repercusión de la noticia a lo largo del tiempo, identificando el momento en el que los medios redujeron su interés en este suceso. Este último aspecto se puede apreciar en la Figura 5.

Cabe destacar las posibilidades que ofrece el corpus como herramienta para determinar el interés en el tiempo que produce una noticia. En el caso de ésta, se aprecia como una semana después del seísmo su interés decreció

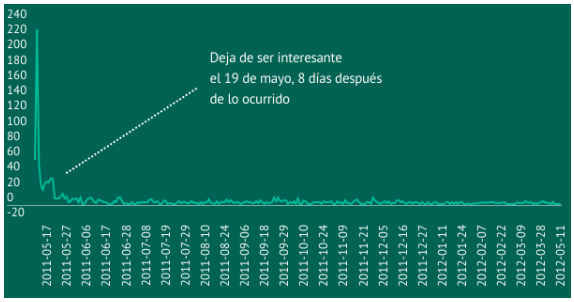


Figura 5: Seguimiento periodístico del terremoto de Lorca durante un año, mostrando el número de noticias publicadas en los distintos días

notablemente en cuanto al número de publicaciones relacionadas, mostrando la limitada vida que puede tener en los medios de comunicación un evento de esta magnitud.

### 5.2 Independencia de Cataluña

En este apartado se presenta un segundo estudio siguiendo las pautas del anterior, con el foco puesto en un tema de relevancia política nacional como es la independencia de Cataluña. Los términos que se emplearon para localizar noticias relacionadas con esta temática fueron: “independencia”, “independència”, “independentismo”, “independentisme”, “independentista”, “procés” y “estatut”. El periodo de estudio se fijó desde el 1 de enero de 2017 hasta el 5 de julio de 2018 (último día en el que se incorporaron noticias al corpus).

Al igual que para el caso del terremoto de Lorca, se ha diseñado una infografía que hace un análisis temporal y espacial de la cobertura del suceso en el corpus.[15] En este análisis se pueden observar datos destacables, como que la cobertura de este tema en el periódico *20 minutos* se ha llevado a cabo de manera prácticamente ininterrumpida en el año y medio a estudio: solo en 6 días de los 445 analizados no se produjo ninguna noticia relacionada con este asunto. También llama la atención detalles como que en Ceuta y Melilla no se hace ninguna mención al proceso en todo el periodo analizado, que Barcelona (1.289 noticias), Valencia (336 noticias) y A Coruña (335 noticias) son las provincias con mayor cobertura del tema, y que en Madrid apenas se encuentran 28 noticias sobre este asunto, un número llamativamente bajo te-

[14] `http://cort.as/-C56d`, accedida en noviembre de 2018.

[15] `http://cort.as/-C56g`, accedida en noviembre de 2018.

niendo en cuenta el volumen de publicaciones de su sección local. En la Figura 6 se puede ver una captura de la infografía centrada en el análisis temporal.

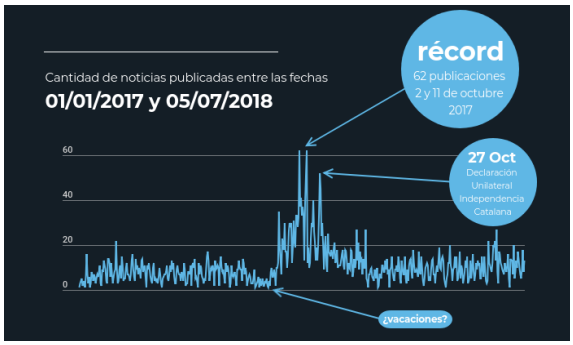


Figura 6: Análisis temporal de las noticias relacionadas con la independencia de Cataluña

## 6 Conclusiones y trabajo futuro

Este artículo presenta el desarrollo de un corpus de noticias periodísticas de gran tamaño que contempla tanto la dimensión temporal como la geográfica de las noticias. Hasta donde tenemos conocimiento, no existe en la actualidad un corpus de noticias con estas características libremente disponible. Este corpus puede servir de fuente para realizar distintos estudios del uso y evolución del lenguaje, además de ser un recurso de utilidad para el periodismo de datos, tal y como se mostró en la Sección 5 mediante dos casos de estudio. Si bien el corpus, por cuestiones de derechos de uso, no está disponible para su distribución, una de las aportaciones de este trabajo es proporcionar las herramientas software necesarias para que cualquier persona que quiera trabajar con él pueda descargarlo y replicarlo en su ordenador.

En este trabajo se ha descrito todo el proceso llevado a cabo para la extracción y obtención de las noticias, dificultades afrontadas para la limpieza y eliminación de duplicados, así como el posterior análisis lingüístico y enriquecimiento del mismo. Se han proporcionado también datos estadísticos del corpus resultante desde las dimensiones temporal y geográfica, desarrollando diversas infografías interactivas que permiten un análisis más profundo y detallado del mismo.

Como trabajo futuro se plantea el desarrollo de nuevos estudios de carácter lingüístico y sociológico que puedan resultar de interés a partir de esta fuente de información. Se proyecta también extender este trabajo para su aplicación a otros corpus de textos en la Web que puedan ser recopilados de manera similar.

## Bibliografía


Broder, A. Z. 1997. On the resemblance and containment of documents. En *Proceedings of the Compression and Complexity of Sequences 1997*, SEQUENCES '97, páginas 21–29, Washington, DC, USA. IEEE Computer Society.

Graff, D. y G. Gallegos. 1995. Spanish news text. Download at Linguistic Data Consortium: `https://catalog.ldc.upenn.edu/LDC95T9`.

Graff, D. y G. Gallegos. 1999. Spanish newswire text, volume 2. Download at Linguistic Data Consortium: `https://catalog.ldc.upenn.edu/LDC99T41`.

Gray, J., L. Chambers, y L. Bounegru. 2012. *The data journalism handbook: How journalists can use data to improve the news*. O'Reilly Media.

Holmes, D. I. 1985. The analysis of literary style. *Journal of the Royal Statistical Society. Series A (General)*, 148(4):328–341.

Indyk, P. y R. Motwani. 1998. Approximate nearest neighbors: Towards removing the curse of dimensionality. En *Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing*, STOC '98, páginas 604–613, New York, NY, USA. ACM.

Leetaru, K. 2011. Culturomics 2.0: Forecasting large-scale human behavior using global news media tone in time and space. *First Monday*, 16(9).

Leskovec, J., A. Rajaraman, y J. D. Ullman. 2014. *Mining of Massive Datasets*. Cambridge University Press, New York, NY, USA, 2nd edición.

Padró, L. y E. Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. En *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey. ELRA.