El impacto del «*Big Data*» en el el lenguaje

Eugenio Martín Fuentes/
Daniel Martín Mayorga

En un artículo de 1976 titulado «Ingeniería e intimidad», Juan Benet nos ilustraba sobre las razones del progresivo desmoronamiento de la apreciación social hacia los ingenieros, antaño considerados poco menos que taumaturgos, y ya en esa fecha –no digamos ahora– desprovistos de toda mística profesional si esta no viene avalada por un puesto de trabajo especialmente bien remunerado. La razón, explica, es que el progreso ha ido despojando a la ingeniería de ropajes superfluos para dejarla en su esencia: una actividad puramente mediadora, que en sí misma no responde a fin alguno. Dicho con las palabras de Benet, lo importante para la sociedad no es la presa, sino el agua que embalsa.

Treinta y muchos años después, la hipótesis benetiana se ha terminado de confirmar, y de manera ya incontestable. El incremento exponencial de la capacidad de procesamiento y almacenamiento informático, unido a factores como el desarrollo de nuevos

servicios e infraestructuras de comunicaciones, ha dado una vuelta de tuerca más a esta sociedad de la información. En efecto, la obtención y acumulación de información de todo tipo ha explotado en los últimos tiempos, hasta el punto de que hoy en día menos del cinco por ciento de los datos que se guardan en los repositorios de todo el mundo tiene una antigüedad mayor a tres años.

Desde que empieza el día, prácticamente todas nuestras acciones tienen un correlato informático que las registra; desde el manejo del ordenador o el teléfono, al uso de cualquier servicio. Y esto es continuo en el caso de las máquinas, todas vinculadas telemáticamente a unidades de control. Es la explosión de los macrodatos (*Big Data*, en su terminología más reconocible), los cuales representan la síntesis misma de la intermediación, tan cercana a la realidad como –casi– solo lo puede estar la realidad misma.

Quizá sería necesario un breve excurso en relación con el concepto *Big Data* que, si bien resulta común en el ambiente tecnológico o de negocios, lo es menos en otros contextos... de momento. Pero, como suele ocurrir en estos casos, la amplitud y versatilidad de la cuestión dificultan una explicación al uso. Por eso vamos a circunscribir este análisis a un campo de aplicación de los macrodatos particularmente interesante: la lengua. Nos ayudará a comprender mejor el concepto y, a la vez, permitirá reflexionar sobre su aplicación al quizá mayor activo que tiene nuestro país, el idioma español.

El lenguaje humano se basa en un sistema de códigos a menudo arbitrarios, en el que el significado y significante tienen normalmente muy poca relación icónica y cuya utilización en muchos contextos resulta con frecuencia ambigua. El lenguaje no se fabrica de acuerdo con unas especificaciones, ni se conduce, ni se condiciona. Surge y se desarrolla espontáneamente como resultado de un proceso natural en el seno de las sociedades. Con el tiempo, instituciones académicas se constituyen en testigos de

su evolución y documentan y contextualizan sus cambios, pero no modulan la lengua ni intervienen en su desarrollo.

El interés del lenguaje como ciencia toma carta definitiva de naturaleza en el siglo XIX, con la acuñación del término «Lingüística». Es entonces, vista bajo el prisma científico, cuando el carácter arbitrario de la lengua adquiere especial relevancia. A diferencia de otras ciencias en las que es posible representar la realidad mediante modelos matemáticos y, en consecuencia, predecir resultados, el lenguaje no se sujeta a reglas. Las palabras y las frases pueden transmitir uno u otro significado según el contexto en que se empleen o incluso el tono con el que se verbalicen; además de otras muchas complejidades que no siempre resultan tan explícitas.

Y puesto que no es posible estructurar un modelo predictivo del lenguaje, no cabe otra opción que recurrir a la observación. Observar, experimentar y concluir son la base de la investigación lingüística.

Sobre este principio, en la pasada década de los sesenta, se desarrolló lo que se conoce hoy en día como «Lingüística de Corpus». Esta metodología se basa en la utilización de fuentes de información (corpus) que aportan evidencias del uso del lenguaje en diferentes contextos; lo que puede ser utilizado, mediante la aplicación de técnicas de procesamiento lingüístico, para sintetizar reglas globales. La Lingüística Computacional (LC) conjuga métodos de análisis del lenguaje con tecnologías de tratamiento por ordenador.

De una forma sencilla, se puede definir un corpus como un conjunto de textos o fragmentos de textos representativos de una lengua, codificado bajo un esquema de metadatos que permite ubicar a cada uno de ellos en un escenario concreto: autor, procedencia geográfica, temática, fecha de creación, etc. Las palabras que forman estos textos son asimismo codificadas y etiquetadas con información que resulta determinante para el investigador (cate-

goría gramatical, acepción dentro del texto cuando se trate de una forma ambigua, etc.). Para un ingeniero, un corpus no sería más que una base de datos. Un cartógrafo, por su parte, lo vería como un mapa a escala del lenguaje.

Construir un corpus es una tarea que conlleva gran trabajo, cuya complejidad no está en la captura de los textos en formato electrónico –algo resuelto por la tecnología– sino en la adecuada selección de las muestras, en su codificación y procesamiento. No hay que olvidar que, en su sentido más general, el corpus debe ser una muestra representativa de una determinada lengua en todas sus facetas.

Otro tanto sucede cuando se trata de etiquetar el corpus. Realizar esta labor de forma manual es impensable si lo que se pretende es procesar millones de palabras. Hay que recurrir en estos casos a la aplicación de principios metodológicos y algoritmos sustentados por una base computacional y por programas de ordenador desarrollados específicamente para ese cometido. Esta técnica está recogida también dentro de la LC.

Los métodos de LC se aplican no solo en la creación de los corpus sino también en su explotación. Con estas técnicas es fácil determinar un amplio rango de estadísticas y características segmentadas del lenguaje: frecuencias de aparición de palabras, coincidencias de unas palabras con otras (*coapariciones*), utilización de determinadas formas en uno u otro país o en una u otra época, etc. La explotación de los corpus permite visualizar la evolución de una determinada lengua, estudiar sus estructuras y analizar tendencias desde una perspectiva tanto cualitativa como cuantitativa. Para el investigador, el corpus es el telescopio con el que, a través de una lente pequeña, alcanza a ver todo un universo.

Los estudios basados en corpus comenzaron a tener protagonismo hace más de cincuenta años, coincidiendo con el cambio paradigmático que supuso la incorporación del ordenador al

ámbito lingüístico. Además de esta circunstancia, también en gran medida deben su despegue al lanzamiento de emblemáticos proyectos de investigación en países anglosajones y escandinavos, a partir de la construcción de grandes corpus lingüísticos electrónicos para el inglés.

En el siglo XX se imponen los métodos de análisis basados en la observación del comportamiento de formas ortográficas (palabras) y su frecuencia de aparición. Sin embargo, estos métodos se ven limitados por la imposibilidad de acceder a repertorios de datos lingüísticos significativos en tamaño y calidad. En aquellos primeros años, las limitaciones en la capacidad para capturar datos y la no menor limitación en su almacenamiento, procesamiento y análisis, impusieron grandes restricciones al desarrollo de la Lingüística de Corpus. A partir de la pasada década de los noventa, sin embargo, la popularización del ordenador personal y, sobre todo, la creciente actividad comunicativa humana soportada por dispositivos digitales ha generado una gran cantidad de textos en formato electrónico. La existencia de estos materiales ha propiciado además el desarrollo de una infraestructura tecnológica específicamente concebida para el procesamiento y explotación de estos recursos.

Llegados nuestros días, somos testigos de la explosión de nuevas formas de comunicación a través de Internet y de las redes sociales, que se convierten a su vez en nuevas fuentes de recursos lingüísticos.

La abundancia de información no es un atributo privativo del campo de la lengua. Cualquier sociedad de la información moderna maneja hoy en día vastos repositorios de datos, tanto públicos como privados. Moviéndonos fuera de la esfera lingüística y situándonos en otros ámbitos, varias disciplinas científicas manejan volúmenes de datos del orden de los petabytes (1 petabyte = 10^{15} = 1.000.000.000.000,000 bytes). La comunidad de física de altas

energías utiliza el gran colisionador de Hadrones o LHC, cerca de Ginebra, para sus investigaciones, que suponen una producción de quince petabytes de datos al año. El telescopio LSST (Large Synoptic Survey Telescope) instalado en Chile, en funcionamiento a partir de 2015, será capaz de observar el cielo al completo cada pocos días; su cámara de 3,2 gigapíxeles generará imágenes con un tamaño de medio petabyte al mes. El Instituto Europeo de Bioinformática (EBI), mantiene un repositorio de secuenciaciones de ADN que alcanzó los nueve petabytes en 2009 y que seguramente revolucionará los campos de la genética y la medicina. Esta capacidad avanzada para la computación ayuda a los investigadores a explorar y extraer información sobre colecciones de datos masivos. Es el denominado «cuarto paradigma» en ciencia, que complementa la teoría, los experimentos y las simulaciones.

El Business Intelligence Lowdown (BIL) establece una clasificación con las bases de datos más grandes del mundo. El WDCC (Centro Mundial de Datos para el Clima) aparece en primer lugar. Almacena unos 6 petabytes y 220 terabytes (1 terabyte = 10^{12} = 1.000.000.000.000 bytes) de datos sobre el clima, predicciones y simulaciones. En segundo lugar se encuentra el National Energy Research Scientific Computing Center (NERSC), organismo de investigación sobre distintos tipos de energía. Su base de datos tiene 2,8 petabytes. En tercer lugar se sitúa AT&T, compañía de telecomunicaciones que almacena 323 terabytes de información.

Google aparece en esta clasificación en cuarto lugar, aunque en realidad nadie conoce con precisión el verdadero tamaño de su base de datos. No hace mucho tiempo, esta compañía sacó a la luz su gran base de datos de la lengua («Culturomics») que incluye un 4 % de todos los textos escritos en el mundo y que se focaliza en el fenómeno lingüístico y cultural. Este gigantesco corpus utiliza la colección de Google Books del inglés y de otros idiomas, cubriendo el espacio temporal de los dos últimos siglos. Sobre esta

gran base de datos, Google es capaz de hacer amplios y llamativos análisis de patrones.

La abundancia de información que ofrece esta época en que nos ha tocado vivir facilita mucho las cosas al lingüista. Ahora se puede agrandar la escala del «mapa» ampliando el tamaño de la muestra y siendo con ello más precisos en los análisis. Se ha pasado casi sin transición de la escasez a la abundancia y, sin embargo, no todas las opiniones son coincidentes: comienzan ya a surgir las primeras críticas. Voces autorizadas de entidades relevantes (Microsoft Research) han manifestado públicamente su escepticismo y expresado su preocupación por el enfoque que se da a esta metodología, que pone el énfasis en la recopilación masiva de datos y relega a un segundo plano la elección de muestras representativas. Desde esta perspectiva, no se trataría solo de una cuestión de volumen; la calidad del dato también debe ser considerada.

Aplicado a lo que nos ocupa, un corpus generalista debe contener una muestra suficientemente representativa del lenguaje en todas sus facetas. Un corpus formado solo por textos de medicina podría llegar a ser un reflejo fiel del lenguaje de los médicos, pero no sería en absoluto representativo de otros colectivos. La selección de textos y el acceso a información de calidad son tareas fundamentales en el desarrollo del corpus.

La tecnología moderna no solo ha facilitado la construcción de grandes bases de datos; también ha propiciado el desarrollo de mecanismos que permiten, mediante el establecimiento de vinculaciones con recursos del tipo datos enlazados (*Open Linked Data*), poner en comunicación entre sí estos repositorios o interconectarlos con recursos abiertos de otra naturaleza. Es posible así tejer una inmensa red de información que el usuario contempla desde fuera como un repositorio de datos único.

La posibilidad de acceder a estas macro fuentes de información y con ello de analizar grandes conjuntos de datos textuales podría

proporcionar al lingüista o al lexicógrafo información de gran utilidad para realizar todo tipo de medidas sobre el léxico: con qué frecuencia se utilizan las palabras, cómo se distribuyen sus apariciones en textos, con qué otras palabras aparecen (coapariciones), etc. También sería de gran utilidad para la selección de léxicos nucleares o diatópicos, agrupación temática de palabras, etc. Explotando las coapariciones de segundo nivel se podría abordar la inducción de sentidos. Usando la información de dependencias obtenida en la fase de análisis sintáctico se podrían proporcionar lo que se conoce como *word sketches*, un primer esquema del comportamiento distribucional de una palabra: sus complementos más típicos, sus modificadores, la combinatoria significativa con otras palabras, etc. Finalmente, pero no por ello menos importante, el estudio de las redes de palabras abriría nuevas vías a las agrupaciones semánticas y a la detección de sentidos.

Estos repositorios de datos no solo son de interés para el investigador de la lengua. También lo han sido tradicionalmente para sectores industriales como el editorial, que ha venido utilizándolos, de una u otra forma y en mayor o menor medida, para la redacción de diccionarios, gramáticas, ortografías y otras obras de carácter normativo. Sin embargo, su interés industrial se ha visto acrecentado en los últimos años con la aparición de unas nuevas tecnologías basadas en el procesamiento del habla (tecnologías lingüísticas) y en la interpretación del lenguaje por las máquinas, cuyo desarrollo ha surgido como una consecuencia directa de la simbiosis entre la informática y la lingüística.

Estas tecnologías han propiciado a su vez la aparición de industrias emergentes que están generalizando el consumo de recursos lingüísticos a todos los niveles. Sus negocios pivotan sobre actividades como reconocimiento automático de contenido, sistemas de pregunta-respuesta, traducción de voz a voz, análisis de lo que se dice en redes sociales, análisis de textos y reconocimiento de habla.

Las empresas que conforman este nuevo sector industrial utilizan el lenguaje como recurso básico en sus productos y servicios y, para obtener la materia prima que necesitan, no tienen más remedio que hacer minería en esos grandes repositorios de datos.

El procesamiento de estas grandes bases de datos queda fuera del alcance de los sistemas informáticos tradicionales. Pero este es un problema al que ya se han enfrentado otras comunidades científicas. Toda organización construida alrededor de la recopilación, el análisis, la monitorización, el filtrado, la búsqueda o la organización de contenido debe enfrentarse al problema de la escala en el tamaño de sus repositorios de datos y a su procesamiento.

Cuando los grandes experimentos del colisionador de Hadrones o el Large Synoptic Survey Telescope, antes mencionados, comenzaron a arrojar datos, fue necesario pensar en una organización muy específica de los sistemas informáticos para su procesamiento. Surgió así un nuevo concepto en las técnicas de explotación de grandes volúmenes de información que se ha dado en llamar «macrodatos» o, en terminología anglosajona, «*Big Data*».

El concepto de «*Big Data*» se aplica a toda aquella información que no puede ser capturada, gestionada y procesada en un tiempo razonable mediante las capacidades del *software* tradicional y se utiliza al hablar de repositorios de datos de petabytes y exabytes (*1 e*xabyte = 10^{18} = 1.000.000.000.000.000.000 bytes).

Según Gartner, «*Big Data* es un repositorio de gran volumen, de alta velocidad y de gran variedad de datos, que requieren nuevas formas de procesamiento para permitir la toma de decisiones mejorada y la optimización de procesos. *Big Data* utiliza estadísticas inductivas con datos de baja densidad de información, cuyo volumen enorme permiten inferir leyes, confiriéndole así algunas capacidades predictivas».

En otras palabras, «*Big Data*» es un concepto genérico que quiere representar una arquitectura informática a la vez que un

modelo de procesamiento, utilizados ambos para explotar datos masivos e inducir resultados a partir de ellos.

En la implementación de esta arquitectura se aplica el concepto de computación distribuida, que resuelve el problema del procesamiento masivo utilizando un gran número de ordenadores organizados en racimos, incrustados en una infraestructura de telecomunicaciones distribuida.

En los sistemas de almacenamiento tradicionales, una base de datos única es consultada por aplicaciones o procesos. En la arquitectura *«Big Data»* la gestión de la base de datos se realiza de forma distribuida entre varios servidores; cada uno de ellos gestiona una parte de la misma y sirve la parte de contenido que le corresponde a la aplicación que interroga. En un segundo paso se compone el conjunto de datos final, mediante la integración de la información obtenida separadamente de cada servidor.

La infraestructura *software* estará compuesta por procedimientos y programas que se encargan de descomponer el problema en subproblemas y repartirlos entre los ordenadores que componen la granja de servidores para, finalmente, componer una solución a partir de las soluciones parciales calculadas por cada ordenador. Para la implementación de la infraestructura software existen distintos modelos de programación, por ejemplo *MapReduce*, empleado en la resolución práctica de algoritmos susceptibles de ser paralelizados.

La fragmentación de los datos y la distribución de las consultas entre servidores que trabajan en paralelo permiten, entre otros, la posibilidad de especialización de máquinas en determinados procesos, lo que redunda en una mayor velocidad de acceso y en mejores prestaciones en el manejo de la información.

En resumen, la investigación, el estudio de la lengua y el desarrollo de la industria que emerge como consecuencia de su maridaje con las nuevas tecnologías requieren cada vez más recursos

lingüísticos. Estos recursos se asocian a muestras de la lengua contenidas en repositorios de textos procesados (corpus) que pueden alcanzar volúmenes por encima de los terabytes. La facilidad de adquirir textos y fragmentos (muestras reales) de las lenguas, ha convertido el problema inicial de la escasez de material en el de la selección dentro de la abundancia y en el del tratamiento eficiente de grandes volúmenes de información. Las técnicas de almacenamiento y procesamiento masivo de datos conocidos como *«Big Data»* encontrarán aquí sin duda un nuevo campo de aplicación.

Concluimos con una pregunta que inevitablemente quedará en el aire: estas cuestiones, aquí planteadas de forma genérica, ¿cómo afectan en concreto a la lengua española? La singular unidad y extensión geográfica de nuestro idioma, así como la vigencia del papel normativo de la Real Academia Española conjuntamente con el resto de las Academias de los países americanos, avalan contemplar con optimismo el futuro. Pero la irrupción de los macrodatos supone un salto tecnológico que filólogos y lingüistas computacionales tendrán que dar del brazo de ingenieros e informáticos. De otro modo, el estándar digital de la lengua española, algo más parecido a una base de datos que a un diccionario, junto con la inmensa industria asociada, puede quedar fuera del control de las instancias que hasta ahora se han ocupado de guardar y enriquecer este legado. Aunque nos pueda parecer imposible.

E. M. F./D. M. M.