Culturómica I: Cultura y lenguaje

Pedro García Barreno
Real Academia Española
Real Academia de Ciencias Exactas, Físicas y Naturales

Resumen: La Inteligencia Artificial (IA) es una de las fuerzas más transformadoras de nuestro tiempo. Si bien puede haber debate sobre si la IA transformará nuestro mundo para bien o para mal, algo en lo que todos estamos de acuerdo es que la IA no sería nada sin los datos masivos o macrodatos. *Big data* e IA se consideran dos gigantes. El aprendizaje automático se considera una versión avanzada de la IA a través de la cual las computadoras inteligentes pueden enviar o recibir datos y aprender nuevos conceptos mediante el análisis de los datos sin asistencia humana. El Gran Colisionador de Hadrones, por ejemplo, generará unos 15 *petabytes* de datos al año. Eso no es nada comparado con lo que sucede cuando mapeamos un cerebro completo, lo que implicará alrededor de un millón de *petabytes* de datos. La astronomía, la química, los estudios climáticos, la genética, el derecho, la ciencia de los materiales, la neurobiología, la teoría de redes o la teoría de partículas son solo algunas de las áreas que ya están siendo transformadas por grandes bases de datos. Ahora esta revolución está llegando a las humanidades. El programa de libros masivos de Google, que ha digitalizado millones de libros, ha desarrollado una aplicación que brinda a los investigadores acceso a una base de datos de miles de millones de palabras en varios conjuntos de idiomas y dos siglos: «datos grandes y extensos». El programa de Google *N-gram Viewer* hace más que brindar una mirada única a la historia de las palabras. Promete cambiar la forma en que los historiadores hacen su trabajo y cambiar nuestra imagen de la historia misma. Un nuevo tipo de alcance, *big data*, va a cambiar las humanidades, a transformar las ciencias sociales y a renegociar la relación entre el comercio mundial y la «torre de marfil». Paralelamente, la arquitectura cognitiva juega un papel vital al proporcionar planos para construir sistemas inteligentes que admitan una amplia gama de capacidades similares a las de los humanos. La arquitectura de red neuronal para aprender vectores de palabras puede manejar más de 100 mil millones de palabras en un día. Una traducción automática neuronal (NMT) traduce

entre varios idiomas, y también puede aprender a realizar puentes implícitos entre pares de idiomas nunca vistos (traducción a ciegas –*zero-shot translation*–) explícitamente durante el entrenamiento, lo que demuestra que el aprendizaje de transferencia y la traducción a ciegas son posibles para la traducción neuronal. Un marco de entrenamiento novedoso para agentes de diálogo con base visual –aprendizaje de refuerzo profundo (RL) para aprender de extremo a extremo en un mundo sintético completamente sin conexión a tierra, donde los agentes se comunican a través de símbolos sin significados preestablecidos– mostró que dos bots inventan su propio protocolo de comunicación sin supervisión humana; ¿*tabula rasa*? Los agentes de RL no solo superan significativamente a los agentes de aprendizaje supervisado, sino que también aprenden a aprovechar las fortalezas de los demás, sin dejar de ser interpretables para los observadores humanos externos. Los bots parlantes –bots-*talk*– recuerdan a los gemelos comunicativos –*twins-talk*–, a la novela postestructuralista o a los lenguajes culturalmente restringidos. Los lenguajes de IA pueden evolucionar a partir de un lenguaje humano natural o pueden crearse *ab initio*.

Palabras clave: Aprendizaje por refuerzo, *Big data* (macrodatos), Lengua Beowulf, Datos fiables, Datos instantáneos, Generación de lenguaje natural, Humanidades digitales, Índice de Busa (Índice Tomístico), Índice del *Ulises* de Joyce, Ley de Zipf, *N-gram*, Programa Google de libros masivos, Reconocimiento óptico de caracteres, Traducción automática neuronal, Traducción a ciegas.

Abstract: Artificial Intelligence (AI) is one of the most transformative forces of our times. While there may be debate whether AI will transform our world in good or evil ways, something we all agree on is that AI would be nothing without big data. Big data and AI are considered two giants. Machine learning is considered as an advanced version of AI through which smart computers can send or receive data and learn new concepts by analyzing the data without human assistance. The *Large Hadron Collider*, for example, will generate about 15 petabytes of data per year. That's nothing compared to what happens when we map a whole brain, which will involve about a million petabytes of data. Astronomy, chemistry, climate studies, genetics, law, materials science, neurobiology, network theory, or particle theory are just a few areas already being transformed by large databases. Now this revolution is coming to the humanities. Google's massive book program, which has digitized millions of books, has spun off an application that gives researches access to a data-base of billions of words across several language set and two centuries: «big-and-long data». Google's program –*N-gram Viewer*– does more than provide a unique look at the history of words. It promises to change how historians do their work and to change our picture of history itself. A new kind of scope –big data– is going to change the humanities, transform the social science, and renegociates the relationship between the

world commerce and the «ivory tower». In parallel, cognitive architecture play a vital role in providing blueprints for building intelligent systems supporting a broad range of capabilities similar to those of human. Neural network architecture for learning word vectors can train more than 100 billion words in a day. A *Neural Machine Translation* (NMT) translates between multiple languages, and NMT can also learn to perform implicit bridging between language pair never seen explicitly during training, showing that transfer learning and zero-shot translation is possible for neural translation. A novel training framework –deep Reinforcement Learning (RL) to end-to-end learn in a completely ungrounded synthetic world, where the agents communicate via symbols with no pre-specified meanings– for visually grounded dialog agents showed that two bots invent their own communication protocol without any human supervision; tabula rasa? RL agents not only significantly outperform supervised learning agents, but learn to play to each other's strengths, all the while remaining interpretable to outside human observers. Bot-talk remembers twins-talk, post-structuralist novel or languages culturally constrained. AI languages can be evolved starting from a natural human language or can be created *ab initio*.

Keywords: Beowulf language, Big-long-smart-fast-data, Busa's Idex (Index Thomisticus), Digital humanities, Google massive book program, N-gram, Natural Language Generation (*NLG*), Neural Machine Translation (*NMT*), Optical Character Recognition (*OCR*), Reinforcement Learning (*RL*), Word Index to James Joyce's Ulyses, Zero-shoot translation, Zipf's Law.

> «La ciencia que se ocupa de datos masivos –*big data*–, utiliza técnicas tan diferentes, que debe distinguirse de los tres paradigmas científicos existentes: teoría, experimentación y computación. Esta ciencia intensiva de datos debe contemplarse como un nuevo cuarto paradigma para la exploración científica».
>
> *James (Jim) Nicholas Gray*

INTRODUCCIÓN

En 1955 se publicaba *The Great Conversation. The Substance of a Liberal Education*, el primer volumen de la primera edición de *Great Books of the Western World*. En el curso de la historia, generación tras generación escribieron libros que han ido ganando un lugar en la lista que ha guiado la Gran Conversación; una cultura de diálogo que ha caracterizado a Occidente. No es que los libros representen la panacea para solucionar los problemas complejos a los que se enfrenta la humanidad, pero representan el punto de partida. La lectura debe ir acompañada de información de calidad sobre la que basar un juicio, y de la habilidad para hacer.

«Conocer no es suficiente, debemos aplicar. Querer no es suficiente; debemos hacer», remachó Johann W. von Goethe. Para ello es indispensable que la Gran Conversación se nutra de un bagaje significativo de información científica y técnica. Lejos del contenido de *El Canon Occidental* de Harold Bloom, David Weatherall, *Regius Professor* de Medicina en la Universidad de Oxford, escribió, también en 1955, con referencia a la medicina, pero que puede ampliarse a la totalidad de nuestras actividades:

> La importancia cada vez mayor de la ciencia en la provisión de atención médica y los difíciles problemas sociales y éticos que surgirán de nuestra nueva capacidad para determinar nuestro futuro hacen que sea esencial que todos seamos más alfabetizados científicamente. Nuestros políticos deben comprender los rudimentos de la evidencia científica, y la sociedad en su conjunto debe estar lo suficientemente bien informada para comprender la mejor manera de lograr una vida saludable y participar en el debate de los complejos problemas que seguirán planteando los avances en la investigación biológica y médica. Este movimiento hacia una mayor conciencia científica tendrá que comenzar en las escuelas, con un mayor apoyo a la educación científica.

En la primera parte de la *Rede Lecture* 1959, Charles Percival Snow se pregunta:

> Muchas veces he estado presente en reuniones de personas que, según los estándares de la cultura tradicional, se consideran altamente educadas y que han expresado con considerable entusiasmo su incredulidad ante el analfabetismo de los científicos. Una o dos veces me han provocado y he preguntado a los «inquisidores» cuántos de ellos podrían describir la Segunda Ley de la Termodinámica. La respuesta fue fría: también fue negativa. Sin embargo, estaba preguntando algo que es el equivalente científico de: ¿Has leído una obra de Shakespeare? Ahora creo que si hubiera hecho una pregunta aún más simple, como: ¿Qué quiere decir con masa o aceleración, que es el equivalente científico de decir, puede leer? –no más de uno de cada diez de los altamente educados habría sentido que estaba hablando el mismo idioma.

En *The Two Cultures: A Second Look*, escrito en 1963, C. P. Snow concluye:

> Los cambios en la educación no van a producir milagros. La parcelación de nuestra cultura nos está volviendo más obtusos de lo necesario. Pode-

> mos reparar en cierta medida la comunicación; pero, como he dicho antes, no vamos a producir hombres y mujeres que entiendan tanto de nuestro mundo como Piero della Francesca, o Pascal, o Goethe, comprendieron el suyo. Con buena fortuna, sin embargo, podemos educar a una gran proporción de nuestras mejores mentes para que no ignoren la experiencia imaginativa, tanto en las artes como en la ciencia, ni ignoren las dotes de la ciencia aplicada, el sufrimiento remediable de la mayoría de sus prójimos humanos, y las responsabilidades que, una vez vistos, no pueden negarse.

Peter Drucker, abogado, consultor, futurista..., escribió en *Innovation and Entrepreneurship*, en 1985:

> De hecho, estamos en las primeras etapas de una gran transformación tecnológica, una que es mucho más radical de lo que los 'futurólogos' más extasiados se dan cuenta hasta ahora, más grande que las mega-tendencias o *Future Shock*. Trescientos años de tecnología llegaron a su fin después de la Segunda Guerra Mundial. Durante esos tres siglos, el modelo de tecnología fue mecánico: los eventos que suceden dentro de una estrella como el sol. Este período comenzó cuando un físico francés casi desconocido, Denis Papin, imaginó la máquina de vapor alrededor de 1680. Terminó cuando replicamos en una explosión termonuclear los eventos dentro de una estrella. Para estos tres siglos el avance de la tecnología significó –como en los procesos mecánicos– más velocidad, mayores temperaturas, mayores presiones. [...] Desde el final de la Segunda Guerra Mundial, sin embargo, el modelo de la tecnología se ha convertido en el proceso biológico, los eventos dentro de un organismo. Y en un organismo, los procesos no se organizan en torno a la energía en el sentido físico del término. Se organizan en torno a la información.

Tal vez espoleados por este texto, un año después, surgía el embrión de los debates sobre la educación y formación del futuro. La AAAS lanzó, en 1989, la primera publicación del *Project 2061*. En la actualidad, *Science for All American*, una colaboración de tres años (en 1986 nos visitó el Cometa Halley por última vez) entre cientos de científicos, matemáticos, ingenieros y otros académicos, tuvo un impacto significativo sobre la reforma educativa al orientar el concepto de *formación científica* (aprender haciendo) y establecer las bases de los estándares educativos en ciencia, tecnología, ingeniería y matemáticas (*Science, Technology, Engineering and Mathematics*, STEM). Un proyecto con un horizonte de 75 años (el Cometa Halley volverá a brillar en el año 2061). George De Boer, director del Proyecto 2061:

> A menudo se olvida, pero es este libro el que lo puso todo en marcha e impregna todo lo demás.

Conectividad o convergencia, diversificación y complejidad creciente son, hoy, los instrumentos culturales. Pero si se siguen las noticias sobre tecnología, Inteligencia Artificial (IA) y *Big Data* (BD, macrodatos) van a la cabeza. IA y BD son la fuerza directriz detrás de las tecnologías innovadoras y disruptivas. Inteligencia Artificial es la tecnología que permite a las computadoras hacer cosas que hasta hace poco tiempo eran privativas de los humanos. Por ejemplo, las computadoras siempre han calculado; ahora aprenden y aportan conclusiones. La IA asume dos actividades: aprendizaje por máquinas y aprendizaje profundo. Lo primero implica la construcción de *software* que aprenda de los datos y aplique ese conocimiento a nuevos conjuntos de datos. El aprendizaje profundo produce redes neurales bioinspiradas en el cerebro humano, e interpreta sonidos e imágenes. La IA está huérfana sin datos, que de ellos aprende. *Big data* se refiere a cantidades masivas de datos disponibles a tal efecto. La IA no es neonata; como concepto y acción lleva décadas en el mercado. La ausencia de un inmenso caladero de datos la hacía poco eficiente. Datos y más datos de imágenes, textos, audio... hacen de la IA una actividad cuasi sin límites. Numerosas actividades se benefician de la pareja IA-BD: economía global, e-comercio, *marketing* digital, robótica (automoción, industria y fabricación, asistentes domiciliarios, medicina y salud. Respecto a este último tema, Larry Brillant, Jeremy Ginsberg *et al*. comentan:

> Al aprovechar la inteligencia colectiva de millones de usuarios, los registros de búsqueda web de Google pueden proporcionar uno de los sistemas de monitoreo de influenza –gripe– más oportunos y de mayor alcance disponibles en la actualidad. Mientras que los sistemas tradicionales requieren de 1 a 2 semanas para recopilar y procesar los datos de vigilancia, nuestras estimaciones se actualizan todos los días. Al igual que con otros sistemas de vigilancia sindrómica, los datos son más útiles como un medio para estimular una mayor investigación y recopilación de medidas directas de la actividad de la enfermedad.

En un futuro no muy lejano los libros que leamos, los *e-mails* que recibamos e incluso alguna canción que escuchemos, serán producto de Generación de Lenguaje Natural (*Natural Language Generation*, NLG). Esto es, la capacidad tecnológica de crear productos humanos mediante IA.

En 2018, Google lanzó una nueva técnica de dominio abierto (*open-source*) para entrenamiento de Procesamiento de Lenguaje Natural (*Natu-*

ral Language Processing, NLP) denominado *Bidirectional Encoder Representations from Transformers* (BERT). Bidireccional refiere la capacidad para comprender la ambigüedad del lenguaje. Difiere de otros modelos de entrenamiento porque aprende del contexto del diálogo (*text analytics*), en vez de utilizar palabras o frases.

Y en abril de 2019, Springer publicó su primera máquina generadora de libros. Por otro lado, *big data* añade *long data* (series históricas), *smart data* (datos con significado) y *fast data* (datos en tiempo real). Sirvan de ejemplos pioneros: Watson (sistema de respuesta a preguntas en dominios abiertos) y Mastor (sistema de traducción automática de voz), de IBM, Siri (asistente personal virtual) de Apple o Alexa de Amazon.

Peter Diamandis anuncia: «el futuro es más rápido de lo que usted piensa»; ahí está Moxie, un producto de Embodied Inc. Moxie incorpora tecnología en la frontera 5.0 que facilita la interacción humano-máquina. Moxie no pretende reproducir una forma humanoide, si sus características funcionales. Por ejemplo, Alexa, Aura y Google Home tienen una *wake word* y responden a una voz que proviene en una dirección específica. Moxie escucha en todo momento y en todas direcciones; ello permite mantener una conversación fluida sin interrupciones; comete fallos y aprende. Se camina a un mundo de robots amigos, compañeros y mentores, comenta Diamandis.

CULTURÓMICA

> «Inteligencia artificial y *Big Data*: una poderosa combinación para el desarrollo futuro».
> *Singularity University*

Erez Lieberman Aiden y Jean-Baptiste Michel inician su libro *Uncharted*:

> Imagínese si tuviéramos un robot que pudiera leer todos los libros en todos los estantes de todas las bibliotecas importantes, en todo el mundo. Leería estos libros a una velocidad super rápida por el de robot, y recordaría cada palabra que había leído, usando su memoria súper infalible de robot. ¿Qué podríamos aprender de este robot historiador?

Tal vez y en principio, posibles reinterpretaciones de algún hecho histórico; y teniendo en cuenta la propuesta de Rudi Keller:

> Detrás de los cambios en el lenguaje hay una «mano invisible».

Aiden y Michel echan mano de un ejemplo: Estados Unidos de Norteamérica ¿es singular o plural? Tras la Declaración de la Independencia en 1776, la primera Constitución Americana –los «Artículos de la Confederación»– fue ratificada en 1781. Por aquella fecha la «nación» era una confederación laxa de estados que operaban a modo de países independientes. El 25 de mayo de 1787 se abrió la «Convención Constitucional». La «nueva» Constitución –un documento de, aproximadamente, 4200 palabras, firmado en Filadelfia el 17 de septiembre de 1787– trata a los Estados Unidos como un plural:

> La traición contra **los** Estados Unidos consistirá únicamente en hacer la guerra contra ellos, o en adherirse a sus enemigos, prestándoles ayuda y consuelo. (Art. III, Sec. 3.ª).

Y la decimotercera enmienda, una de las tres *Reconstruction Admendments* adoptadas entre 1865 y 1870:

> Ni la esclavitud ni la servidumbre involuntaria, salvo como castigo por un delito por el cual la parte haya sido debidamente condenada, existirá dentro de **los** Estados Unidos o en cualquier lugar sujeto a su jurisdicción.

El hito con mayor resonancia en el imaginario popular sobre el desplazamiento del plural al singular es la Guerra Civil americana (1861-1865). El *Washington Post*, en 1887, recogía:

> Hubo una época hace unos años en que se hablaba de Estados Unidos en plural. Los hombres decían «los Estados Unidos son» – «los Estados Unidos tienen» – «los Estados Unidos eran». Pero la guerra cambió todo eso [...] La rendición del general Lee significó una transición del plural al singular.

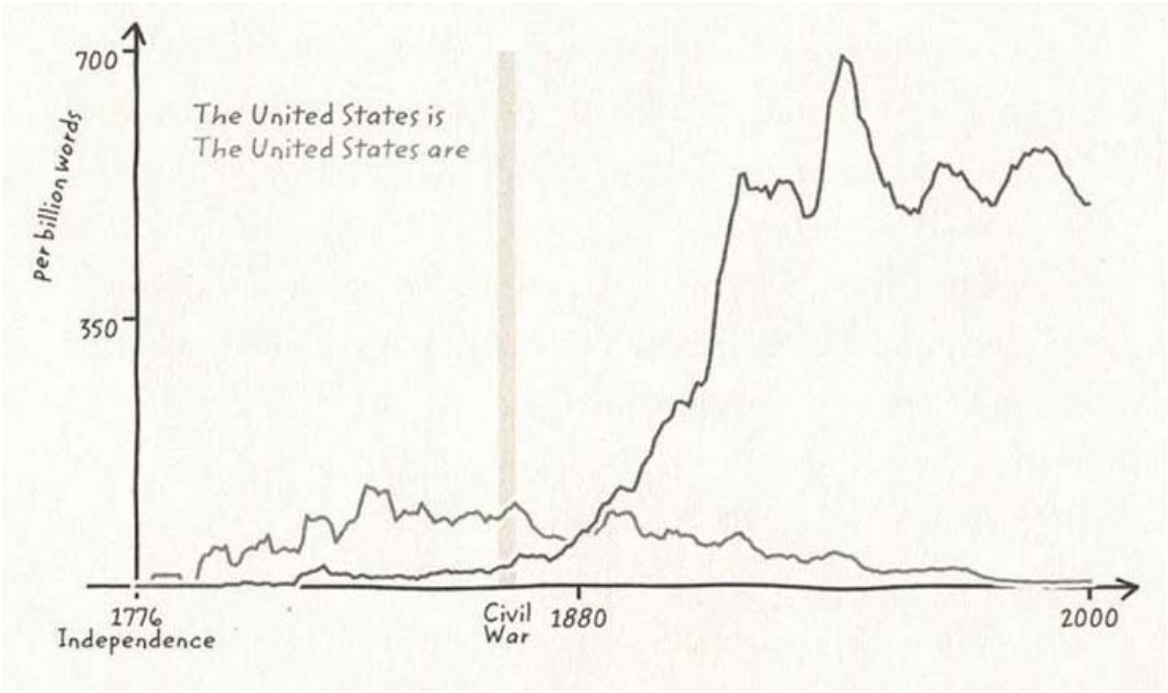
Ward W. Briggs recoge una frase de Gildersleeve, en 1909:

> «Estados Unidos son», dijo uno, «Estados Unidos es», dijo otro.

Mas, tal vez, quién más ha influido en señalar el año 1865 como la transición «plural - singular» haya sido James McPherson (n. 1936), expresidente de la prestigiosa *American Historical Association* y una leyenda entre los historiadores. Su libro *Battle Cry of Freedom: The Civil War Era* ganó el Premio Pulitzer 1989. En él puede leerse:

> Antes de 1861, las dos palabras 'Estados Unidos' generalmente se traducían como un sustantivo plural: 'los Estados Unidos son una república'. La guerra marcó una transición de los Estados Unidos a un sustantivo singular.

El criterio de autoridad de James McPherson ¿está fuera de toda duda? Erez Aiden y Jean-Baptiste Michel echan mano de su robot. El resultado se plasma en la figura que aparece en la página 4 del libro referido y aquí reproducida. El eje vertical muestra la frecuencia de ambas frases –*The United States is* y *The United States are*– en cuanto aparecen, por término medio, cada mil millones de palabras escritas durante el año en cuestión. Un gráfico como el mostrado «aclara» cuando la gente incorporó el singular en su hablar y escribir diarios.



En primer lugar, la transición fue gradual; comenzó en la década de 1810 y continuó hasta la de 1980. Más de siglo y medio. En segundo lugar, la Guerra Civil no marcó la transición. La forma singular no fue la preponderante hasta 1880. Hoy, afirman Aiden y Michel, millones de personas, en todo el mundo, se aproximan a la historia de una manera nueva, disruptiva:

> A través de los ojos digitales de un robot [...] *Big data* influye en el cambio de las humanidades; transformar las ciencias sociales y renegociar la relación entre el mundo del comercio y la torre de marfil.

Compañías como Google, Facebook o Amazon «leen» todo lo que viaja por las redes sociales. Registrar la cultura es el núcleo de su negocio. Como colectivo, la humanidad produce cinco zettabites de datos cada año: 40.000.000.000.000.000.000.000 bites (1 zetta = 10 21 bites). Esto es *big data*. La punta del iceberg. Los datos producidos por el *Homo sapiens* se doblan anualmente. *Big data* es, cada vez, más *big*. Escribe Samuel Arbesman:

> Pero no importa qué tan grandes sean esos datos o qué información obtengamos de ellos, siguen siendo solo una instantánea: un momento en el tiempo. Por eso creo que debemos dejar de atascarnos solo en *big data* y comenzar a pensar en datos a largo plazo. Por datos 'largos', me refiero a conjuntos de datos que tienen un recorrido histórico masivo, que lo llevan desde los albores de la civilización hasta el día de hoy [...] Por lo tanto, debemos agregar datos largos a nuestro conjunto de herramientas de *big data*. Pero no asuma que los datos largos son únicamente para analizar cambios «lentos». Los cambios rápidos también deben verse a través de esta lente, porque los datos largos proporcionan 'contexto' [...] La idea general de los datos largos no es realmente nueva. Campos como la geología y la astronomía o la biología evolutiva, donde los datos abarcan millones de años, dependen de largas escalas de tiempo para explicar el mundo actual. A la historia misma se le está dando el tratamiento de datos largos [...] *Big data* puede decirnos lo que necesitamos saber para los ciclos de exageración de hoy. Pero los datos largos pueden llegar a nuestro pasado... y ayudarnos a trazar un camino hacia nuestro futuro.

Big & Long data permiten a los investigadores plantear experimentos antes no soñados. Por ejemplo, analizar todos y cada uno de los libros impresos, desde el Misal de Constanza hasta el último conocido. «*This book is the story of one of those experiments*», escriben Aiden y Michel (*Uncharted*, pg. 15). La historia comenzó años antes.

1911. R. C. Eldridge publica un listado de frecuencias de [seis mil] palabras calculadas a partir del texto de un diario.

1935. Aquellas frecuencias sirvieron de base para los cálculos de George K. Zipf, recogidos en su libro *Psycho-Biology of Language*, primera de sus publicaciones sobre regularidades, ahora conocidas como Ley de Zipf; aunque representa un redescubrimiento. Algunos han sugerido que la Ley debería denominarse «Regularidad de Ayres-Condon-Dewey-Eldridge-Estoup-Hanley Joos-Zipf».

1937. Miles L. Hanley, *Word Index to James Joyce's Ulysses*. La obra de Joyce es un libro de 730 páginas que recogen un texto de 260.430 palabras, que el autor indexa alfabéticamente.

1939. George K. Zipf publica *Human Behaviour and the Principle of Least Effort.* En [2. On the Economy of Words/II. The Question of Vocabulary Balance/A. Empiric Evidence of Vocabulary Balance] puede leerse:

> La novela *Ulises* de James Joyce, con sus 260.430 palabras, representa una muestra considerable de discurso continuo que puede decirse con justicia que ha servido con éxito en la comunicación de ideas. El Dr. Miles L. Hanley y sus asociados han elaborado un índice del número de palabras diferentes que contiene, junto con las frecuencias reales de sus respectivas apariciones, con métodos ejemplares, y han argumentado con toda propiedad que todas las palabras son diferentes en algún sentido fonéticamente en la forma completamente flexionada en la que aparecen (así, las formas, *give, gives, gave, given, giving, giver, gift* –dar, da, dio, dado, dando, dador, regalo– representan siete palabras diferentes y no una palabra en siete formas diferentes). Al índice publicado anteriormente se ha agregado un apéndice de las manos cuidadosas del Dr. M. Joos, en el que se establece toda la información cuantitativa que es necesaria para nuestros propósitos presentes. Para el Dr. Joos no solo nos dice que hay 29.899 palabras diferentes en las 260.430 palabras corrientes; también clasifica esas palabras en el orden decreciente de su frecuencia de ocurrencia y nos dice la frecuencia real, *f*, con la que ocurren los diferentes rangos, *r*. Consultando este apéndice encontramos, por ejemplo, que la décima palabra más frecuente (*r* = 10) aparece 2.653 veces (*f* = 2.653); o que la palabra número 100 (*r*= 100) ocurre 265 veces (f = 265). De hecho, el apéndice nos dice la frecuencia real de ocurrencia, *f*, de cualquier rango, *r*, desde *r* = 1 hasta *r* = 29,899, que es el rango terminal de la lista, ya que el *Ulises* contiene solo ese número de palabras diferentes. Es evidente que la relación entre los distintos rangos, *r*, de estas palabras y sus respectivas frecuencias, *f*, es potencialmente bastante instructiva sobre todo el asunto del equilibrio del vocabulario, no solo porque involucra las frecuencias con las que aparecen las diferentes palabras, sino también porque el rango terminal de la lista nos dice el número de palabras diferentes en la muestra. [...] Hemos encontrado una clara correlación entre el número de palabras diferentes en el *Ulises* y la frecuencia de su uso, en el sentido de que se aproximan a la ecuación simple de una hipérbola equilátera: *r* x *f* = C, en la que *r* se refiere al rango de la palabra en el *Ulises* y *f* a su frecuencia de ocurrencia (como ignoramos para el tamaño actual de º C).” [...]
>
> RESUMEN. En cuanto a los datos empíricos mismos, hemos presentado suficiente, creo, para establecer más allá de toda duda la presencia de orden en el habla humana. Así, independientemente de las palabras físicas particulares utilizadas, y de sus significados particulares, la relación entre

> el número n de palabras diferentes y sus frecuencias f es aparentemente la misma para todo el habla. incluso, presumiblemente, si los diferentes grupos no tienen dos palabras físicas ni dos significados en común.
>
> A lo largo de nuestra discusión, señalamos dos tendencias constantes del habla. La primera tendencia fue en la dirección de reducir las magnitudes de las entidades del habla correlacionando las entidades de menor tamaño con las clases de ocurrencia más frecuente; llamamos a esta tendencia la Ley de la Abreviación. La segunda tendencia fue en la dirección de disminuir, o minimizar, el número n de diferentes clases de actividad realizadas; llamamos a esta tendencia la Ley de los Rendimientos Decrecientes (luego la llamaremos simplemente el 'n mínimo').

1946. Roberto Busa, S. J., teólogo experto en la obra de Tomás de Aquino, planteó que el estudio de la concordancia de todas las palabras (15.666.000) de la obra aquiniana podría ayudarle en su trabajo. IBM lideraba el ascenso imparable de la tecnología de computadoras. Busa intuyó que la nueva tecnología debería ser la herramienta adecuada. Presentó su proyecto al presidente de IBM, Thomas J. Watson, Jr., que asignó al ingeniero Paul Tasman al proyecto.

En 1951, en el XVIII *World Conference of Documentation* celebrado en Roma, Busa exhibió el volumen titulado: *S. Thomae Aquinatis Hymnorum Ritualium: Varia Specimina Concordantiarum: A First Example of a Word Index Automatically Compiled and Printed by IBM Punched Card Machines* (Milano: Bocca 1951).

Tras 30 años de trabajo, en 1980, los 56 volúmenes del *Index Thomisticus*, que incorpora una lematización completa del texto, veían la luz. Aquel mismo año Busa escribía:

> La vida académica actual parece estar más a favor de muchos proyectos de investigación a corto plazo que deben publicarse rápidamente, en lugar de proyectos que requieren equipos de compañeros de trabajo colaborando durante décadas. Pero, volviendo a lo que acabo de decir, para poner en práctica el procesamiento electrónico de oraciones humanas como tal, se necesita mucha más inducción. El magnífico acervo de métodos matemáticos que tenemos hoy tiene que estar basado en censos lingüísticos de textos naturales de millones de palabras. A veces se aplica una cantidad espléndida de matemáticas a una base demasiado pequeña de datos lingüísticos. Sería mucho mejor acumular resultados centímetro a centímetro sobre una base de un kilómetro de ancho, que acumular un kilómetro de investigación sobre una base de un centímetro.

El *Index* de Busa dio paso a un nuevo campo: *digital humanities* (humanidades digitales).

1996. *Stanford Digital Library Technologies Project*. Objetivo: proyectar la biblioteca del futuro. Integrar el universo de los libros en la *World Wide Web*. Tras una serie de intentos, Larry Page y Sergey Brin desarrollaron un *little search engine*; un buscador denominado BackRub que, pronto, lo denominarían Google.

2004. Google anuncia que su misión es organizar la información del planeta; en una gran parte recogida en forma de libro u otras fuentes impresas. Page y Marissa Mayer (entonces directora de productos; en 2013 CEO de Yahoo) inician la digitalización de libros. Escanear uno de 300 páginas consume 40 minutos. Cuando Mary Sue Coleman, presidente de la Universidad de Míchigan (allí se graduó Page), escuchó la pretensión de escanear los 7 millones de libros que componen su biblioteca pensó en unos mil años. Page ofreció los servicios de Google y sugirió que la tarea se podía realizar en seis años. La biblioteca de la Universidad de Michigan es una pequeña muestra de total: unos 130 millones según el catálogo creado por Google.

La siguiente tarea: implementar un sistema de escáner no-destructivo. Un artilugio similar al «dedo gordo del bibliotecario» que pasara, incansablemente día y noche, página tras página mientras las cámaras tomaran imágenes del texto. También utilizaron un proceso denominado Reconocimiento Óptico de Caracteres (*Optical Character Recognition*, OCR) por el que un programa informático localiza e identifica cada una de las letras contenidas en una imagen, a la vez que convierte la imagen digitalizada en un texto sin formato. El resultado es un archivo de texto que contiene el libro completo. Nueve años después de anunciar el proyecto Google había digitalizado 30 millones de libros; uno de cada cuatro editados desde que la imprenta de Gutenberg imprimiera el primero. El reto concluiría en 2020.

Cuando Google publicitó en 2004 su intención de digitalizar todos los libros publicados, la industria del libro se puso nerviosa. Los abogados aparecieron nada más empezar. En septiembre de 2005, la *Authors Guild*, representando a un sin fin de autores presentó la primera querella. Un mes después se presentaron los abogados enviados por la *Americam Association of Publishers* representando a las megaeditoriales McGraw-Hill, Penguin USA, Simon & Schuster, Pearson Education y John Wiley. En 2006 se unieron editoriales francesas y alemanas. En 2007, la competencia a Google, representada por Microsoft, preparó una demanda sobre la base de que «la estrategia de Google viola sistemáticamente los derechos de *copyright* y mina los incentivos de creación». *Google Book Search Settle-*

ment Agreement fue una propuesta entre Authors Guild, *Association of American Publishers* y Google. Representó el inicio de una larga batalla legal que concluyó en 2016, a favor de Google. Otras decisiones judiciales respaldaron las actividades de *Hathi Trust Digital Library* (https://www.hathitrust.org/) o del *Project Gutenberg* (https://www.gutenberg.org/). En nuestro entorno, las de la Biblioteca Virtual Miguel de Cervantes (http://www.cervantesvirtual.com/).

2005. Aiden y Michel coinciden en Harvard. El *Harvard's Program for Evolutionary Dynamics* (PED) [o ¿*Program for «rEvolutionary» Dynamics*?] representa un paraíso de creatividad, arte y ciencia fundado por el carismático matemático y biólogo Martin A. Novak. PED es un lugar donde se congregan matemáticos, lingüistas, investigadores sobre el cáncer, religiosos, psicólogos, novelistas, ... o físicos, para pensar sobre nuevas maneras de abordar el mundo. De las preguntas allí planteadas, una de ellas les llamó la atención: «*Why do we say drove and not drived*?»

Sobre la pregunta en cuestión plantearon crear una especie de lente, no para observar objetos físicos sino el cambio histórico.

El lenguaje, pensaron, es un aspecto de la cultura fácil de definir y medir. Un gran microcosmos para estudiar la cultura como un todo. Además, escriben, es uno de los más precoces antecedentes de *big data*, y la escritura, como registro fósil, de *long data*. Teniendo la Ley de Zipf como referencia, acogieron a dos doctorandos que, durante meses –aún no se había desarrollado el *Google Books*–, leyeron textos en inglés antiguo, el lenguaje de Beowulf, e inglés medieval, el lenguaje de Chaucer. Detectaron 177 verbos irregulares que pudieron rastrear más allá de mil años. Los 177 verbos comenzaron como irregulares en el inglés antiguo. En tiempos del inglés medieval, cuatro siglos después, solo 145 de las formas irregulares habían pervivido; los 32 restantes se habían regularizado. El inglés moderno mantiene, exclusivamente, 98 formas regulares. Sin embargo, ninguno de los 12 verbos más frecuentes se ha regularizado; han resistido 12 siglos de presión de la regla *-ed*. Por el contrario, de los 12 menos frecuentes, 11 de ellos han tomado la forma regular. Para Aiden y Michel los datos hablan:

> Algo similar a la selección natural ha influido en la cultura humana, dejando su huella entre los verbos.

Por supuesto, añaden, que el caso de los verbos irregulares no es equiparable a lo que sucede en la evolución biológica. En esta, un rasgo determinado puede requerir miles o millones de años para conseguir la aptitud de un organismo determinado. Para los verbos irregulares el rasgo evolutivo

lo representa la frecuencia de uso. A partir de ella puede estimarse la pauta de desaparición de un determinado verbo irregular. Parafraseando la vida media de una sustancia radiactiva, Aiden y Michel escriben:

> La fórmula era simple y hermosa: la vida media de un verbo escala como la raíz cuadrada de su frecuencia. Un verbo irregular cien veces menos frecuente se regularizará diez veces más rápido.

Por ejemplo, verbos cuyas frecuencias se sitúan entre 1 / 100 y 1 / 1000 –verbos como *drink* o *speak*– tienen una vida media de, aproximadamente, 5400 años (comparable a la vida media del C 14 [5715 años], el isótopo de referencia en la datación). El ocaso de la forma irregular drove (drive) tendrá que esperar 7800 años. Si la predicción se cumple, solo 83 de los 177 verbos irregulares seguirán siendo irregulares en el año 2500. Los resultados aparecieron en *Nature*, en octubre de 2007. Por su parte, Mark Pagel *et al*. plantean como un índice de predicción de las tasas de evolución léxica la frecuencia del uso de las palabras. Nuestra cultura, ¿obedece leyes determinísticas?

Una nueva edición del *Google Books Ngram Corpus* fue publicada un año después por Yuri Lin, Jean-Baptieste Michel, Erez Lieberman Aiden *et al*. Poco después apareció *English as she will be spoke*, en *New Scientist* por M. Erard.

Con motivo de la Exposición Mundial 1939, en Nueva York, ingenieros de *Westinghouse Electric & Manufacturing Company* enterraron una cápsula del tiempo que, entre otros objetos, contenía un manual de la lengua inglesa que describía palabras, gramática y fonética del inglés americano del siglo xx. El objetivo: que los estudiosos, 5000 años después, comprendieran el idioma de sus antepasados que, seguramente, les resultaría tan incomprensible como el hitita para nosotros. Para Michael Erard no habría que esperar tanto tiempo. Coincide con Aiden y Michel.

El inglés de Beowulf era incomprensible para Chaucer y pocos ingleses cultos en la actualidad son capaces de leer las obras originales de Shakespeare. Estamos hablando de 1600 años. Lejos de un *Panglish* los lingüistas especulan que las lenguas evolucionan a menudo por cambios explosivos, máxime cuando aparecen dialectos derivados de una lengua troncal. En tal caso, nuevos vocablos y cambios fonológicos de los existentes pueden observarse en el transcurso de una generación; ello, quizás, por un efecto fundador lingüístico o por el deseo de establecer una identidad social distintiva. También fue explosivo el fenómeno evolutivo biológico del Cámbrico.

En el mismo sentido puede interpretarse la publicación de Mark Pagel, que describe cómo un nuevo campo en expansión de los estudios filogénicos y comparativos de la evolución del lenguaje utiliza conceptos, datos y modelos estadísticos inspirados en la genética para explorar las propiedades lingüísticas de tipo similar. Pagel está interesado, según escribe:

> Luego pasaré a describir el trabajo reciente en cuatro áreas de la evolución del lenguaje en las que los enfoques de modelado estadístico han comenzado a arrojar resultados. Estos incluyen la reconstrucción de las filogenias del lenguaje y su relación con los árboles genéticos; investigaciones de la velocidad, el tempo y la profundidad temporal de la evolución del lenguaje; influencias sociales del lenguaje; y estudios de la estructura del lenguaje.

En relación con este ensayo para Chistopher Beedham y de acuerdo con las opiniones de Steven Pinker, la gramática universal de Noam Chomsky y el estructuralismo de Ferdinand de Saussure, deben considerarse tres aspectos de las irregularidades de la lengua inglesa:

> Verbos irregulares o fuertes, verbos transitivos y plurales de sustantivos irregulares.

Marc Pagel, en busca de una explicación general para la variación en las tasas de reemplazamiento estudia el «nivel de expresión» de una palabra; esto es, la frecuencia con que el vocablo se utiliza en la conversación del día a día. El lenguaje está dominado, de acuerdo con Zipf, por un número limitado de palabras utilizadas con frecuencia sobre un remanente menos empleado. Pagel *et al*. han encontrado en las lenguas indoeuropeas que las palabras que evolucionan lentamente son aquellas con los niveles de expresión más altos; las que se utilizan con mayor frecuencia en el día a día. Para el hablar cotidiano el inglés echa mano de palabras cuyo origen se remonta al inglés más antiguo. Como en el trabajo de Lieberman *et al*. donde los verbos irregulares ingleses mantienen su morfología ancestral y son, con mucho, los más empleados. Pagel *et al*. sugieren que algunos de los más persistentes replicadores culturales –memes (Ver: Susan Blackmore)– evolucionan tan lentos como genes.

Desde otro punto de vista más complejo, James M. Hughes *et al*. estudian la evolución de la literatura estudiando los patrones cuantitativos de influencia lingüística. Indican que su trabajo se relaciona, aunque de forma bastante diferente del de Michel *et al*. Tampoco faltaron opiniones

contrarias. Lingüistas y lexicógrafos expresaron su escepticismo respecto a los métodos y resultados (Ver: Ben Zimmer). Dan Cohen, director del *Roy Rosenzweig Center for History and New Media* y considerado un líder de las humanidades digitales, se refirió al *n-gran viewer* como una *gateway drug* para las humanidades digitales. Viviana Fratini *et al.*, a partir del CREA-RAE concluyeron que la correlación entre irregularidad morfológica y frecuencia no es válida para el sistema verbal español. También Anita Gerrini se muestra reticente. Tal vez, uno de los trabajos más recientes (junio 2019) sobre correlaciones entre irregularidad morfológica y frecuencia sea el de Shije Wu *et al.*, trabajo que hace referencia al inmediatamente antes citado (aunque señala sus limitaciones) e ignora el de Michel *et al.* Shijie Wu *et al.* estudian 28 lenguas; concluyen que la correlación entre irregularidad y frecuencia es más evidente [*robust*] cuando la irregularidad se considera como una propiedad de lexemas [*stems/paradigms*] más que como una propiedad de formas individuales de palabras.

2007. Aviva Aiden, esposa de Erez, fue invitada a *Googleplex* –el cuartel general de Google– para recibir un premio para *Women in Computer Science*. Erez acudió a la oficina de Peter Norvig, director de investigación de Google, con la pretensión de ampliar el estudio sobre la evolución de los verbos irregulares ingles a todas las palabras de la biblioteca digitalizada de Google:

> A Norvig no le gusta hablar mucho. De hecho, lo único más difícil de leer que la colección de libros digitales de Google es la impenetrable cara de póquer de Norvig mientras te escucha hablar [...] Después de escuchar a Erez presentar nuestro lanzamiento de una hora, Norvig finalmente mostró sus cartas. 'Todo esto suena genial, pero ¿cómo podemos hacerlo sin violar los derechos de autor?'.

Aiden y Michel se refieren en varias ocasiones a *Lexical, Loquacious Love*. Karen Reimer tomó el texto completo de una novela romántica y lo alfabetizó. Si una palabra aparece varias veces en la novela, aparece varias veces en su libro que no tiene sintaxis ni frases. Es un listado de palabras en orden alfabético recogidas en 346 páginas y 25 capítulos; no 26 porque no hay palabra alguna que empiece por «x». Todo ello partiendo de la Ley de Zipf, «la transmutación alfabética de Reimer pone de manifiesto un mundo a primera vista invisible», comentan los autores. Frecuencias de palabras, los átomos léxicos que componen la novela. Con estas dos premisas plantearon crear una base de datos en la sombra que contendría cada palabra y cada frase de todos y cada uno de los libros publicados en inglés.

Esas palabras y frases –el término elegante en ciencias de la computación es *n-gram*– incluyen 2,314159... (un 1-gram), coca cola (un 2-gram), o *the United States of America* (un 5-gram).

> Para cada palabra y frase, el registro consistiría en una larga lista de números, mostrando la frecuencia con la que ese n-grama en particular apareció en los libros, año tras año, retrocediendo cinco siglos.

Había una restricción: la estadística de Lander-Waterman. Por su relevancia en la secuenciación de genomas se han desarrollado estrategias que permiten reconstruir un texto mediante el ensamblaje de pequeños fragmentos. Erez y Jean-Baptiste encontraron la solución:

> Nuestra sombra no incluiría datos de frecuencia para palabras y frases que se hayan escrito solo unas pocas veces. Con esta modificación, reconstruir los textos completos sería matemáticamente imposible.

Tiempo después Erez y Jean-Baptiste escribieron una carta conciliadora al *The Honorable* Denny Chin, *United States District Judge*, abogando por el carácter no consuntivo de su estrategia:

> ¿Qué sugerirle a Norvig?... N-grams fue nuestra respuesta. Norvig pensó en esta idea por un minuto y decidió que podría valer la pena intentarlo. Estábamos dentro. De repente tuvimos acceso a la colección de palabras más grande de la historia.

Pero, qué es una palabra:

> Una palabra en inglés es 1-gram que aparece, en promedio, al menos una vez en cada billón de 1 gram de texto en inglés.

Tras cuatro años de trabajo Erez Lieberman Aiden, Jean-Baptiste Michel y un numeroso equipo conceptualmente transversal publicó, en enero de 2011, un artículo seminal y un exhaustivo material de soporte en línea.

Su primer objetivo: las irregularidades verbales en lengua inglesa habían quedado atrás. Google Books ofrecía nuevas expectativas. Aden y Michel tomaron un corte de sus datos: todos los libros publicados entre 1990 y 2000. Esta muestra contenía más de 50 mil millones de 1-grams. Aplicando el concepto de «palabra», el resultado fue: 1.489.337 palabras. La primera edición completa (1928) del *Oxford English Dictionary* (ODL) lista 446.000 palabras. Lo que está en un diccionario es una palabra; si no

está, no lo es. El lexicón oficial en 1990 constaba de algo más de 550.000 palabras; más que el vigente ODL. En cualquier caso, un tercio del N-grámico. Según esta comparación el 52 % de la lengua inglesa es «materia léxica oscura». Sin embargo, entre 1950 y 2000 la lengua inglesa entró en un periodo de crecimiento, casi doblando el arsenal léxico. De hecho, cerca de 8400 palabras entraron cada año (un ritmo aproximado de 20 nuevas palabras al día). El lenguaje cambia y crece. Al parecer por tres motivos: la sociedad está más interconectada; existe un progreso evidente en ciencia y tecnología, en especial la medicina; la diversificación cultural. ¿Cuál es el límite del tamaño del lexicón de una lengua determinada?

> Culturómica –*culturomics*– es la aplicación de recopilación y análisis de datos de alto rendimiento para el estudio de la cultura humana [...] Los resultados de *culturomics* son un nuevo tipo de evidencia en las humanidades. Al igual que con los fósiles de criaturas antiguas, el desafío de la culturómica radica en la interpretación de esta evidencia [...] Estos, junto con miles de millones de otras trayectorias que los acompañan, proporcionarán una gran cantidad de huesos a partir de los cuales reconstruir el esqueleto de una nueva ciencia.

concluyen los autores de la publicación seminal.

En enero de 2011, la *American Dialect Society* votó *app* como la palabra del año 2010. En la categoría *Least likely to succeed* la ganadora fue *culturomics*.

Niklas Luhmann compara el reto y las oportunidades que supone la comunicación por computadora –refiriéndose a *Google Ngram viewer*– para la sociedad actual, con lo que supuso para las sociedades arcaica y moderna el desarrollo de la escritura y de la imprenta, respectivamente. Los profesionales de las humanidades reaccionan con una mezcla de emoción y frustración. Anthony Grafton, un historiador de la Universidad de Princeton, comenta:

> La técnica es un «nuevo punto de partida» para el análisis histórico en lugar de un reemplazo. Cuando escucharon por primera vez sobre el enfoque de la «culturómica» de las humanidades, muchos académicos reaccionaron «como si esto fuera la venida del anticristo». Pero mi reacción es, ¡Dios mira esta nueva herramienta!

Y Jon Orwant, uno de los coautores del artículo de referencia:

> Este es un llamado de atención a las humanidades de que hay un nuevo estilo de investigación que complementa los estilos tradicionales.

Por su parte, Vered Silber-Varod *et al*. concluyen:

> La culturómica es un campo de investigación emergente, que se basa en métodos de análisis cuantitativo. Los autores sugieren que agregar sistemáticamente un aspecto cualitativo a un análisis culturómico puede mejorar considerablemente el potencial de obtener hallazgos perspicaces a partir del análisis del discurso de *big data*, y proporciona un enfoque para seleccionar la combinación adecuada de métodos cuantitativos y cualitativos.

O Steffen Roth:

> [...] los hallazgos sugieren adoptar una posición escéptica sobre algunos de los sentidos comunes más frecuentes de las tendencias en la diferenciación funcional y las correspondientes autodefiniciones de la sociedad.

Con motivo de la conferencia *Shared Horizons: Data, Biomedicine, and the Digital Humanities*, los *National Institutes of Health*, el *National Endowment for the Humanities* y la *National Library of Medicine* convocaron, en Maryland, en la primavera del año 2013, se citó a un grupo de investigadores con un amplio abanico de intereses: matemáticas, historia del arte, lenguas africanas, ciencia computacional, microbiología, retórica, física cuántica, poesía, zoología, música, ciencias sociales, arquitectura, astrofísica... En noviembre del 2008 la Unión Europea lanzó *Europeana*.

Emma Marris apunta que «algunos investigadores piensan que la evolución de los lenguajes puede comprenderse tratándolos como genomas, pero muchos lingüistas no quieren oír hablar de ello». En cualquier caso, Mark Pagel *et al*. sugieren que algunos de los más persistentes replicadores culturales –memes– evolucionan de manera similar a algunos genes. Culturómica incluye el sufijo *-ómica* –neologismo proveniente del inglés [*omics*] utilizado inicialmente en biología para referirse al estudio de una totalidad: genómica, proteómica...– La cultura toda es digitalizable: escritura, pintura, escultura, arquitectura, música.... Ello abre la puerta a una nueva aproximación a nuestra historia. En cualquier caso, en el artículo de *Nature*, Lieberman Erez Aiden expresa su respeto por las aproximaciones tradicionales a las humanidades:

> Creo que deberían usar los mejores métodos disponibles, y todos ellos. Y creo que eso incluye leer cuidadosamente los textos y tratar de entender lo que piensan los autores.

En septiembre de 2011, Kalev H. Leetaru, un analista de *big data* a partir de textos masivos archivados, publicó Culturomics 2.0. En la «Introducción» puede leerse:

> El campo emergente de «*Culturomics*» busca explorar amplias tendencias culturales a través del análisis computarizado de vastos archivos de libros digitales, ofreciendo nuevos conocimientos sobre el funcionamiento de la sociedad humana (Michel, *et al*., 2011). Sin embargo, los libros representan la «historia digerida» de la humanidad, escrita con el beneficio de la retrospectiva. Las personas toman medidas en función de la información imperfecta que tienen disponible en ese momento, y los medios de comunicación capturan una instantánea del entorno de información pública en tiempo real (Stierholz, 2008). Las noticias contienen mucho más que solo detalles fácticos: una variedad de influencias culturales y contextuales impactan fuertemente en cómo se enmarcan los eventos para la audiencia de un medio, ofreciendo una ventana a la conciencia nacional (Gerbner y Marvanyi, 1977). Un creciente cuerpo de trabajo ha demostrado que medir el «tono» de esta conciencia en tiempo real puede pronosticar con precisión muchos comportamientos sociales amplios, que van desde las ventas de taquilla (Mishne y Glance, 2006) hasta el mercado de valores en sí (Bollen, *et al*., 2011). ¿Puede el tono público de los datos de noticias globales pronosticar comportamientos aún más amplios, como la estabilidad de las naciones, la ubicación de los líderes terroristas, o incluso ofrecer una nueva perspectiva sobre el conflicto y la cooperación entre países, con la misma precisión con la que predice las ventas de películas o los movimientos de las acciones? Este estudio hace uso de un archivo traducido de 30 años de informes de noticias de casi todos los países del mundo, aplicando una variedad de enfoques de análisis de contenido computacional que incluyen minería de tonos, geocodificación y análisis de redes, para presentar «*Culturomics* 2.0». El enfoque tradicional de Culturomics trata cada palabra o frase como un objeto genérico sin significado asociado y mide solo el cambio en la frecuencia de su uso a lo largo del tiempo. El enfoque de *Culturomics* 2.0 presentado en este documento se centra en ampliar este modelo al dotar al sistema de un conocimiento de alto nivel sobre cada palabra, centrándose específicamente en el «tono de las noticias» y la ubicación geográfica, dada su importancia para la comprensión de la cobertura de noticias. La traducción de referencias geográficas textuales en coor-

> denadas mapeables y la cuantificación del «tono» latente de las noticias en datos numéricos computables permite explorar una clase completamente nueva de preguntas de investigación a través de los medios de comunicación que no son posibles a través del enfoque tradicional del recuento de frecuencia.

En cualquier caso, el *Google Books Corpus* ha recibido diversas críticas (Pechenick *et al.*):

> Es tentador tratar las tendencias de frecuencia de los conjuntos de datos de Google Books como indicadores de la popularidad «verdadera» de varias palabras y frases. Hacerlo nos permite sacar conclusiones cuantitativamente sólidas sobre la evolución de la percepción cultural de un tema determinado, como el tiempo o el género. Sin embargo, el corpus de Google Books sufre una serie de limitaciones que lo convierten en una oscura máscara de popularidad cultural. Una cuestión principal es que el corpus es en efecto una biblioteca, que contiene uno de cada libro. Un solo autor prolífico puede, por lo tanto, insertar notablemente nuevas frases en el léxico de Google Books, ya sea que el autor sea muy leído o no. Con esto entendido, el corpus de Google Books sigue siendo un conjunto de datos importante que se considera más parecido a un léxico que a un texto. Aquí, mostramos que una característica problemática distinta surge de la inclusión de textos científicos, que se han convertido en una parte cada vez más sustantiva del corpus a lo largo del siglo xx. El resultado es una oleada de frases típicas de los artículos académicos pero menos comunes en general, como las referencias al tiempo en forma de citas. Utilizamos métodos teóricos de la información para resaltar estas dinámicas mediante el examen y la comparación de contribuciones importantes a través de una medida de divergencia de conjuntos de datos en inglés entre décadas en el período 1800-2000. Encontramos que solo el conjunto de datos de ficción en inglés de la segunda versión del corpus no se ve muy afectado por los textos profesionales. En general, nuestros hallazgos cuestionan la gran mayoría de las afirmaciones existentes extraídas del corpus de Google Books y señalan la necesidad de caracterizar completamente la dinámica del corpus antes de usar estos conjuntos de datos para sacar conclusiones generales sobre la evolución cultural y lingüística.

NOTAS

Traducción automática mediante *Google Traductor*. Con modificaciones. https://translate.google.com/?sl=en&tl=es&op=translate

El presente texto fue motivado por la ponencia «Culturómica y Cript'ia'fasia», Sesión: *Inteligencia Artificial: El Valor de los Datos*, Madrid: Real Academia de Ingeniería, 19 junio 2019.

«Culturómica» en el sentido tratado en este ensayo se aparta por completo de «culturómica» en cuanto estrategia de cultivo microbiano de alto rendimiento; técnica utilizada para la exploración del microbioma humano (Ami Diakite, Grégory Dubourg, Niokhor Dione, Pamela Afouda, Sara Bellali, Issa Isaac Ngom et *al.*, *Nature Research* - Scientific Reports, 2020). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7295790/pdf/41598_2020_Article_66738.pdf

BIBLIOGRAFÍA

AAAS. American Association for the Advancement of Science. https://www.aaas.org/

Tomás de Aquino (1225-1274). «Doctor Angélico». Teólogo y filósofo perteneciente a la Orden de Predicadores. Considerado de máximo representante de la enseñanza escolástica. Uno de los pensadores más influyentes de la Iglesia Católica. Obra complete en la Biblioteca de Autores Cristianos.

Amazon.com, Inc. Compañía tecnológica multinacional norteamericana. Fundada por Jeffery (Jef) Preston Bezos (n. 1964), en 1994 con el nombre de Cadabra, Inc. (1994-1995). Considerada una de las cuatro grandes compañías tecnológica junto con Apple, Facebook y Google. https://secfilings.nasdaq.com/filingFrameset.asp?FilingID=13184158&RcvdDate=2/1/2019&CoName=AMAZON%20COM%20INC&FormType=10-K&View=html. **Ver**: Matthew A. Russell, *Mining the Social Web. Data Mining Facebook, Twitter, Lindekin, Google+, Github, and Moore*, 2nd. ed., O'Reily-Strata Making Data Work, 2014.

American Dialect Society.

Samuel Arbesman, «Stop hyping big data and start paying attention to "long data"», *Wired* 01.29.13.

Christopher Beedham, «Irregularity in language: Saussure versus Chomsky versus Pinker», Word 2002; 53 (3): 341-367. https://www.tandfonline.com/doi/abs/10.1080/00437956.2002.11432533

Beowulf (en español Beovulfo). Poema épico anónimo escrito en inglés antiguo (o anglosajón, hablado en Inglaterra, aproximadamente, entre los años 425 y 1125) en verso aliterativo (Aliteración es la reiteración o repetición de sonidos –fonemas– semejantes en un texto o fragmento literario). Escrito entre los siglos VIII-XII, su importancia como epopeya es equiparable al Cantar de los nibelungos germano, el Cantar del mío Cid español, la Canción de Roldán francesa o el Libro de las Conquistas irlandés.

Susan Blackmore, *The Meme Machine*, con un prólogo por Richard Dawkins, Oxford: Oxford University Press 2000. Traducción al castellano –*La Má-*

quina de los Memes, prólogo de Richard Dawkins– por Monserrat Baste-Kraan para Paidos Ibérica, Barcelona, 2000. *Ibídem*, «The Third Replicator» (The essay is the subject of this week's forum discussion among the humanists and scientists at On the Human, a project of the National Humanities Center), *The New York Times*, August 22, 2010.

Harold Bloom, *The Western Canon. The Books and School of the Ages*, New York: Harcourt Brace & Co., 1994. Traducción al español –*El Canon Occidental. La Escuela y los Libros de Todas las Épocas*– de Damián Alou para Editorial Anagrama, Barcelona,1995. En 1994, Harold Bloom publicaba *El Canon Occidental*. El Prefacio y Preludio comienza: «Este libro estudia a veintiséis escritores, necesariamente con cierta nostalgia, puesto que pretendo aislar las cualidades que convierten a estos autores en canónicos, es decir, en autoridades en nuestra cultura [...] La selección no es tan arbitraria como puede parecer.» Tras estudiar los veintiséis elegidos en relación con Shakespeare, incluye un Apéndice siguiendo el criterio de Giambatista Vico que, en sus Principios de una Ciencia Nueva, postulaba un ciclo de tres fases –Teocrática, Aristocrática, Democrática–, seguidas de un caos del cual finalmente emergería una Nueva Edad Democrática. La Edad Teocrática incluye 52 autores, 142 la Aristocrática, 159 la Democrática y 444 la Caótica; en total 797 autores. Si a ellos se suman los 26 del Canon encontramos una recopilación de 823 autores, todos ellos representantes de «una», exclusive y excluyente cultura.

John Bohannon, «Google opens books to new cultural studies», *Science* 2010; 330 (6011): 1600. https://science.sciencemag.org/content/330/6011/1600.abstract

Ibídem, «Google Books, Wikipedia, and the future of Culturomics», *Science* 2011; 331(6014): 135.

Johan Bollen, Huina Mao, Xiao-Jun Zeng, «Twitter mood predicts the stock market», *Journal of Computational Science*, 2011; 2 (1): 1-8. http://dx.doi.org/10.1016/j.jocs.2010.12.007

Ward W. Briggs, Jr., ed., *Soldier and Scholar: Basil Ñanneau Gildersleeve and the Civil War*, Charlottesville and London: University Press of Virginia, 1998, pg. 22. https://books.google.es/books?id=BlXxEVcAzQYC&printsec=frontcover&vq=grammaticalconcord&hl=es&source=gbs_ge_summary_r&cad=0#v=onepage&q=grammatical-concord&f=false

Larry Brillant, «Detecting influenza epidemics using search engine query data», *Nature* 2009; 457 (7232): 1012-1014. http://dx.doi.org/10.1038/nature07634

Roberto Busa, «The Annals of Human Computing: The Index Thomisticus», *Computers and the Humanities* 1980; 14: 83-90. http://www.alice.id.tue.nl/references/busa-1980.pdf

Cámbrico, explosion del. Stephen Jay Gould, *Wonderful Life. The Burgess Shale and the Nature of History*, New York: W. W. Norton, Co., 1898. Versión

castellana –*La Vida Maravillosa. Burgess Shale y la Naturaleza de la Historia*– de Joandomènec Ros para Editorial Crítica/Col. Drakontos, Barcelona, 1991.

Geoffrey Chaucer (1343-1400). Autor de los Cuentos de Canterbury, es considerado el poeta inglés más importante de la Edad Media y el primero en ser sepultado en el Rincón de los Poetas de la Abadía de Westminster.

Noam Chomsky, «On certain formal properties of grammars», *Information and Control* 1959; 2: 137-167. http://somr.info/lib/Chomsky_1959.pdf

Ibídem, «Review of Skinner's verbal behavior», *Language* 1959; 35 (1): 26-58. http://www.biolinguagem.com/ling_cog_cult/chomsky_1958_skinners_verbalbehavior.pdf

Jason Chumtong, David Kaldewey, «Beyond the Google Ngram viewer: Bibliographic databases and journal archives as tools for the quantitative analysis of scientific and meta-scientific concepts», *Forum International Wissenschaft* (FIW) Working Paper No. 08, Universität Bonn August 2017. https://www.researchgate.net/publication/319313734_Beyond_the_Google_Ngram_Viewer_Bibliographic_Databases_and_Journal_Archives_as_Tools_for_the_Quantitative_Analysis_of_Scientific_and_Meta-Scientific_Concepts

Cleverbot. Creada por Rollo Carpentier (1965), actual director de Existor Ltd. Cleverbot participó en una prueba de Turing, junto a personas, en el Techniche Festival 2011, siendo calificado de ser 59,3 % humano. Los participantes humanos consiguieron 63,3 %. https://www.cleverbot.com/; https://www.cleverbot.com/api/

Dan Cohen, citado en Harvard University Press / Blog, 29 June 2011. https://harvardpress.typepad.com/hup_publicity/2011/06/culturomics-close-reading-and-casaubon.html#more

Committee on Research at the Intersection of the Physical and Life Sciences, *Research at the Intersection of the Physical and Life Sciences*, Washington, D. C.: The National Academies Press. https://www.nap.edu/read/12809/chapter/1

Computacionalismo. «¿Podría pensar una máquina? ¿Podría la mente misma ser una máquina de pensar? La revolución informática transformó la discusión de estas preguntas, ofreciendo nuestras mejores perspectivas hasta ahora para las máquinas que emulan el razonamiento, la toma de decisiones, la resolución de problemas, la percepción, la comprensión lingüística y otros procesos mentales característicos. Los avances en computación plantean la posibilidad de que la mente en sí misma sea un sistema computacional, una posición conocida como teoría computacional de la mente (CTM). Los computacionalistas son investigadores que respaldan la CTM, al menos en su aplicación a ciertos procesos mentales importantes». *Stanford Encyclopedia of Philosophy*, editor principal: Edward N. Zalta. https://plato.stanford.edu/entries/computational-mind/

Conectionismo o conexionismo. «El conexionismo es un movimiento en la ciencia cognitiva que espera explicar las habilidades intelectuales utilizando redes neuronales artificiales (también conocidas como "redes neuronales"). Las redes neuronales son modelos simplificados del cerebro compuestos por un gran número de unidades (los análogos de las neuronas) junto con pesos que miden la fuerza de las conexiones entre las unidades. Estos pesos modelan los efectos de las sinapsis que unen una neurona con otra. Los experimentos con modelos de este tipo han demostrado la capacidad de aprender habilidades como el reconocimiento facial, la lectura y la detección de estructuras gramaticales simples». *Stanford Encyclopedia of Philosophy*, editor principal: Edward N. Zalta. https://plato.stanford.edu/entries/connectionism/

Constitución para los Estados Unidos de Norteamérica. https://constitutionus.com/

CREA-RAE. Corpus de Referencia del Español Actual. Banco de datos [CREA] *on line*. Contiene, según el estudio de V. Fratini *et al.*, 154.212.661 palabras procedentes de libros contemporáneos (45 % del corpus), periódicos y revistas (45 %) y transcripciones de radio y TV (10 %). Todo ello hace poco comparable los resultados con los de Lieberman *et al.* que lo retrotraen 1600 años. http://corpus.rae.es/creanet.html

Culturomics. http://www.culturomics.org. **Ver**: Zhiwen Hu, «Culturomics: Science in Culture», Open Repository on Cultural Property. Think Globally, Act Locally 2016-01-19. orcp.hustoj.com?p=2082. Harvard University Press/ Blog, «Culturomics, close reading, and casaubon», 2011. https://harvardpress.typepad.com/hup_publicity/2011/06/culturomics-close-reading-and-casaubon.html

Richard Dawkins, *The Selfish Gene*, Oxford, GB: Oxford University Press, 1976.

George De Boer. En: Kathy Wren, «Before the common core. There was Science for All Americans», *Science* 2014; 345 (6200): 1012-1013.

DialogFlow. Comprada por Google en septiembre de 2016. Es una plataforma de comprensión del lenguaje natural que permite a desarrolladores (y no desarrolladores) diseñar e integrar fácilmente interfaces de usuario conversacionales inteligentes y sofisticadas en aplicaciones móviles, aplicaciones web, dispositivos y bots. Una vez implementado, el bot continúa aprendiendo de las conversaciones con los usuarios gracias a *Machine Learning*. Incluye soporte para español y es gratuito además de estar integrado con múltiples plataformas. https://dialogflow.com/

Peter H. Diamandis, «A new era of social robots», *TechBlog*, May 30, 2021. https://www.diamandis.com/blog/embodied-moxie

Peter H. Diamandis, Steven Kotler, *The Future is Faster Than You Think. How converging technologies are transforming business, industries, and our lives – Exponential Technology Series*, USA: Simon and Schuster, 2020. https://mail.

google.com/mail/u/0/?tab=wm&ogbl#inbox/FMfcgxwDrbxDlmpDcWZ-MklNtjNdLWGrQ

Peter F. Drucker, *Innovation and Entrepreneurship*, New York Harper & Row, 1985. Perfectbound, ed. «Introduction. II», pg. 3.

R. C. Eldridge, *Thousand Common English Words: Their Comparative Frequency and what Can be Done with Them*, Clement Press, 1911 (Original: University of California; digitalizado (Google Books): enero 2008). Six Thousand Common English Word, Buffalo, NY: Clement Press, 1911.

Michael Erard, «English as she will spoke», *New Scientist* 26 March 2008. https://www.newscientist.com/article/mg19726491-300-how-global-success-is-changing-english-forever/

Europeana, «A European cultural heritage platform for all», *Digital Single Market-EU*. https://www.europeana.eu/portal/es

Existor, «Coversation data», *Cleverbot Data for Machine Learning*; January 15, 2016. https://www.existor.com/products/cleverbot-data-for-machine-learning/

Facebook. Compañía estadounidense que ofrece servicios de redes y medios sociales en línea con sede en Menlo Park, California. Su sitio web fue lanzado en febrero de 2004 por Mark Elliot Zuckerberg (n. 1984) y otros compañeros: E. Saverin, A. McCollum, D. Moskovitz y C. Hughes.

Viviana Fratini, Joana Acha, Itziar Laka, «Frequency and morphological irregularity are independent variables. Evidence from a corpus study of Spanish verbs», *Corpus Linguistics and Linguistic Theory* 2014; 10 (2): 289-314. https://pdfs.semanticscholar.org/0c91/51b788b4ac421a865b96b95e181504a9be00.pdf?ga=2.80627735.142475736.1565341358-64397969.1559468322

George Gerbner, George Marvanyi, «The many worlds of the world's press», *Journal of Communication*1 977; 27 (1): 52-66. http://dx.doi.org/10.1111/j.1460-2466.1977.tb01797.x

Anita Gerrini, «Analyzing culture with Google Books: An idea whose time has come?» *Pacific standard: The society of society*, Jun 4, 2017. https://psmag.com/economics/culturomics-an-idea-whose-time-has-come-34742

Ibídem, «Analyzing culture with Google Books: Is it Social Science?» *Pacific Standard* Aug 7, 2011.

Jeremy Ginsberg, Matthew H Mohebbi, Rajan S Patel, Lynnette Brammer, Mark S Smolinski, «Detecting influenza epidemics using search engine query data», Nature 2008; 457 (7232): 1012-1014. https://www.researchgate.net/publication/23484549_Detecting_Influenza_Epidemics_Using_Search_Engine_Query_Data

Google Inc. Empresa fundada por Larry Page y Sergey Brin el 4 de septiembre de 1998. Estrenó en Internet su motor de búsqueda el 27 de septiembre de 1999. El nombre *Google* se inspiró en el término «gúgol», nombre de un número acuñado en 1938 por Milton Sirotta, un niño de nueve años sobrino del mate-

mático estadounidense Edward Kasner, que anunció el concepto numérico en su libro *Mathematics and the Imagination* (E. K. & James Newman, New York: Simon & Schuster, 1940). 1 gúgol = 10^{100}.

Google Book Search Settlement Agreement fue una propuesta entre *Authors Guild, Association of American Publishers* y *Google*, en la resolución de *Authors Guild et al. v. Google*, querella de los primeros alegando haber infringido el *copyright* por parte de Google. El acuerdo fue propuesto inicialmente en 2008, pero aparcado en 2011 por la juez Denny Chin (*United States Circuit Judge*):

CONCLUSION. In the end, I conclude that the ASA [Amended Settlement Agreement] is not fair, adecuate, and reasonable [...] *The motion for final approval of the ASA is denied, without prejudice to renewal in the event the parties negotiate a revised settlement agree ment. The motion for an award of attorneys1 fees and costs is denied, without prejudice.*

(https://www.copyright.gov/docs/massdigitization/statements/gbs_opinion.pdf). En 2013 fue rechazada la demanda *Authors Guild et al. v. Google* (https://www.publishersweekly.com/pw/by-topic/digital/content-and-e-books/article/60006-google-wins-court-issues-a-ringing-endorsement-of-google-books.html). Finalmente, el 18 abril 2016 el Tribunal Supremo rechazó la apelación (https://www.nytimes.com/2016/04/19/technology/google-books-case.html).

Anthony Grafton, citado en John Bohannon, 2011.

James (Jim) Nicholas Gray (1944-2012), *Jim Gray Summary Home Page*. https://jimgray.azurewebsites.net/

Great Books of the Western World, 1ª ed. (54 vols.), 1952, Robert M. Hutchins, Editor in Chief. 2ª ed (60 vols.), 1990, Mortimer J. Adler, Editor in Chief. Chicago: Encyclopædia Britannica, Inc. La 1ª ed. dedicaba varios volúmenes a la cultura científica (Copérnico, Kepler, Galileo, Harvey, Newton o Farady, entre otros). Los volúmenes 55 y 56 de la 2ª ed. («20th Century Philosophy and Religion», «20th Century Natural Science») incluyen, entre otros, a Alfred N. Whitehead, Bertrand Russell, Ludwig Wittgenstein, Henri Poincaré, Max Planck, Albert Einstein, Arthur Eddington, Niels Bohr, Godfrey H. Hardy, Werner Heisenberg, Erwin Schrödinger, Theodosius Dobzhansky o Conrad H. Waddington. https://ebooks.adelaide.edu.au/l/literature/gbww/index.html

Michael Hahn, Marco Baroni, «*Tabula* nearly *rasa*: Probing the linguistic knowledge of character-level neural language models trained on unsegmented text», *arXiv1906.07285[cs.CL]*. https://arxiv.org/pdf/1906.07285.pdf

Miles L. Hanley, *Word Index to James Joyce's Ulysses*, Madison: Univ. Wisconsin Press, 1937.

James M. Hughes, Nicholas J. Foti, David C.Krakauer,Daniel N. Rockmore, «Quantitative patterns of stylistic influence in the evolution of literature»,

Proceedings of the National Academy of Sciences USA 2012; 109 (20): 7682-7686. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3356644/

Humanidades digitales. Área de la actividad académica en la intersección de las tecnologías de la computación o digitales y las humanidades. La definición del campo está continuamente reformulándose. *Debates in the Digital Humanities* (vol. 2, 2016) reconoce esta dificultad:

> Junto con los archivos digitales, los análisis cuantitativos y los proyectos de construcción de herramientas que una vez caracterizaron el campo, DH ahora abarca una amplia gama de métodos y prácticas: visualizaciones de grandes conjuntos de imágenes, modelado 3D de artefactos históricos, tesis 'nacidas digitales', activismo de eqtiqueta y su análisis, juegos de realidad alternativa, espacios de creación móviles y más. En lo que se ha llamado DH de la 'gran carpa', a veces puede ser difícil determinar con alguna especificidad qué implica, precisamente, el trabajo de humanidades digitales.

Sus orígenes se retrotraen a las décadas de los años 1930 y 1940 con los trabajos pioneros de Josephine Miles y Roberto Busa. *The Digital Humanities Manifesto* 2.0; http://www.humanitiesblast.com/manifesto/Manifesto_V2.pdf.

Centre for Digital Humanities; *The Alliance of Digital Humanities Organizations (*ADHO); https://en.wikipedia.org/wiki/Alliance_of_Digital_Humanities_Organizations

European Association for Digital Humanities; https://eadh.org/. En el ámbito hispánico, los trabajos se iniciaron en España, en 1971, con Francisco A. Marcos Marín y en México con Luis Fernando Lara, ambos vinculados a la escuela de Pisa (Italia);

Inteligencia artificial (IA). La primera definición de Inteligencia artificial se debe a John McCarthy (1927-2011), profesor en el *Dartmouth College* que, en 1956 organizó en su sede, con Marvin L. Minsky (*Harvard University*), Nathaniel Rochester (IBM) y Claude E. Shannon (*Bell Telephone Labs*.), la conocida como «Conferencia de Darmouth»:

> En principio, cada aspecto del aprendizaje o cualquier otra característica de la inteligencia se puede describir con tanta precisión que se puede hacer una máquina para simularlo (*Every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it*).

IA va unida a aprendizaje por máquinas, aprendizaje profundo, redes neurales artificiales, procesamiento del lenguaje natural, robótica, internet de las cosas y *big data*, que conforma la denominada «Sociedad 5.0 o sociedad superinteligente»; tal es el énfasis puesto por Japón para paliar los efetos del envejecimiento (Ver: M. Kovacic). Alphabet y alphaGo-DeepMind de Google, Alexa de Amazon, Siri-HomePod y CoreML de Apple, Caffe 2 y PyTorch de Facebook, o Cortana y Cognitive Toolkit de Microsoft, son los nombres comerciales más extendidos. En España, Telefónica desarrolla el

programa Aura. En resumen, se trata de la convergencia entre el espacio físico y el ciberespacio.

Interlingua. Un tipo de lenguaje artificial internacional auxiliar itálico, basado en vocablos comunes a la mayoría de los idiomas de Europa occidental y en una gramática anglorrománica simplificada. Frank P. Gopsill, *International languages: a matter for Interlingua*. Sheffield, England: British Interlingua Society, 1990.

Melvin Johnson, Mike Schuster*, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, «Google's multilingual neural machine translation system: enabling zero-shot translation», *Transactions of the Association for Computacional Linguistic* 2017; 5: 339-351. https://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00065

Martin Joos, *Appendix to Hanley's Word Index*, citado por George K. Zipf en *Human Behaviour and the Principle of Least Effort* (2. On the economy of words. II. The question of vocabulary balance. A. Empiric evidence of vocabulary balance).

James Joyce, *Ulysses*, Paris: Sylvia Beach Whitman (Shakespere and Co.), 1922. Inicialmente publicado en partes (marzo 1918-diciembre 1920) por la revista americana *The Little Review*.

Rudi Keller, *On Language Change. The Invisible Hand in Language* (Translated by Brigitte Nerlich. Original en alemán, 1990), London&New York: Routledge, 1994. Published in the Taylor &rancis e-Library, 2005. https://epdf.pub/on-language-change-the-invisible-hand-in-language.html

Mateja Kovacic, «Sociaedad 5.0: La Sociedad jaopesa superinteligente como modelo global», *Vanguardia Dossier* 2019; 71: 56-76.

Eric S. Lander, Michael S. Waterman, «Genomic mapping by fingerprinting random clones», *Genomic* 1988; 2 (3): 231-239. [«The physical map is assembled by first "fingerprinting" a large number of clones chosen at random from a recombinant library and then inferring overlaps between clones with sufficiently similar fingerprints»]. https://dornsife.usc.edu/assets/sites/516/docs/papers/msw_papers/msw-081.pdf

Kalev Leetaru, «Culturomics 2.0: Forecasting large-scale human behavior using global news media tone in time and space», *First Monday* 2011; 16 (9): https://doi.org/10.5210/fm.v16i9.3663

David W. Letcher, «Culturomics: A new way to see temporal changes in the prevalence of words and phrases», *American Institute of Higher Education* - The 6th International Conference, Charleston, SC - April 6-8, 2001.

Erez Lieberman (n. 1980). Tras casarse con Aviva Presser, en 2005, él y ella añadieron a sus apellidos «Aiden» (en Hebreo «Edén»). Firma sus trabajos: Erez Aiden, Erez Lieberman, Erez Lieberman-Aiden o Erez Lieberman Aiden. Perteneciente, entre otras, a la Division of Health Sciences and Technology, MIT, publicó

en 2009: «Comprehensive mapping of long-range interactions revealsfolding principles of the human genome», *Science* 326 (5950): 289-293. https://pdfs.semanticscholar.org/ca99/4823723e34e8b2c7c44848ad85ae2c7cf0be.pdf

Erez Aiden, Jean-Baptiste Michel, *Uncharted. Big Data as a Lens on Human Culture*, New York: Riverhead Books / Penguin Group (USA), 2013; «1. Through the looking glass», pg. 1-3.

Erez Lieberman, Jean-Baptiste Michel, Joe Jackson, Tina Tang & Martin A. Nowak. «Quantifyying the evolutionary dynamics of language». *Nature* 2007; 449 (7163): 713-716. http://www.nature.com/nature/journal/v449/n7163/full/nature06137.html

Erez Lieberman-Aiden, Jean-Baptiste Michel, To The Honorable Denny Chin, United States District Judge, September 3, 2009.

Erez Lieberman Aiden, Jean-Baptiste Michel, «Culturomics, Ngrams, and New Power Tools for Science», *Google Research Blog* Oct. 18, 201.

Yuri Lin, Jean-Baptiste Michel, Erez Lieberman Aiden, Jon Orwant, Will Brockman, Slav Petrov, «Syntactic annotation for the Google Books Ngrams Corpus», Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pg. 169-174, Jeju, Republic of Korea, 8-14 July 2012. https://www.aclweb.org/anthology/P12-3029

Niklas Luhmann, *Die Gesellschaff der Gesellscgaft, Frankfurt am Main*, 1997. Traducción: Javier Torres Nafarrete, y Darío Rodríguez Mansilla, Marco Ornelas Esquinca, Rafael Mesa Iturbe, *La Sociedad de la Sociedad*, México: Editorial Herder, S. de R. L. de C. V. / Formación electrónica: Quinta del Agua Ediciones, S. A. de C. V., 1024 pg. https://circulosemiotico.files.wordpress.com/2012/10/la-sociedad-de-la-sociedad-niklas-luhmann.pdf

Emma Marris, «The language barrier», *Nature* 2008; 453 (7194): 446-448. https://www.researchgate.net/publication/5351934_Language_The_language_barrier

John McCarthy, Marvin L. Minsky, Nathan Rochester, Claude E. Shannon, *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence*, Dartmouth, NH: August 31, 1955.

Proponemos que se lleve a cabo un estudio de inteligencia artificial de 2 meses y 10 hombres durante el verano de 1956 en Dartmouth College en Hanover, New Hampshire. El estudio debe proceder sobre la base de la conjetura de que cada aspecto del aprendizaje o cualquier otra característica de la inteligencia puede, en principio, describirse con tanta precisión que se puede hacer una máquina para simularlo. Se intentará descubrir cómo hacer que las máquinas utilicen el lenguaje, formen abstracciones y conceptos, resuelvan tipos de problemas que ahora están reservados a los humanos y se mejoren a sí mismos. Creemos que se puede lograr un avance significativo en uno o más de estos problemas si un grupo cuidadosamente seleccionado de científicos trabaja juntos durante un verano.

http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html

James McPherson, *Battle Cry of Freedom: The Civil War Era* [6th. vol., Oxford History of United States series], Oxford University Press, 1988.

Jean-Baptiste Michel, Yuan K Shen, Aviva P. Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, Erez Lieberman Aiden [pertenecen a 17 instituciones], «Quantitative analysis of culture using millions of digitized books», *Science* 2011; 331 (6014): 176-182. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3279742/. https://www.researchgate.net/publication/49688894_Quantitative_Analysis_of_Culture_Using_ Millions_of_Digitized_Books

Consultar análisis publicación en: David W. Letcher, «Culturomics: A new way to see temporal changes in the prevalence of words and phrases», *American Institute of Higher Education* - The 6 th International Conference, Charleston, SC - April 6-8, 2011 (vol. 4 (1): 228-235). https://web.archive.org/web/20160303215026/http://www.amhighed.com/documents/charleston2011/AIHE2011_Proceedings.pdf#page=228

Marc Miquel-Ribé, David Laniado, «Wikipedia culture gap: quantifying content imbalances across 40 language editions», *Frontiers in Physics* 2018; 6 / Article 54. https://www.frontiersin.org/articles/10.3389/fphy.2018.00054/full

Meme. «La nueva sopa es la sopa de la cultura humana. Necesitamos un nombre para el nuevo replicador, un sustantivo que transmita la idea de una unidad de transmisión cultural o una unidad de imitación. 'Mimene' proviene de una raíz griega adecuada, pero quiero un monosílabo que suene un poco como 'gen'. Espero que mis amigos clasicistas me perdonen si abrevio mimene a 'meme'». («*The new soup is the soup of human culture. We need a name for the new replicator, a noun which conveys the idea of a unit of cultural transmission, or a unit of imitation. 'Mimene' comes from a suitable Greek root, but I want a monosyllable that sounds a bit like 'gene'. I hope my classicist friends will forgive me if I abbreviate mimene to 'meme'*»). Richard Dawkins, *The Selfish Gene*, Oxford: Oxford University Press, 1976; pg. 206.

Misal de Constanza. Impreso en 1449 o 1450 por Johannes Gutenberg. Primer libro impreso a gran escala mediante el sistema de tipos móviles. Ostenta el carácter de icono por simbolizar el comienzo de la «Edad de la Imprenta»

Gilad Mishne, Natalie Glance, «Predicting movie sales from blogger sentiment,» *Proceedings ofAAAI-CAAW-06: AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs, 2006*. http://www.aaai.org/Papers/Symposia/Spring/2006/SS-06-03/SS06-03-030.pdf

Ngram. Un *n-gram* es una subsecuencia de n elementos de una secuencia dada. Utilizado en el estudio del lenguaje natural, en el estudio de las secuencias de genes o en el estudio de las secuencias de aminoácidos.

Peter Norvig. Stuart Russell & Peter Norvig, *Artificial Intelligence: A Modern Approach*, (A leading textbook in AI), 3rd. ed., Global edition, UK: Pearson

Education Limited, 2016. https://www.amazon.com/Artificial-Intelligence-Modern-Approach-3rd/dp/0136042597

Peter Norvig & Sebastian Thrun, 1er. MOOC: «Artificial Intelligence for Trading Program», 2011. https://sites.google.com/site/aiclass2011archive/; https://eu.udacity.com/course/intro-to-artificial-intelligence--cs271. **Ver**: Andrew Ng and Jennifer Widom, Origins of the modern MOOC (xMOOC). http://www.robotics.stanford.edu/~ang/papers/mooc14-OriginsOfModernMOOC.pdf

Martin Andrea Novak. Profesor austriaco (n. 1965), en la Universidad de Harvard desde 2003, de *Biología matemática*. Especialista en el papel de la cooperación en el proceso evolutivo. En 2001 publicó, en colaboración con Roger Highfield, *SuperCooperators: Altruism, Evolution and Why We Need Each Other to Succeed*, Simon & Schuster (Traducción –*Supercooperadores*– en Ediciones B, 2012).

Jon Orwant, citado en John Bohannon, 2010.

Larry Pagel, Sergey Brin, ver: Google Inc.

Marc Pagel, «Human language as a culturally transmitted replicator», *Nature Reviews Genetics* 2009; 10: 405-415.

Marc Pagel, Quentin D. Atkinson, Andrew Meade, «Frequency of word-use predicts rates of lexical evolution throughout Indo-European history», *Nature* 2007; 449 (7163): 717-721. https://www.researchgate.net/publication/5916092_Frequency_of_Word-Use_Predicts_Rates_of_Lexical_Evolution_Throughout_Indo-European_History

Eitan Adam Pechenick, Christopher M. Danforth, Peter Sheridan Dodds, «Characterizing the Google Books Corpus: Strong limits to inferences of socio-cultural and linguistic evolution», *PLoS ONE* 2015; 10 (10): e0137041.

Steven Pinker, *Words and Rules. The Ingredients of Language*, New York: Basic Books, 1999.

Ibídem. *The Language Instinct: How the Mind Creates Language*, New York: William Morrow and Company, 1994 / *El Instinto del Lenguaje: Como la Mente Construye el Lenguaje,* José Manuel Igoa (traductor), Madrid: Alianza Ensayo 2012.

Recompensa en IA (*chatbots*). La mayor parte de los sistemas de inteligencia artificial aprenden por refuerzo, es decir, se les premia cuando realizan una acción que les ayuda a lograr un objetivo o a completar una tarea. Este método de aprendizaje ha demostrado ser extremadamente eficaz cuando el objetivo es que la inteligencia artificial aprenda a realizar una tarea concreta. Sin embargo, no resulta tan adecuado cuando se pretende que la inteligencia artificial aprenda a ser realmente autónoma y tomar decisiones sin una orden previa directa. En el caso de los chatbots negociadores no hubo recompense por mantenerse fieles a la lengua inglesa.

Karen Reimer (Eve Rhymer), *Legendary, Lexical, Loquacious Love*, Chicago, Il: Sara Ranchouse Publishing, 1996.

Replicador. «En la discusión de la evolución, un replicador –*replicator*– es una entidad (como un gen, un meme o el contenido de un disco de memoria de computadora) que puede copiarse, incluidos los cambios que pueda haber sufrido. En un sentido más amplio, un replicador es un sistema que puede hacer una copia de sí mismo, sin copiar necesariamente ningún cambio que haya sufrido. Los genes de un conejo son replicadores en el primer sentido (un cambio en un gen puede heredarse); el conejo en sí es un replicador solo en el segundo sentido (una muesca hecha en su oreja no se puede heredar», K. E. Drexler, «Glossary», pág. 288.

Eve Rhymer (Karen Reimer)

Steffen Roth, «Fashionable functions: A Google Ngram view of trends in Functional differentiation (1800-2000)», *International Journal of Technology and Human Interaction* 2014; 10 (2): 34-58. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.692.7206&rep=rep1&type=pdf

Ferdinand de Saussere, *Cours de Linguistique Générale*, Charles Bally, Albert Sechehaye, eds., 1916 / *Course in General Linguistics*, Roy Harris (traductor), La Salle, Ill: Open Court, 1983.

Svenka Savić, *How Twins Learn to Talk: A Study of the Speech Development of Twins from One to Three*, New York and London: Academic Press, 1980. Translated into English by Vladislava Fclabov.

Shared Horizons: *Data, Biomedicine, and the Digital Humanities*, Project Dir.: Neil Fraistat, MITH-NEH-NLM Genomics Workshop, University of Maryland, August 30, 201. https://drum.lib.umd.edu/bitstream/handle/1903/14721/SharedHorizons.FinalReport.082113.pdf

Vered Silber-Varod, Yoram Eshet-Alkalai. Nitza Geri, «Culturomics: Reflections on the potential of big data discourse analysis methods for identifying research trends», *Online Journal of Applied Knowledge Management* 2016; 4 (1): 82-98. http://www.iiakm.org/ojakm/articles/2016/vol-ume4_1/OJAKM_Volume4_1pp82-98.pdf

Singularity University. [Going] *through SU changes the way you view the world and, for me, it says that you're someone who is playing a much bigger game*." - Dr. Peter H. Diamandis CoFounder & Executive Chairman. https://su.org/

Charles P. Snow, *The Rede Lecture 1959*: 1. The Two Cultures. 2. Intellectuals as Natural Luddites. 3. The Scientific Revolution. 4. The Rich and the Poor, Cambridge University Press. http://s-f-walker.org.uk/pubsebooks/2cultures/Rede-lecture-2-cultures.pdf

Ibídem, *The Two Cultures: And a Second Look. An Expanded Version of The Two Cultures and The Scientific Revolution*, Cambridge University Press 1963. *Las Dos Culturas y un Segundo Enfoque. Versión Ampliada de Las Dos Culturas y la Revolución Científica*, Madrid: Alianza Editorial, 1977.

Ibídem, *The Two Cultures*, with Introduction by Stefan Collini, Cambridge University Press/Canto, 1993.

STEM. *Science, Technology, Engineering, Mathematics.* https://www.ed.gov/stem

Katrina Stierholz, «What old news tells us that data does not: The uses of news reports in monetary policy research», *On The Record: A Forum on Electronic Media and the Preservation of News* (23 October), New York Public Library, New York City, 2008. http://www.crl.edu/sites/default/files/attachments/events/Stierholz-What%20Old%20News%20Tells%20Us%20That%20Data%20Does.pdf

Eös Szathmáry, «Chemes, genes, memes: a revised classification of replicators», C. L. Nehaniv, ed., *Mathematical and Computational Biology: Computational Morphogenesis, Hierarchical Complexity, and Digital Evolution*. *Lectures on Mathematics in the Life Science* 1999; 26: 1-10.

Tabula rasa. Francis Fukuyama, en el epígrafe *The Tabula Rasa Filled In*, escribe:

> En 1959, Noam Chomsky sugirió que había «estructuras profundas» subyacentes a la sintaxis de todos los idiomas; la idea de que estas estructuras profundas son aspectos innatos y genéticamente programados del desarrollo del cerebro es ampliamente aceptada hoy en día. Son los genes y no la cultura los que aseguran que la capacidad de aprender idiomas aparezca en algún momento del primer año de desarrollo del niño [...] [Pero] la idea de la tabula rasa es un caos. La investigación en neurociencia cognitiva y psicología ha reemplazado la pizarra en blanco con una visión del cerebro como un órgano modular lleno de estructuras cognitivas altamente adaptadas.

Ver: Michael Hahn, Marco Baroni, *Tabula casi rasa*.

Jamie Trinidad, «'Culturomics' and international law research», *ESIL Conference Papers Series* 2014; 4 (3): 1-15. https://www.academia.edu/8347117/Culturomics_and_International_Law_Research_ESIL_2014_Conference_Paper_final_draft_appears_in_ch.15_of_ESIL_Select_Proceedings_Hart_2017_?auto=download

Giambattista Vico (1668-1774), *Principi di Una Scienza Nuova D'Intorno Alla Comune Natura Delle Nazioni*, Nápoles 1744. Traducción al español –*Principios de una Ciencia Nueva sobre la Naturaleza Común de las Naciones*, I-IV– por Manuel Fuentes Benot, para M. Aguilar Editor, Buenos Aires, 1956.

William Shi-Yuan Wang (n. 1933), Prof. Emérito de Lingüística, Jefe Dept. Lenguaje y Ciencias Cognitivas, Hong Kong Polytechnic University. Citado por Tao Gong *et al.*, 2018.

Washington Post, 24 abril 1887, pg. 4.

David Weatherall, *Science and the Quiet Art. Medical Research and Patient Care*, Oxford: Oxford University Press-Oxford Medical Publications, 1995: pg. 347.

Wikipedia, «Language creation in artificial intelligence», *Wikipedia* 14 April 2019. https://en.wikipedia.org/wiki/Language_creation_in_artificial_intelligence

Word2vec (*Skim-gram*). Con modelo *Skip-Gram* lo que se quiere decir es: dado un conjunto de frases (también llamado corpus) el modelo analiza las pala-

bras de cada sentencia y trata de usar cada palabra para predecir que palabras serán vecinas. Por ejemplo, a la palabra Caperucita le seguirá Roja con más probabilidad que cualquier otra palabra. Gonzalo Ruíz de Villa, *Introducción a Word2vec* (skim-gram model). https://medium.com/@gruizdevilla/introducci%C3%B3n-a-word2vec-skip-gram-model-4800f72c871f

Palabra del Día *(Word of the Day)*, «Daily updates on the latest technology terms», *TechTarget IT Knowledge Exchange*, July 15, 2019. https://itknowledgeexchange.techtarget.com/

Shijie Wu, Ryan Cotterell, Timoyhy J. O'Donnell, «Morphological irregularity correlates with frequency», *Proceedings of the 57th. Annual Meeting ACL*, Firence, Italy 2019. https://pdfs.semanticscholar.org/934b/3c32bed67ee2b1b66ee5855394c0a372f9cf.pdf?_ga=2.72239891.142475736.1565341358-64397969.1559468322

Ben Zimmer, «Life in These, uh, This United States», *Language Log*, November 24, 2005. http://itre.cis.upenn.edu/~myl/languagelog/archives/002663.html

Ibídem, «When physicists do linguistics. "Is English 'cooling'"? A scientific paper gets the cold shoulder», *Boston Globe* February 10, 2013. https://www.bostonglobe.com/ideas/2013/02/10/when-physicists-linguistics/ZoHNxhE6uunmM7976nWsRP/story.html

George K. Zipf (1902-1950). Lingüista y filólogo norteamericano. Ocupó la jefatura del Departamento de Literatura Alemana, en Harvard, durante las décadas de los años 1930 y 1940. Estudió frecuencias estadísticas en diferentes lenguas. Es el epónimo de la Ley de Zipf, una ley empírica formulada utilizando estadística matemática que establece que mientras solo unas pocas palabras se utilizan con frecuencia la mayoría del lexicón se usan rara vez. Esta afirmación se expresa: Pn ~ 1 / n a, donde Pn representa la frecuencia de una palabra en la posición n-ésima (cuando las palabras se ordenan de mayor a menor frecuencia) y a es casi 1. Esto significa que el segundo elemento se repetirá aproximadamente con una frecuencia de 1/2 de la del primero, y el tercer elemento con una frecuencia de 1/3 y así sucesivamente. La Ley de Zipf es una ley potencial (cuando una cantidad es proporcional a otra cantidad elevada a un exponente fijo o potencia). En la Ley de Zipf las dos cantidades son rango y frecuencia, y el exponente es 1. *The Psycho-Biology of Language*, Boston: Houghton Mifflin, 1935. *Human Behaviour and the Principle of Least Effort*, Reading, MA: Addison-Wesley, 1949.

Una revisión de estas ideas en: Willem Levelt, *A History of Psycholinguistics*, Oxford: Oxford University Press, 2012. Nelson H. F. Beebe, *A Bibliography of Publications about Benford's Law, Heaps' Law, and Zipf's Law*, Salt Lake City: University of Utah, 2013. Una ley no empírica, pero más precisa, derivada de los trabajos de Claude Shannon fue descubierta por Benoît Mandelbrot. Si las cantidades pertenecen a una estructura geométrica y el exponente no es un número entero, la estructura subyacente es un fractal.