INTELIGENCIA ARTIFICIAL & BIG DATA.
CULTURA Y LENGUAJE. *
Parte I.

Pedro R. García Barreno

"*Distinguishing it from the three existing science paradigms of theory, experimentation, and computation. The techniques and technologies for data-intensive science are so different that it is worth distinguishing it as a new, fourth paradigm for scientific exploration*."
James (Jim) Nicholas Gray.

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 versión 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 mapa an whole brain, which will involve about a million petabytes of data. Astronomy, chemistry, climate studies, genetics, law, materials science, neurobiology, network theory, or paticle 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 acces to a database of billions of words across several language set and two centuries: "big-and-long data". Google's program – Ngram 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 renegociate 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 suporting a broad range of capabilities similar to those of humans. 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 supervisión (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 humans 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.

I. 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:

> "*The increasingly important of science in the provision of health care, and the difficult social and ethical issues that will stem from our newfound ability to determine our futures, makes it essential that all of us become more scientifically literate. Our politicians must undersatand the rudiments of scientific evidence, and society as a whole must be sufficiently well informed to understand how best to achieve a healthy life and participate in debating the complex issues that will continue to be posed by advances in biological and medical research. This movement toward greater scitific awareness will have to start in schools, with better support for science education*."

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

> *"A good many times I have been present at gatherings of people who, by the standards of the traditional culture, are thought highly educated and who have with considerable gusto been expressing their incredulity at the illiteracy of scientists. Once or twice I have been provoked and have asked the company how many of them could describe the Second Law of Thermodynamics. The response was cold: it was also negative. Yet I was asking something which is the scientific equivalent of: Have you read a work of Shakespeare's? I now believe that if I had asked an even simpler question — such as, What do you mean by mass, or acceleration, which is the scientific equivalent of saying, Can you read? — not more than one in ten of the highly educated would have felt that I was speaking the same language."*

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

> "*Changes in education are not going to produce miracles. The división of our culture is making us more obtuse tan we need be: we can repair communication to some extent: but, as I have said before, we are not going to turn out men and women who understand as much of our world as Piero della Francesca did of his, or Pascal, or Goethe. With good fortune, however, we can educate a large proportion of our better minds so that they are not ignorant of imaginative experience, both in the arts and in science, nor ignorant either of the endowments of aplied science, of the remediable suffering of most of their Fellow humans, and the responsibilities which, once they are seen, cannot be denied".*

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

> "*We are indeed in the early stages of a major technological transformation, one that is far more sweeping that the most ecstatic of the 'futorologists' yet realice, greater tan Megatrends or Future Shock. Three hundred yearss of techhnology came to an end after World ar II. During those three centuries the model for technology was a mechanical one: the events that go on inside astar such as the sun. This period began when an otherwise almostunknown French physicist, Denis Papin, envisaged the steamengine around 1680. They ended when we replicated in thenuclear explosion the events inside a star. For these three centuriesadvance in technology meant—as it does in mechanical processes—more speed, higher temperatures, higher pressures.* […] *Since the end of World War II, however, the model of technology has become the biological process, the events inside an organism. And in an organims, processes are not organized around energy in the physicist's meaning of the term. They are organized around information*."

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" y establecer las bases de los estándares educativos en ciencia, tecnología, ingeniería y matemáticas (*Science, Technology, Engineeringand Mathematics*, STEM). Un proyecto con un horizonte de 75 años (el Cometa Halley volverá a brillar en el año 2061). Para George De Boer, director del Proyecto 2061:

> *"It's often forgotten, but it's this book that got it all going and just pervades everything else."*

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) 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, Jeremy Ginsberg *et al*. comentan:

> "*Harnessing the collective intelligence of millions of users, Google web search logs can provide one of the most timely, broad-reaching influenza monitoring systems available to-*

> *day. Whereas traditional systems require 1–2 weeks to gather and process surveillance data, our estimates are current each day. As with other syndromic surveillance systems, the data are most useful as a means to spur further investigation and collection of direct measures of disease activity*."

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 (*natural 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* (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, o *Siri* (asistente personal virtual), de Apple. Peter Diamandis anuncia que "el futuro es más rápido de lo que usted piensa". Cultura (PARTE I) y lenguaje (PARTE II) son dos terrenos paradigmáticos para el encuentro entre inteligencia artificial y *big data*.

II. CULTURÓMICA

"*Artificial Intelligence and Big Data:*
A Powerful Combination for Future Growth"
Singularity University.

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

> "*Imagine if we had a robot that could read every book on every shelf o every major library, all over the world. It would read these books at a super-fast robot speed and remember every single word that it had read, using its super-infallible robot memory. What could we learn from this robot historian*?"

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:

> "*Treason against the United States, shall consist only in levying war against* ***them****, or in adhering to* ***their*** *enemies, giving them aid and comfort*" (Art. III, Sec. 3ª).

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

> "*Neither slavery nor involuntary servitude, except as a punishment for crime whereof the party shall have been duly convicted, shall exist within the United States, or any place subject to* ***their*** *jurisdiction*."

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:

> "*The was a time a few years ago when the United States was spoken of in the plural number. Men said 'the United States are' - 'the United States have' - the United States were'. But the war changed all that [...] The surrender of Gen. Lee meant a* ***transition from the plural to the*** singular."

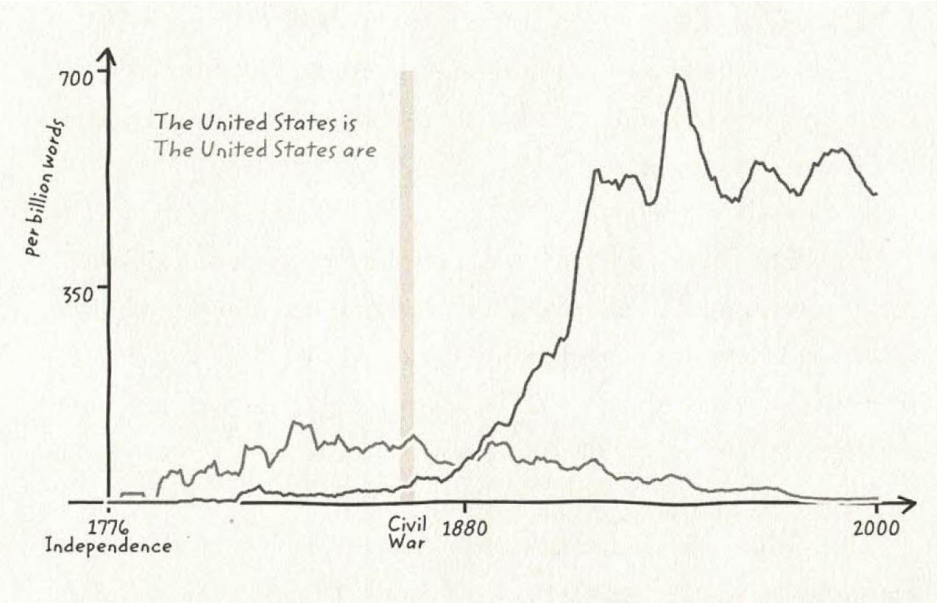
Ward W. Briggs recoge una frase de Gildersleeve, en 1909:

> "*'United States are,' said one, 'United States is' said another*."

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:

> "*Before 1861 the two words 'United States` were generally rendered as a plural noun: 'the United States are a republic' The war marked a transition of the United States to a singular noun*."

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 los años 1810 y continuó hasta los años 1980. Más de siglo y medio. También, 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:

> "*Through the digital eyes of a robot [...] Big data is going to the change of humanities, transform the social sciences, and renegotiate the relationship between the world of commerce and the ivory tower*."

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:

> "*But no matter how big that data is or what insights we glean from it, it is still just a snapshot: a moment in time That's why I think we need to stop getting stuck only on big data and start thinking about* ***long data.*** *By 'long' data, I mean datasets that have massive historical sweep – taking you from the dawn of civilization to the present day [...] So we need to add long data to our big data toolkit. But don't assume that long data is solely for analyzing "slow" changes. Fast changes should be seen through this lens, too – because long data provides 'context' [...] The general idea of long data is not really new. Fields such as geology and astronomy or evolutionary biology – where data spans millions of years – rely on long timescales to explain the world today. History itself is being given the long data treatment [...] Big data may tell us what we need to know for hype cycles today. But long data can reach into our past … and help us lay a path to our future*."

"*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 Kingsley Zipf, recogidos en su libro *Psycho-Biology of Language*, primera de las publicaciones de Zipf sobre regularidades, ahora conocidas como Ley de Zipf. Recalcar que representa un redescubrimiento de los hallazgos de otros. 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:

> "*James Joyce's novel Ulysses, with its 260,430 running words, represents a sizable sample of running speech that may fairly be said to have served successfully in the communication of ideas. An index to the number of different words therein, together with the actual frequencies of their respective occurrences, has already been made with exemplary methods by Dr. Miles L. Hanley and associates who have quite properly argued that all words are different in any way 'phonetically' in the fully inflected form in which they occur (thus the forms, give, gives, gave, given, giving, giver, gift represent seven different words and not one word in seven different form).*
>
> *To the above published index has been added an appendix from the careful hands of Dr. M. Joos, in which is set forth all the quantitative information that is necessary for our present purposes. For Dr. Joos not only tell us that there are* ***29,899 different words in the 260,430 running words****; he also ranks those words in the decreasing order of their frequency of occurrence and tell us the actual frequency,* ***f****, with which the different ranks,* ***r****, occur. By consulting this appendix we find, for example, that the 10th most frequent word (****r*** *= 10) occurs 2,653 times (****f*** *= 2,653); or that the 100th word (****r****= 100) occurs 265 times (****f*** *= 265). In fact, the appendix tell us the actual frequency of occurrence,* ***f****, of any rank,* ***r****, from* ***r*** *= 1 to* ***r*** *= 29,899, which is terminal rank of the list, since the Ulysses contains only that number of different words.*
>
> *It is evident that the relationship between the various rank,* ***r****, of these words and their respective frequencies,* ***f,*** *is potentially quite instructive about the entire matter of vocabulary balance, not only because it involves the frequencies with which the different words occur but also because the terminal rank of the list tells us the number of different word in the sample.*
>
> *[...]*
>
> *We have found a clearcut correlation between the number of different words in the Ulysses and the frequency of their usage, in the sense that they approximate the simple equation of an equilateral hyperbola:* ***r*** *x* ***f*** *=* ***C****, in which* ***r*** *refers to the word's rank in the Ulysses and* ***f*** *to its frequency of occurrence (as we ignore for the present size of º* ***C****)."*
>
> *[...]*
>
> *SUMMARY. As to the empiric data themselves, we have presented enough, I think, to establish beyond doubt the presence of orderliness in human speech. Thus, regardless of the particular physical words used , and regardless of their particular meanings, the ratio between the* ***n*** *number of different words and their* ***f*** *frequencies is apparently the same for all speech groups, even, presumably, if the different groups have no two physical words and no two meanings in common.*
>
> *[...]*
>
> *Throughout our discussion, we pointed out two consistent tendencies of speech. The first tendency was in direction of reducing the magnitudes of the speech entities by correlating the entities of smaller size with the classes of more frequent occurrence; we called this tendency the Law of Abbreviation. The second tendency was in the direction of decreasing, or minimizing, the* ***n*** *number of different classes of activity performed; we called this tendency the Law of Diminishing Returns (later we shall call it simply the '****n*** *minimum').*"

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., quién asigno 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 incopora una lematización completa del texto, veían la luz. Aquel mismo año Busa escribía:

> "*Today's academic life seems to be more in favor of many short-term research projects which need to be published quickly, rather than of projects requiring teams of coworkers collaborating for decades. But, going back to what I have just said, to put into practice the electronic processing of human sentences as such, much more induction is needed. The magnificent store of mathematical methods we have today has to be based on linguistic censuses of natural texts of millions of words. Sometimes a splendid amount of mathematics is applied to too small a base of linguistic data. It would be much better to build up results one centimetre at a time on a base one kilometre wide, than to build up a kilometre of research on a one-centimetre base*."

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 podía realizarse 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 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 mega-editoriales 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 Settlement 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, 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 del 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; los otros 79 se mantienen en el lenguaje, pero en forma regular. 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:

"*The formula was simple and beautiful: The half-life of a verb scales as the square root of its frequency. An irregular verb that is one hundred times less frequent will regularize ten times as fast*".

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 aparece "*English as she will be spoke*", en *New Scientist*. Con motivo de la Exposición Mundial 1939, en New 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 sus 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 como 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:

"*I shall then move on to describe recent work in four areas of language evolution in which statistical modelling approaches have begun to return results. These include the reconstruction of language phylogenies and their relationship to genetic trees; investigations of the rate, tempo and time-depth of language evolution; social influences of language; and studies of the structure of language*."

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 Saussurean, deben considerarse tres aspectos de las irregularidades de la lengua inglesa:

"*Irregular or strong verbs, non-passivizable transitive verbs, and irregular noun plurals*".

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 Indo-Europeas 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- 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 anteriormente 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:

> "*Norvig does not like to say much. In fact, the only thing harder to read than Google's digital books collection is Norvig's impenetrable poker face as he listen to you talk [...] After listening to Erez present our hour-pitch, Norvig finally showed his cards. 'This all sounds great, but how can we do it without violating copyright*?'".

Aiden y Muchel 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". 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 publi-

cados 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).

> "*For each word and phrase, the record would consist of a long list of numbers, showing how frequently that particular n-gram appeared in books, year after year, going back five centuries*".

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:

> "*Our shadow would not include frequency data for words and phrases that had been written only a handful times. With this modification, reconstructing the full texts would be mathematically impossible*".

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.

> "*What to suggest to Norvig*?"... "*Ngrams were our answer. Norvig thought about this idea for a minute and decided it might be worth a shot. We were in. Suddenly we had access to the biggest collection of words in history*."

Pero, qué es una palabra:

> "*An English word is a 1-gram that appears, on average, at least once in every billion 1-grams of English text*."

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 *supporting online material*.

Su primer objetivo, las irregularidades verbales en lengua inglesa había 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 *Ngrámico*. 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 arsenal léxico. De hecho, cerca de 8400 palabras entraron cada año (un ritmo 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, y la diversificación cultural. ¿Cuál es el límite del tamaño del lexicón de una lengua determinada?

"*Culturomics is the application of high-throughput data collection and analysis to the study of human culture [...] Culturomics results are a new type of evidence in the humanities. As with fossils of ancient creatures, the challenge of culturomics lies in the interpretation of this evidence [...] These, together with billions of other trajectories that accompany them, will furnish a great cache of bones from which to reconstruct the skeleton of a new science*", 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 al *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:

"*The technique is a 'new starting point' for historical analysis rather than a replacement. When they first heard about the 'culturomics' approach to the humanities, many scholars reacted 'as if this were the coming of the antichrist'. But my reaction is, God look at this new tool*!".

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

"*This is a wake-upcall to the humanities that there is a new style of research that on complement the traditional styles*."

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

"*Culturomics is an emerging research field, which relies on quantitative analysis methods. Authors suggest that systematically adding a qualitative aspect to a culturomics analysis may considerably improve the potential of gaining insightful findings out of big data discourse analysis, and provides an approach for selecting the appropriate mix of quantitative and qualitative methods*."

O Steffen Roth:

"[...] *the findings suggest adopting a skeptical position on some of the most frequent common senses of trends in functional differentiation and corresponding self-definitions of society*."

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, 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*.

Frente al "intentamos encontrar patrones matemáticos en la naturaleza" de Martin Novak, Emma Marris apunta que "algunos investigadores piensan que la evolución de los lenguajes puede com-

prenderse tratándolos como genomas, pero muchos lingüistas no quieren oir hablar de ello". En cualquier caso Mark Pagel *et al*. sugieren que algunos de los más persistentes replicadores culturales –nemes- 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. 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:

> "*I think you should use the best methods available - and all of them. And I think that includes carefully reading texts and trying to get behind what authors think*."

REFERENCIAS Y NOTAS

*Con motivo de la ponencia "Culturómica y Cript'ia'fasia", Sesión: "*Inteligencia Artificial: El Valor de los Datos*", Real Academia de Ingeniería, 19 junio 2019.Ver: www.pedrogarciabarreno.es. "4. Escritos varios. Ensayos. Integración Cultural_ III_ Culturómica.".

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 tecnológicas 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.
https://www.webpages.uidaho.edu/~stevel/504/Mining-the-Social-Web-2nd-Edition.pdf .

American Dialect Society.
https://www.americandialect.org/American-Dialect-Society-2010-Word-of-the-Year-PRESS-RELEASE.pdf

Samuel Arbesman, "Stop hyping big data and start paying attention to 'long data'", *Wired* 01.29.13.
https://www.wired.com/2013/01/forget-big-data-think-long-data/

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 literari). 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.

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, Arsitocrática, 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
John Bohannon, "Google Books, Wikipedia, and the future of Culturomics", *Science* 2011; 331(6014): 135.

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 =grammatical con-cord&hl=es&source=gbs_ge_summary_r&cad=0#v=onepage&q=grammatical-concord& f=false

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.
http://s-f-walker.org.uk/pubsebooks/pdfs/Stephen_Jay_Gould_Wonderful_Life_The_Burgess.pdf
Versión castellana –*La Vida Maravillosa. Burgess Shale y la Naturaleza de la Historia*- de Joan-domènec 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

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.

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

Culturomics.
http://www.culturomics.org.

Zhiwen Hu, "Culturomics: Science in Culture", *Open Repository on Cultural Prfoperty. Think Globally, Act Locally* 2016-01-19.
orcp.hustoj.com?p=2082

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

DPLA, *Digital Public Library of America.*
https://dp.la/

Peter Diamandis, "Over the next three months, I am beyond excited to give you a sneak peek into

my new upcoming book, *The Future is Faster Than You Think*!"
https://mail.google.com/mail/u/0/?tab=wm&ogbl#inbox/FMfcgxwDrbxDlmpDcWZMklNtjNdLWGrQ

Peter F. Drucker, *Innovation and Entrepreneurship*, New York Harper & Row, 1985. Perfectbound, ed. "Introduction. II", pg. 3.
http://www.untag-smd.ac.id/files/Perpustakaan_Digital1/ENTREPRENEURSHIP%20__Innovation%20and%20entrepreneurship.PDF.

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.

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: Eduardo Saverin, Andrew McCollum, Dustin Moskovitz y Chris 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.

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.
https://psmag.com/economics/culturomics-an-idea-whose-time-has-come-34742

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 matemático 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 infringimiento del 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.publishers week-ly.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.

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 US*A 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. https://dhdebates.gc.cuny.edu/ projects/debates-in-the-digital-humanities-2016) reconoce esta dificultad:

> "*Along with the digital archives, quantitative analyses, and tool-building projects that once characterized the field, DH now encompasses a wide range of methods and practices: visualizations of large image sets, 3D modeling of historical artifacts, 'born digital' dissertations, hashtag activism, and the analysis thereof, alternate reality games, mobile makerspaces, and more. In what has been called 'big tent' DH, it can at times be difficult to determine with any specificity what, precisely, digital humanities work entails*."

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. http://www.centrefordigitalhumanities.nl/minor-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). http://www.humanidadesdigitales.org/inicio.htm

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

Eric S. Lander, Michael S. Waterman, "Genomic mapping by fingerprinting random clones", *Genomic* 1988; 2 (3): 231-239.
https://dornsife.usc.edu/assets/sites/516/docs/papers/msw_papers/msw-081.pdf
["*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*"].

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 reveals folding 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.
https://cases.justia.com/federal/district-courts/new-york/nysdce/1:2005cv08136/273913/303/0.Pdf?ts=1253286807

Erez Lieberman Aiden, Jean-Baptiste Michel, "Culturomics, Ngrams, and New Power Tools for Science", *Google Research Blog* Oct. 18, 201.
https://www.google.com/search?source=_hp&ei=kBNHXfbTDMTYaMPXo_gH&_q=erez+_lieberman+jean+baptiste+michel+culturomics+ngrams+and+new+power+tools+for+science&oq=erez+lieberman+jean+baptiste+michel+culturomics+ngrams+and+new+power+tools+for+science&gs_l=psy-ab.3...13353.51724..52000...1.0..0.326._13944.0j80j5j2.....0....1..gws-wiz.....10..35i39j0i131j0j0i67j0i20i263j0i10j0i10i203j0i203j0i22i30j33i160j33i21j33i10.8V9GhmVx6Zo&ved=0ahUKEwj224zS3OnjAhVELBoKHcPrCH8Q4dUD CAU&uact=5.

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 July2012.
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 Soceidad*, 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.

Benoît B. Mandelbrot, The Fractal Geometry of Nature, San Francisco: W.H. Freeman, 1985. Multifractals and 1/f Noise: Wild Self-Affinity in Physics, New York: Springer, 1999.

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

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.
Suporting online material for...: www.sciencemag.org/cgi/content/full/science.1199644/DC1
Consultar análisis de la 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

Misal de Constanza. Libro impreso en 1449 o 1450 por Johannes Gutenberg. Se considera el primer libro impreso a gran escala mediante el sistema de tipos móviles. Tiene el estatus de icono por simbolizar el comienzo de la «Edad de la Imprenta»

Ngram. Un "n-grama" 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 en 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

Jon Orwant, citado en John Bohannon, 2010.

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

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

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

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.

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;sequence =1.

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/volume4_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 Co-Founder & 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. *Ibídem, The Two Cultures*, with Introduction by Stefan Collini, Cambridge University Press/Canto, 1993. *Ibídem, Las Dos Culturas y un Segundo Enfoque. Versión Ampliada de Las Dos Culturas y la Revolución Científica*, Madrid: Alianza Editorial, 1977.

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

Tokenización. El analizador léxico es la primera fase de un compilador. Su principal función consiste en leer los caracteres de entrada y elaborar como salida una secuencia de componentes léxicos que utiliza el analizador sintáctico para hacer el análisis. Convierte el programa fuente en una cadena de tokens (elemento básico del lenguaje o unidad léxica indivisible). Para reconocer el token usa un patrón, una regla que describe como se forman las cadenas que corresponden a un token. https://sites.google.com/site/compiladoresaplcrl/home

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.

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. "When physicists do linguistics. 'Is English "cooling"? A sicientific 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: $P_n \sim 1 / n^a$, donde P_n 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.
https//:books.google.es/books?id=m-XDCwAAQBAJ&pg=PT162&lpg=PT162&dq=Zipf+Martin+Joos+Hanley%27s+word+index&source=bl&ots=Dm21qs6-3B&sig=ACfU3U2rkKygmN8DBSc7uidYDlOWeuTnjQ&hl=es&sa=X&ved=2ahUKEwjmu7mk_OHjAhXy2eAKHY8uBGIQ6AEw

DnoECAgQAQ#v=onepage&q=Zipf%20Martin%20Joos%20Hanley's%20word%20index&f=false
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.

Pedro R. García Barreno, M.D., Ph.D., MBA.
de la Real Academia Española
de la Real Academia de Ciencias de España
del Comité Científico de FIDE
Madrid, septiembre 2019.