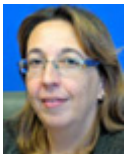
“Culturomics: una nueva forma de estudiar la frecuencia de uso de las palabras y frases y su evolución a lo largo del tiempo. Ngrams sobre trabajo, prevención y riesgo”



Talavera Pedrol, Núria
Departamento de Organización de Empresas / Universitat Politècnica de Catalunya / Av. Diagonal 647, Planta 10, Barcelona / España
+34 93 405 44 69 / nuria.talavera@upc.edu

ABSTRACT

Este artículo describe el concepto Culturomics, un nuevo campo de estudio cuyo objetivo es la cuantificación de tendencias culturales a partir del análisis de millones de libros digitalizados, y utiliza la herramienta Google Labs’ Ngram Viewer para examinar cómo ha evolucionado a lo largo del tiempo el uso de algunas palabras relacionadas con el trabajo, la prevención y el riesgo, en los libros de habla hispana digitalizados por Google durante los últimos años. La herramienta ofrece información sobre más de cinco millones de libros publicados entre los años 1800 y 2000, en siete lenguas, y permite consultar la frecuencia de uso de términos individuales y frases de hasta cinco palabras.

Palabras clave
Culturomics, Trabajo, Internet, Ngram Viewer, Tendencias

INTRODUCCIÓN

El análisis de pequeñas colecciones de textos cuidadosamente seleccionados es una técnica utilizada habitualmente por los académicos para hacer inferencias sobre las tendencias que han guiado el pensamiento humano a lo largo de la historia. En la actualidad, la evolución de las tecnologías de la información y la comunicación y el desarrollo de Internet, permiten abordar estos estudios con una orientación cuantitativa al ser posible el acceso a enormes cantidades de datos en formato digital y su posterior análisis mediante computadores.

Durante las últimas décadas **Google** ha construido un corpus de textos digitalizados que representa aproximadamente el 4% de los libros publicados en la historia de la humanidad. Este corpus contiene alrededor de 500 billones de palabras (361 billones en inglés, 45 billones en español, 45 billones en francés, 37 billones en alemán, 13 billones en chino, 35 billones en ruso y 2 billones en hebreo) extraídas de libros publicados entre los años 1500 y 2000, procedentes de más de 40 bibliotecas de universidades repartidas por todo el mundo.

En diciembre de 2010, Google pone a disposición de los internautas, y de forma gratuita, la herramienta **Ngram Viewer**, que permite interactuar con este inmenso

corpus, que jamás podría ser leído por un humano. Esta herramienta permite consultar con qué frecuencia se ha utilizado un determinado **n-gram** a lo largo del tiempo, siendo un **1-gram**, una cadena de caracteres ininterrumpidos por un espacio (pueden ser palabras, números, abreviaturas, etc.). Un n-gram es una secuencia de 1-grams.

La frecuencia de uso se calcula dividiendo el número de veces que el n-gram aparece en los textos de un determinado año por el número total de palabras en el corpus para dicho año.

CULTUROMICS: UNA DISCIPLINA EMERGENTE

Inmediatamente después del lanzamiento de la herramienta Ngram Viewer de Google, Michel, et al., (2011) publican en Science el artículo "*Quantitative Analysis of Culture Using Millions of Digitized Books"* en el que presentan el concepto Culturomics como una nueva disciplina para el análisis cuantitativo de tendencias lingüísticas, culturales y sociales con base en libros, periódicos y textos digitalizados disponibles en Internet. En este artículo, Michel et al. utilizan este enfoque para estudiar el tamaño del léxico de la lengua inglesa, la evolución de algunas formas gramaticales, y algunos fenómenos culturales, como la velocidad a la que olvidamos el pasado, los efectos de la censura, la evolución de la popularidad de los personajes célebres según su ocupación.

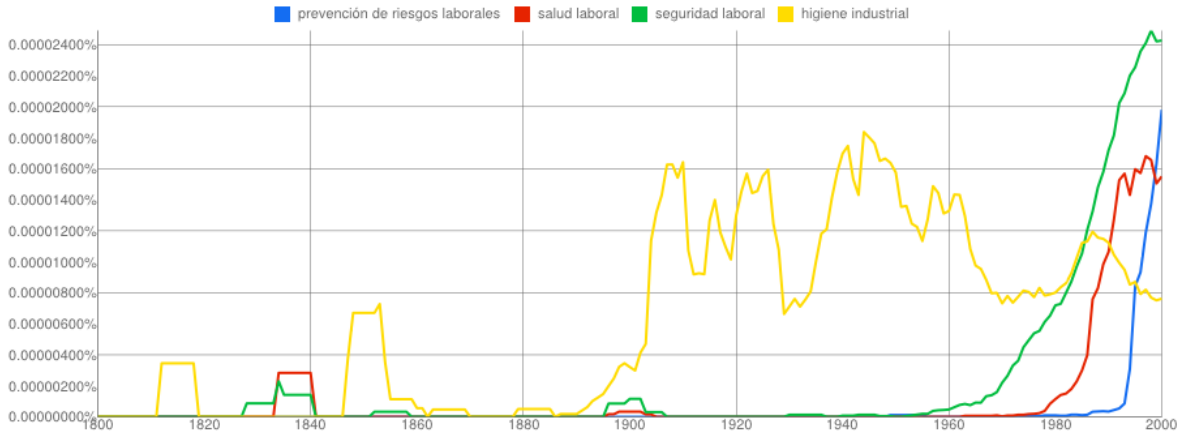
Entre sus curiosos hallazgos encontramos los siguientes: 1) la lengua inglesa crece a un ritmo de 8.500 palabras por año; 2) existen muchas más palabras que las que aparecen en cualquier diccionario; 3) los verbos irregulares evolucionan hacia formas regulares; 4) el recuerdo de los sucesos pasados se desvanece a un ritmo cada vez más rápido; 5) la duración de la fama de los personajes populares es cada vez inferior; 6) los efectos de la censura en la supresión de sucesos y personas se observan de forma evidente en los gráficos.

En Marzo de 2012 Marzo de 2012., el equipo de trabajo de Culturomics lanza un nuevo artículo en el que analizan el uso de las palabras en diferentes campos y en el que afirman que las palabras viven en un mundo competitivo, en el que tienen que pelear por su supervivencia contra sus sinónimos, variantes ortográficas y palabras relacionadas. Analizando los datos correspondientes al período 1800-2008, se encontraron patrones llamativos no sólo en inglés, sino también en español y en hebreo. Según los autores, se ha producido un cambio dramático en la tasa de natalidad y la mortalidad de las palabras: las muertes han aumentado y los nacimientos han disminuido.

Así, la lengua inglesa sigue creciendo pero a un ritmo inferior al identificado en el artículo anterior. La tasa de crecimiento se está desacelerando en parte porque el lenguaje ya es tan rico que la utilidad marginal de las nuevas palabras está disminuyendo: las cosas existentes ya están bien descritas. También se observa que las palabras que acaban de nacer gozan de mayor popularidad que las habituales, seguramente porque se utilizan para describir algo realmente nuevo (iPod, Internet, Twitter, etc.). Por otra parte, las mayores tasas de mortalidad de las palabras, son en gran medida cuestiones de homogeneización. Los correctores ortográficos evitan la aparición de versiones diferentes de la misma palabra. Los autores identifican asimismo un punto de inflexión en el ciclo de vida de las nuevas palabras: a los 30-50 años de su nacimiento pueden entrar a formar parte del léxico de largo plazo o caer en el precipicio del desuso.

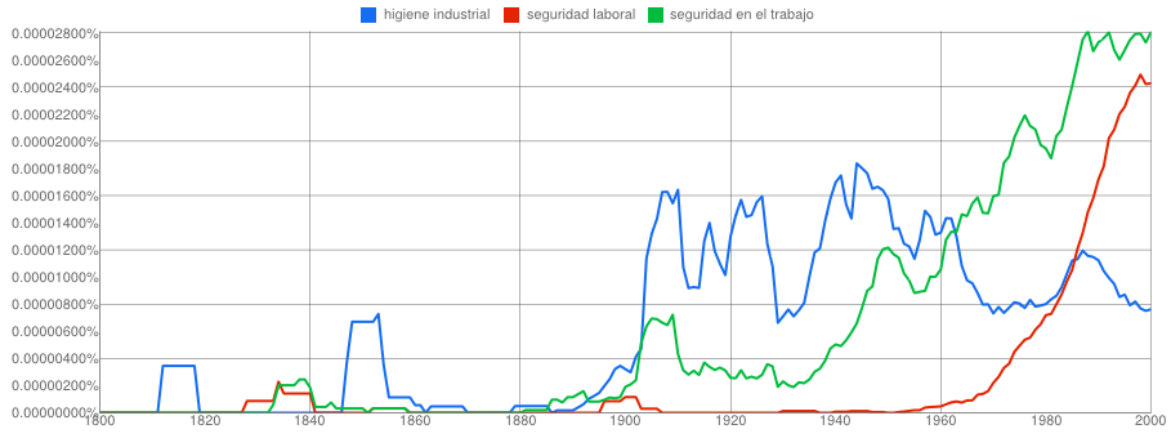
CULTUROMICS, TRABAJO Y PREVENCIÓN DE RIESGOS

Basándonos en el corpus de lengua española y en la herramienta Ngram Viewer podemos estudiar la frecuencia de aparición de algunos términos o conjuntos de términos relacionados con el trabajo y la prevención de riesgos laborales en los libros escaneados por Google. El gráfico siguiente muestra y compara la frecuencia de uso de los n-grams prevención de riesgos laborales, salud laboral, seguridad laboral e higiene industrial. Podemos ver que el 2-gram *higiene industrial* es el que ha registrado con gran diferencia más apariciones a lo largo del período estudiado (1800-2000). El 2-gram *seguridad laboral* prácticamente no aparece hasta 1960 y al cabo de unas décadas empiezan las apariciones de los términos *salud laboral* (hacia 1980) y *prevención de riesgos laborales* (hacia 1990), cuyo uso aumenta de forma espectacular (probablemente debido a la promulgación de la Ley 31/1995 de prevención de riesgos laborales en España), superando en los tres casos la frecuencia de uso de los términos *higiene industrial*.



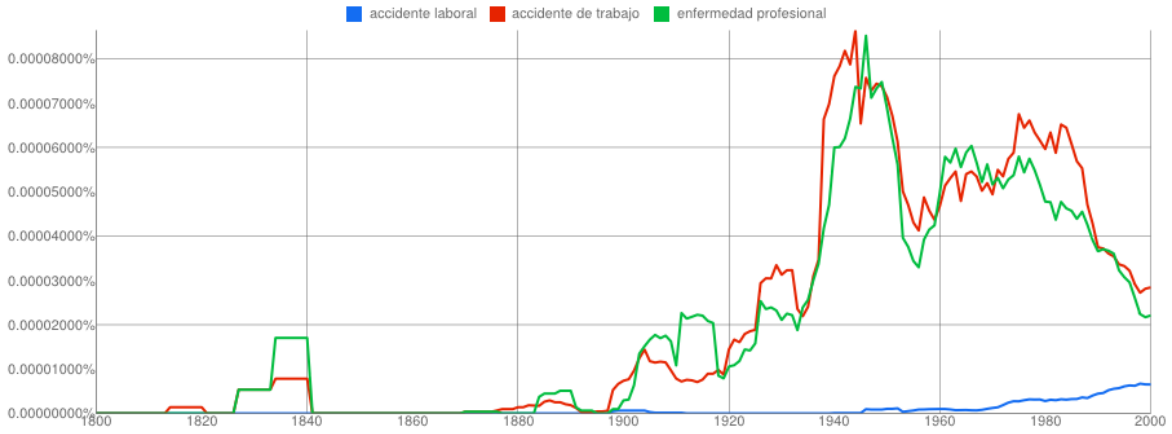
Evolución de la frecuencia de uso de los n-grams prevención de riesgos laborales, salud laboral, seguridad laboral e higiene industrial

Del gráfico anterior podríamos concluir erróneamente que la preocupación de los autores por la seguridad en el ámbito laboral no surge hasta 1960. Sin embargo si probamos con el 3-gram *seguridad en el trabajo* nos damos cuenta de que a pesar de que aproximadamente hasta 1960 su frecuencia de uso es inferior al 2-gram *higiene industrial,* su evolución es claramente creciente y a partir de esta fecha lo supera con creces. Observamos por tanto que el adjetivo *laboral* aparece y se populariza en las últimas décadas del siglo XX.



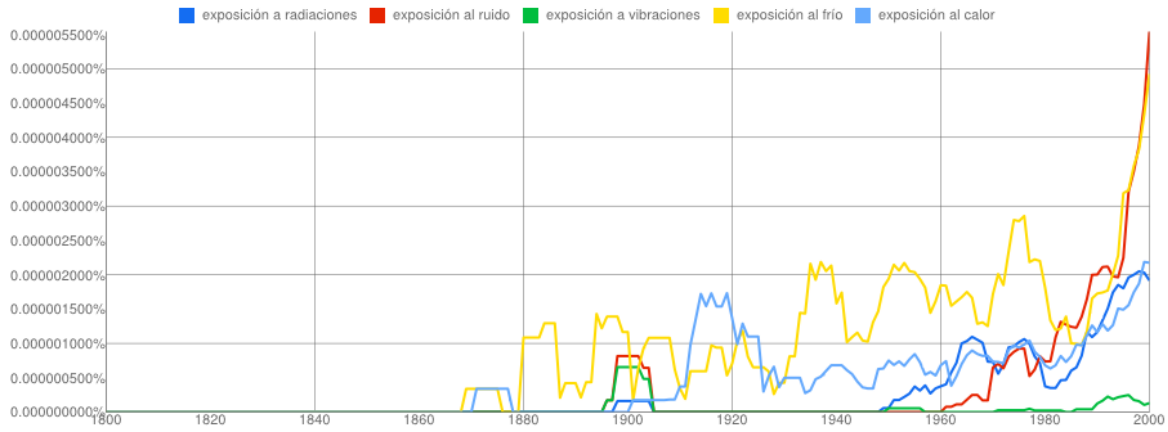
Evolución de la frecuencia de uso de los n-grams higiene industrial, seguridad laboral y seguridad en el trabajo

En el gráfico siguiente observamos que *accidente de trabajo* y *enfermedad profesional* muestran evoluciones similares con un pequeño pico entre los años 1830 y 1840, unos años de silencio y un resurgimiento a partir de 1880. El 3-gram *accidente laboral* no aparece prácticamente hasta 1970.



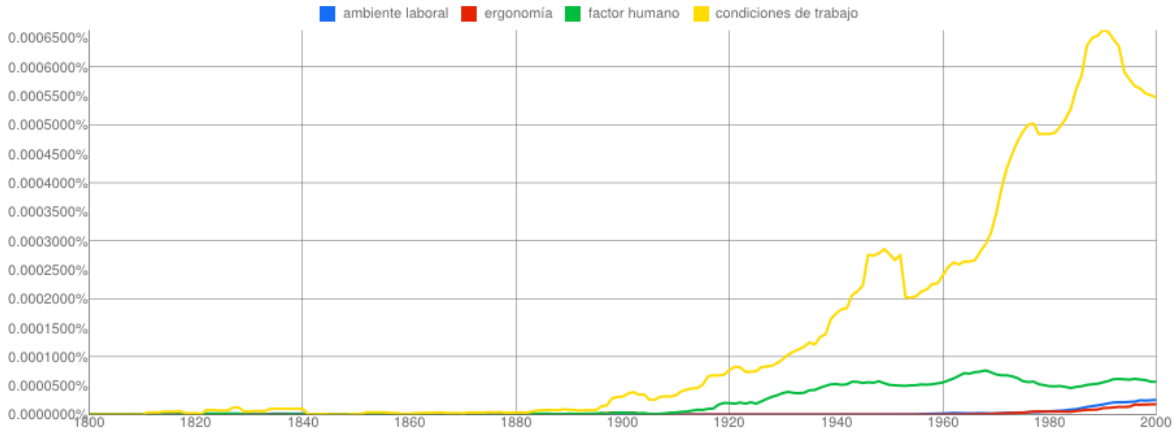
Evolución de la frecuencia de uso de los n-grams accidente laboral, accidente de trabajo y enfermedad profesional

Atendiendo a los factores de riesgo higiénico, si comparamos la evolución de la frecuencia de los 3-grams *exposición a radiaciones*, *exposición al ruido*, *exposición a vibraciones*, *exposición al frío* y *exposición al calor* podemos ver que la cadena de caracteres *exposición al frío* es la que se ha repetido con mayor frecuencia y de forma más uniforme a lo largo del tiempo, seguida de la *exposición al calor*. La exposición a radiaciones cobra interés a partir de 1950 (probablemente debido a las bombas atómicas lanzadas por Estados Unidos en 1945 contra las ciudades japonesas de Hiroshima y Nagasaki), la *exposición al ruido* crece de forma significativa a partir de 1960 y la *exposición a vibraciones* prácticamente no aparece hasta 1990 (a excepción de un curioso pico registrado alrededor de 1900).



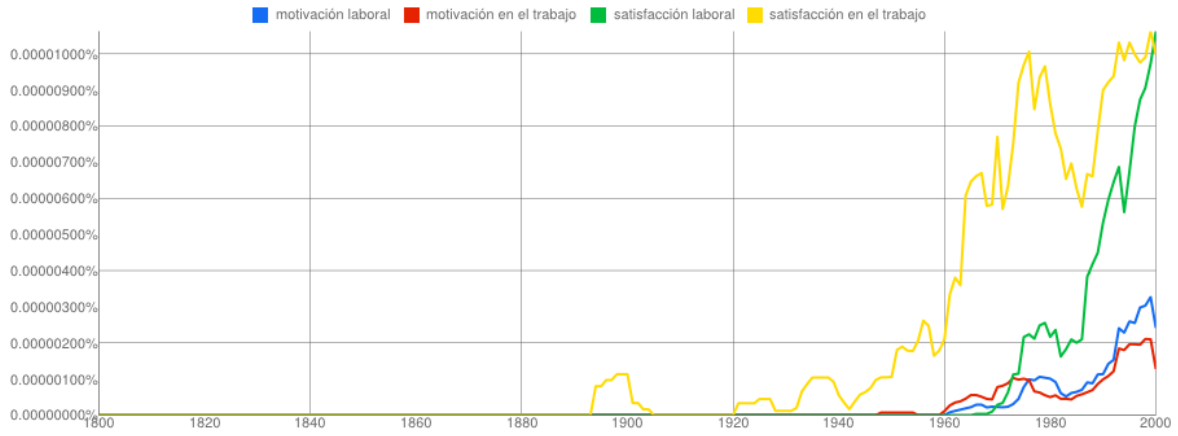
Evolución de la frecuencia de uso de los n-grams exposición a radiaciones, exposición al ruido, exposición a vibraciones, exposición al frío y exposición al calor

Respecto al ambiente laboral, la ergonomía, el factor humano y las condiciones de trabajo, observamos que el 3-gram *condiciones de trabajo* supera enormemente los otros tres conceptos. La palabra *ergonomía* y el texto *ambiente laboral* aparecen de forma tardía (alrededor de 1990) y adoptan una pendiente similar, creciente pero muy suave.



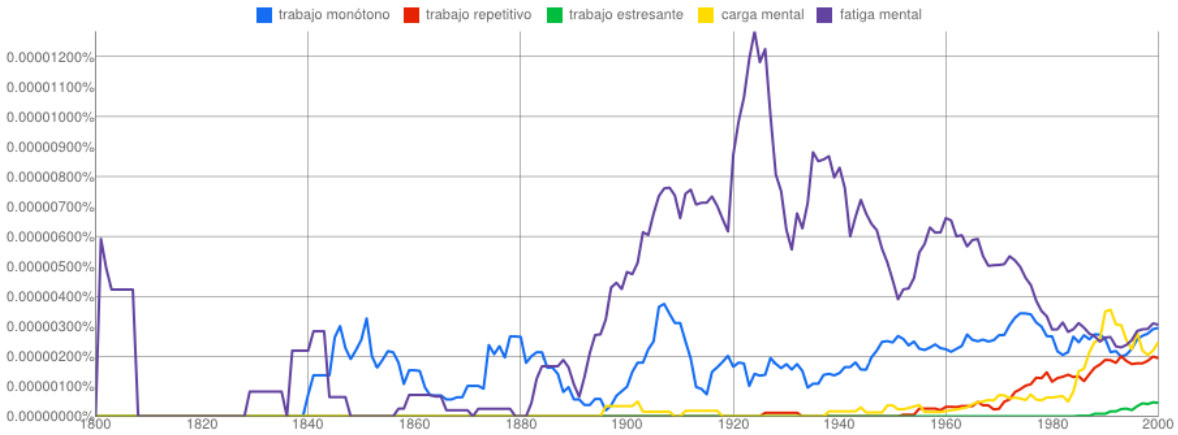
Evolución de la frecuencia de uso de los n-grams ambiente laboral, ergonomía, factor humano y condiciones de trabajo

Veamos ahora la evolución de la frecuencia de algunos factores psicosociales como la motivación y la satisfacción laboral. En el gráfico siguiente podemos ver que, en el ámbito laboral, el concepto satisfacción (*satisfacción en el trabajo* y *satisfacción laboral*) se ha utilizado antes y con mayor frecuencia que el concepto motivación (*motivación en el trabajo* y *motivación laboral*).



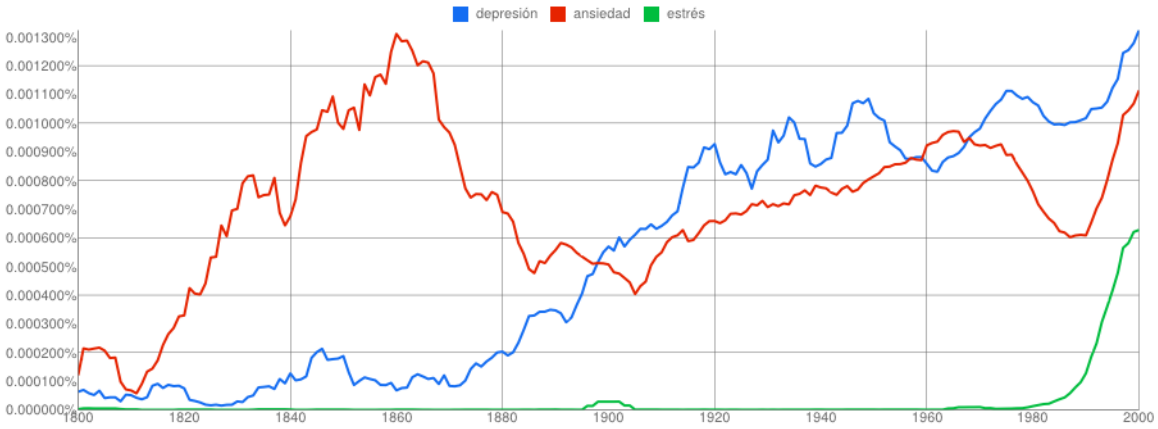
Evolución de la frecuencia de uso de los n-grams motivación laboral, motivación en el trabajo, satisfacción laboral y satisfacción en el trabajo

En el gráfico siguiente se comparan los 2-grams *trabajo monótono*, *trabajo repetitivo*, *trabajo estresante*, *carga mental* y *fatiga mental.* El conjunto de términos *fatiga mental* es el que experimenta una mayor frecuencia de ocurrencia a partir de la última década del siglo XIX, alcanzando un elevado pico alrededor de 1925 y disminuyendo su frecuencia a partir de entonces para converger con los demás términos hacia el año 2000. El 2-gram *trabajo monótomo* aparece en 1840 y se mantiene oscilando alrededor de un valor medio ligeramente ascendente. Finalmente los 2-grams *trabajo repetitivo* y *carga mental* aparecen con pendiente creciente alrededor de 1960 y el conjunto *trabajo estresante* a partir de 1990.



Evolución de la frecuencia de uso de los 2-grams trabajo monótono, trabajo repetitivo, trabajo estresante, carga mental y fatiga mental

A continuación, y par acabar con los ejemplos, podemos ver la evolución de la frecuencia de uso en los libros escaneados de los 1-gram *depresión*, *ansiedad* y *estrés*. Observamos que en el siglo XIX se escribe muchísimo más sobre la ansiedad que sobre la depresión y que, a principios del siglo XX, el término *depresión* cobra protagonismo y a partir de ahí se mantiene prácticamente siempre por encima del término *ansiedad*. El término estrés aparece en las últimas décadas del siglo XX.



Evolución de la frecuencia de uso de los 1-grams depresión, ansiedad y estrés

RESUMEN

Los continuos avances en computación y tecnologías de la información y comunicación han generado la oportunidad de manipular extensos conjuntos de datos culturales que hasta el momento eran inaccesibles para el análisis. El concepto *Culturomics* se refiere al estudio del comportamiento humano y las tendencias culturales mediante la recolección y el análisis cuantitativo de textos digitalizados. La herramienta de Google Ngram Viewer permite obtener en unos segundos, información sobre el uso que se ha hecho de un conjunto de cadenas de caracteres en el corpus de libros digitalizados por la empresa (un 4% de los libros publicados en la historia de la humanidad).

En el ámbito de la prevención de riesgos laborales y áreas afines nos puede ayudar a comprender cómo ha evolucionado la preocupación de la sociedad por los temas relacionados con la seguridad y la salud en el trabajo, a conocer el impacto que han tenido determinados sucesos (accidentes, descubrimientos, promulgación de nuevas normativas, etc.) en la 'memoria' colectiva, a identificar tendencias y, quizás en un futuro, a anticipar respuestas.

Sin duda nos encontramos ante una nueva forma de abordar el estudio de las ciencias de la humanidad cuyo verdadero reto será la interpretación de las evidencias observadas al analizar los datos. Por otra parte, a medida que el corpus aumente su tamaño, incorporando un mayor porcentaje de libros digitalizados y otros textos (periódicos, manuscritos, mapas, etc.) aumentará también su capacidad de reflejar la realidad de nuestra sociedad

REFERENCIAS

Michel J. B., Shen Y. K., Aiden A. P., Veres A., Gray M. K., Google Books Team. Pickett J. P., Hoiberg D., Clancy D., Norvig P., Orwant J., Pinker S., Nowak M. A., Aiden

E. L. (2011). Quantitative analysis of culture using millions of digitized books. Science 331, 176–182.

M. Petersen, J. Tenenbaum, S. Havlin, and H. E. Stanley, (2012) "Statistical Laws Governing Fluctuations in Word Use from Word Birth to Word Death," Nature Scientific Reports 2, 313.

Letcher, David W. (2011). "Culturomics: A New Way to See Temporal Changes in the Prevalence of Words and Phrases". American Institute of Higher Education 6th International Conference Proceedings 4 (1): p. 228.

Leetaru, Kalev H. (2011). "Culturnomics 2.0: Forecasting Large-Scale Human Behavior Using Global News Media Tone In Time And Space". First Monday 16 (9)

Bohannon, John (14 January 2011). "Google Books, Wikipedia, and the Future of Culturomics". Science 331 (6014): p. 135

Schwartz, Tim (1 April 2011). "Culturomics: Periodicals Gauge Culture's Pulse". Science 332 (6025): pp. 35–36

Morse-Gagné, Elise E. (1 April 2011). "Culturomics: Statistical Traps Muddy the Data". Science 332 (6025): p. 35.