UNICIENCIA 24
pp. 133-141
2010

VARIACIÓN DEL ÍNDICE DE NIEBLA USANDO UN *CORPUS* OBTENIDO A PARTIR DE LOS LIBROS DIGITALIZADOS POR GOOGLE

Felipe Ovares Barquero
Escuela de Informática, Universidad Nacional

José Alberto Rubí Barquero
Escuela de Filosofía, Universidad Nacional

RESUMEN

En lingüística, principalmente en el idioma inglés, se usa el Índice de Niebla de Gunning para determinar la legibilidad de un texto. El índice estima los años de educación formal necesarios para comprender el texto en una primera lectura. Un Índice de 11 años apunta a una persona con el colegio finalizado, (Gunning, 1973). Analizamos en esta investigación la variación del Índice al cambiar la forma de obtener uno de los parámetros. En la fórmula original se consideran "palabras complejas" las que tienen tres o más sílabas. En su lugar utilizamos "palabras desconocidas" que son aquellas cuyo uso es poco familiar, según un corpus construido durante la investigación, partiendo de millones de libros digitalizados por Google y la Universidad de Harvard. Aunque la variación de los resultados dependerá del valor asignado para determinar si una palabra es desconocida la investigación es pionera en el uso de un corpus para calcular el Índice de Niebla.

Palabras clave: Índice de Niebla Gunning, análisis de textos, *corpus*, facilidad de lectura, sílabas, lingüística computacional, digitalización de libros, minería de datos.

ABSTRACT

In linguistics, especially in the English language, the Gunning Fog Index is used to determine the readability of text. The said Index estimates the number of years of formal education needed to comprehend text on the first reading. Therefore, the resulting index of 11 years describes a high school graduate, (Gunning, 1973). In our study we analyzed the variation of the Index by changing the way one of the parameters is obtained. In the original formula "complex words", those which contain three or more syllables are considered. Instead, we used "unknown words", those which use is not very familiar, according to a corpus built during the study, constituting of millions of books digitized by Google and Harvard University. Although the variation of the results will depend on the assigned value to determine if a word is unknown, the study is pioneer in the use of a corpus to calculate the Fog Index.

Keywords: Gunning Fog Index, Text Analyzer, *Corpus*, Readability, Syllables, Computational Linguistic, Book Scanning, Data Mining.

INTRODUCCIÓN

El análisis de textos, en cualquier idioma, es una práctica que sigue creciendo desde la incursión de las computadoras como herramientas para el procesamiento de textos. Los procesadores, como el Word de Microsoft, efectúan el análisis del texto que el usuario va introduciendo. Subraya con rojo las palabras que pueden tener errores ortográficos y con verde las que podrían tener problemas de congruencia, entre otros. Estas facilidades, bien atendidas, reducen al mínimo los errores. Sin embargo, otros análisis podrían incorporarse para adaptar los textos al nivel de comprensión de los lectores. Vamos a presentar algunos algoritmos y fórmulas para el tratamiento de textos que extienden otros existentes.

Algunos lenguajes de programación como C/C++, PHP, entre otros, ofrecen una amplia variedad de facilidades para el manejo

Recibido 15 de noviembre, 2010– aceptado 20 de noviembre, 2010

de hileras de caracteres. Gracias a estas características, la lingüística computacional se ha extendido rápidamente para constituirse en un campo de investigación importante. Nos interesa para esta investigación el *corpus* lingüístico asistido por computadoras y obtenido a partir de millones de libros digitalizados por Google. Un *corpus* lingüístico es un conjunto, normalmente muy amplio, de ejemplos reales de uso de una lengua, en nuestra investigación este conjunto está compuesto por las palabras del idioma español (Cantos, 2001).

DEL CARÁCTER AL PÁRRAFO

Cuando se escribe una columna para un diario lo usual es atenerse a cierta cantidad de caracteres. Por ejemplo, 4 500 caracteres, incluyendo espacios en blanco, es una columna típica. Podría someterse a una cantidad de palabras, digamos 750. Sin embargo, el conteo de caracteres se adapta mejor al espacio de la columna casi sin cambios notables en el texto original. Cuando se trata de palabras, un texto de 750 oscila entre los 4 200 y los 4 800 caracteres. Los editores de texto muestran la cantidad de palabras y de caracteres. Word©, el editor de Microsoft, presenta en el análisis, además de la cantidad de palabras y de caracteres con y sin espacios, la cantidad de páginas, de párrafos y de líneas. Esta información es útil para adecuarse a ciertos formatos. En las últimas versiones de Word, la cantidad de páginas y de palabras está siempre en la esquina inferior izquierda. Word también ofrece estos datos para segmentos de texto marcados. Algunos escritores profesionales suelen someterse al rigor de escribir una cantidad de palabras cada día, por ejemplo, el escritor norteamericano Stephen King (2000) afirma que su cuota es de 2 000. Una novela promedio tiene 50 000 palabras impresas en unas 200 páginas. El editor minimalista FocusWriter 1.3.1 permite configurar, en las preferencias, el objetivo diario cuyo valor por defecto es 1 000.

Los editores ofrecen otras facilidades, tal es el caso de la revisión de la ortografía. Cada vez que se introduce una palabra en el texto, el editor revisa si esa palabra la tiene en su diccionario, si no existe la subraya con rojo para que el usuario tome alguna decisión. Otra de las facilidades consiste en mostrar sinónimos cuando el usuario lo solicita. También analiza la concordancia. Con estas herramientas lo normal sería que un texto acabe sin estos errores. Esto no implica la presencia de otros problemas que un filólogo hallaría al darle una ojeada. Por ahora, este trabajo es para un experto humano. Los editores no lo pueden hacer, pero en el futuro se irán acercando a este objetivo. Algunas ayudas son sencillas, por ejemplo, el editor podría indicarle al usuario: (1) la palabra "que" está siendo muy usada y marcarlas. (2) la frase tiene demasiadas palabras. (3) la palabra recién introducida no es usual según el *corpus* del editor.

ANÁLISIS DE COMPRENSIÓN DE UN TEXTO

Algunos periódicos, como el *New York Times* y la revista *Time*, publican sus artículos luego de analizarlos y ajustarlos. El objetivo consiste en que sean entendibles por lectores con cierta cantidad de años de educación, definida según sus expectativas. Para este análisis consideré el Índice de Niebla de Gunning[1], (Gunning, 1973). Veamos de qué se trata.

Al escribir un texto es deseable que la lectura sea comprensible para la mayoría de los lectores. Constituye una tarea fundamental para los editores de los periódicos y de los libros. En general, resulta una práctica sana. Existen, por supuesto, notables excepciones de escritores que disfrutan escribiendo para un público al cual le puede ofrecer textos complejos. No es el caso habitual. En los agradecimientos de su libro *Historia del tiempo* Stephen W. Hawking (1988) cuenta que alguien le dijo que cada ecuación incluida reduciría las ventas a la mitad. Al final, no resistió la tentación, agregó la famosa equivalencia entre la masa y la energía dada por la expresión de la teoría de la relatividad de Albert Einstein $E=mc^2$. Y comenta: "Espero que esto no asuste a la mitad de mis potenciales lectores".

Si nos colocamos en la situación de los periódicos, vemos que ellos deben garantizarse un público meta. Este objetivo se logra sometiendo los textos a un análisis. Al definir el lector meta, por ejemplo, mediante algún grado de

1 *Gunning Fox Index*. Podríamos interpretarlo así: ¿Cuán nublado está un texto?

escolaridad, velan para que sus publicaciones cumplan con ese requisito.

El **Índice de Niebla de Gunning**, un buen intento, es una prueba de lectura que permite conocer el grado de dificultad de un texto. El resultado nos brinda un nivel de lectura o escolaridad. Cuanto mayor sea el valor obtenido, mayor será la dificultad para entender el texto.

El índice se calcula con el siguiente algoritmo (Gunning, 1973):

1. Seleccione un texto, uno o más párrafos completos, de alrededor de 100 palabras. No omita oraciones. Las oraciones terminan en un punto.

2. Determine la longitud media de las oraciones. Divida el número de palabras por el número de oraciones.

3. Cuente las palabras "complejas", las que tienen más de tres sílabas.

4. Añadir la longitud media de la oración y el porcentaje de palabras complejas.

5. Multiplique el resultado por 0,4.

$$\text{Índice} = 0{,}4 \left(\left(\frac{\textit{palabras}}{\textit{oraciones}} \right) + 100 \left(\frac{\textit{palabras complejas}}{\textit{palabras}} \right) \right)$$

El Índice de Niebla da la cantidad de años de educación que el lector requiere para entender el texto. El índice castiga las frases largas con palabras complejas de tres o más sílabas. Un texto comprensible, por lo tanto, se fundamenta en lo contrario, frases cortas y palabras breves.

Como referencia, el *New York Times* tiene un Índice de Niebla promedio de 11 a 12 años y la revista *Time* sobre 11. La documentación técnica, en general, oscila entre 10 y 15, y la profesional casi nunca supera los 18 *Usingenglish.com glossary definition*. (13 de enero de 2011). Recuperado de <http://www.usingenglish.com/showdef.php?p=fog-index.html>.

Marcamos, en los siguientes ejemplos, las palabras complejas en *itálica* para resaltarlas:

Ejemplo 1: Un pequeño texto, dos haikus de Ovares (2010) para mostrar el cálculo de la función

1
Sapos en coro,
celebrando la lluvia,
toda la noche.

2
En una gota,
por la rama del rosal,
cae la luna.

Nº.	Oraciones	Palabras	Complejas
1	Sapos en coro, *celebrando* la lluvia, toda la noche.	9	1
2	En una gota, por la rama del rosal, cae la luna.	11	0
	Totales	20	1
	Promedio	10	
	Índice de Niebla	**6**	

Tabla 1. Análisis del ejemplo 1

El Índice de Niebla = 0.4 ((20 / 2) + 100 (1/20)) = **6**

En el texto (Ovares, 2007) anterior tenemos dos oraciones que constan de 9 palabras cada una. Una de ellas se considera "palabra compleja" porque tiene más de 3 sílabas: *ce-le-bran-do*. Es un texto que requiere, según el cálculo del Índice de Niebla, 6 años de escolaridad.

Ejemplo 2: Para este otro ejemplo, analizamos las dos primeras oraciones del capítulo I[2] de *El ingenioso hidalgo de Quijote de la Mancha*

2 El título de este capítulo es: *Que trata de la condición y ejercicio del famoso y valiente hidalgo don Quijote de la Mancha.*

En un lugar de la Mancha, de cuyo nombre no quiero acordarme, no ha mucho tiempo que vivía un hidalgo de los de lanza en astillero, adarga antigua, rocín flaco y galgo corredor. Una olla de algo más vaca que carnero, salpicón las más noches, duelos y quebrantos los sábados, lantejas[1] los viernes, algún palomino de añadidura los domingos, consumían las tres partes de su hacienda.

Nº.	Oraciones	Palabras	Complejas
1	En un lugar de la Mancha, de cuyo nombre no quiero *acordarme*, no ha mucho tiempo que vivía un *hidalgo* de los de lanza en *astillero*, *adarga antigua*, rocín flaco y galgo *corredor*.	33	6
2	Una olla de algo más vaca que *carnero, salpicón* las más noches, duelos y *quebrantos* los *sábados*, *lantejas* los viernes, algún *palomino* de *añadidura* los *domingos*, *consumían* las tres partes de su *hacienda*.	33	10
	Totales	66	16
	Promedio	33	
	Índice de Niebla	**22,9**	

Tabla 2. Análisis del ejemplo 2

El Índice de Niebla = 0.4 ((66 / 2) + 100 (16/66)) = **22,9**

En este ejemplo también tenemos 2 oraciones, pero mucho más largas y con 16 palabras complejas. El resultado indica que se requieren 22,9 años. Es un valor alto. Deja claro, el Índice de Niebla, que este famoso libro es difícil de entender.

El índice disminuye mediante un ligero cambio en el texto. Observamos, con el respeto de Don Miguel de Cervantes y tan solo para mostrar el caso, que cambiando algunas comas por puntos el resultado baja casi diez años.

En un lugar de la Mancha, de cuyo nombre no quiero acordarme. No ha mucho tiempo que vivía un hidalgo de los de lanza en astillero. Adarga antigua, rocín flaco y galgo corredor. Una olla de algo más vaca que carnero. Salpicón las más noches. Duelos y quebrantos los sábados. Lantejas los viernes y algún palomino de añadidura los domingos, consumían las tres partes de su hacienda.

Nº.	Oraciones	Palabras	Complejas
1	En un lugar de la Mancha de cuyo nombre no quiero *acordarme*	12	1
2	No ha mucho tiempo que vivía un *hidalgo* de los de lanza en *astillero*	14	2
3	*Adarga antigua* rocín flaco y galgo *corredor*	7	3
4	Una olla de algo más vaca que *carnero*	8	1
5	*Salpicón* las más noches	4	1
6	Duelos y *quebrantos* los *sábados*	5	2
7	*Lantejas* los viernes y algún *palomino* de *añadidura* los *domingos consumían* las tres partes de su *hacienda*	17	6
	Totales	67	16
	Promedio	9,57	
	Índice de Niebla	**13,4**	

Tabla 3. Análisis del ejemplo 2 modificado

Pasamos, con el maquillaje de los puntos, de 2 oraciones a 7. El nuevo Índice de Niebla es

de 13,4 años contra al anterior de 22,9. Vemos, con este ejemplo, como la evaluación castiga las oraciones largas. Además, si cambiáramos las "palabras complejas" por sinónimos más cortos el Índice continuaría disminuyendo.

ALGORITMOS

Para calcular el Índice de Niebla leemos el texto desde un archivo plano, en nuestro caso es un archivo tipo txt. Lo separamos en oraciones y las almacenamos en una base de datos. En esta investigación empleamos MySQL (Butcher & Maslakowski, 2001). Al finalizar la lectura del texto se tiene en la tabla todas las oraciones, si se hallan repetidas se guarda solo una. Para cada oración se cuentan las palabras que forman la oración, las palabras complejas y las palabras desconocidas.

Para encontrar las palabras complejas es necesario dividir cada palabra en sílabas y luego contarlas. Si tiene tres o más, la palabra se considera "compleja". Entre más palabras complejas se hallen en un texto más alto será el valor del Índice. Usamos un algoritmo que recibe una palabra y la devuelve dividida en sílabas, por ejemplo, para almohada devuelve al-mo-ha-da. Con otra función, contamos las sílabas, para este caso, devuelve 4. El índice considera "compleja" la palabra almohada, pero no lo hace con galgo (gal-go), con 2 sílabas. Sería difícil encontrar una persona que desconozca la palabra almohada, sucedería lo contrario con galgo, como se observa el Gráfico 1. Para, almohada-galgo, vemos que la consideración de complejidad no es tan sutil como debería. Este detalle nos hizo considerar una modificación en la forma de calcular el Índice de Niebla y comparar los resultados.

Para efectuar la comparación, programamos el algoritmo para el cálculo del Índice de Niebla de Gunning original con la fórmula presentada en (1). Luego, usamos nuestra propuesta mostrada en (2), la cual consiste en buscar en el *corpus* las palabras. Cambiamos las "palabras complejas" de la fórmula original por las "palabras desconocidas". Así, cualquier palabra con un uso bajo en el *corpus* la consideraremos "desconocida" o poco familiar para un lector meta, obviando la cantidad de sílabas. En la fórmula original sustituimos las palabras complejas por las desconocidas así:

$$\text{Índice} = 0{,}4 \left(\left(\frac{palabras}{oraciones} \right) + 100 \left(\frac{palabras\ complejas}{palabras} \right) \right)$$

Con este cambio podemos comparar los resultados y establecer si existe una variación importante. Es interesante destacar que el nuevo cálculo dependerá del valor establecido para clasificar una palabra con uso bajo. Al estudiar el *corpus* obtenido vemos que este valor se puede establecer entre 10 000 y 20 000. Lo mismo sucede con el Índice actual que considera a una palabra compleja cuando tiene tres o más sílabas, el cálculo varía si fuera de cuatro o más.

CORPUS LINGÜÍSTICO DEL ESPAÑOL

La modificación introducida para el cálculo consiste en agregar un *corpus* lingüístico del español. Para el análisis, ahora, cada palabra de la oración la buscamos en una tabla de la base de datos que hemos denominado *CorpusEsp*, si no está registrada o tiene un valor bajo (por ejemplo 20 000) se le considera "desconocida". Esta variante castiga aquellas palabras que en el *corpus* están registrados con un uso poco frecuente. En nuestro *corpus* "almohada" tiene un uso alto (registra 116 358 apariciones), mientras que "galgo" no (18 779), por lo tanto será considera una "palabra desconocida" y almohada no. En el Gráfico 1 se muestra el uso de ambas palabras según el Google Books Ngram Viewer.

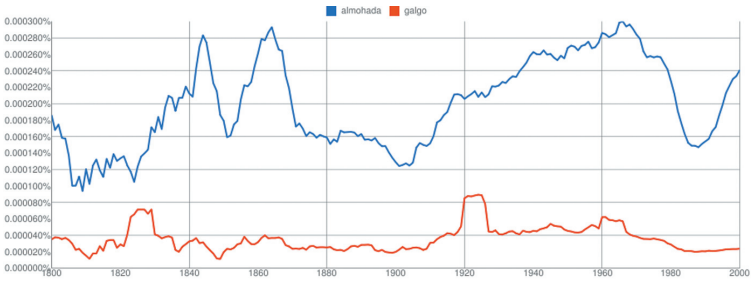

Grafico 1. Uso de las palabras almohada y galgo entre los años 1800 al 2000

Para el estudio construimos un *corpus* a partir de los libros en español digitalizados por Google (Michel y Lieberman, 2010), cuya información está disponible en el sitio web de Google Books Ngram Viewer <http://ngrams.googlelabs.com/datasets>. Más adelante detallaremos la construcción de este *corpus* que llamaremos corpus Google y que contiene más de un millón de palabras, lo cual permite conocer con bastante exactitud la frecuencia de uso de las palabras.

CONSTRUCCIÓN DEL *CORPUS*

Google y la Universidad de Harvard, por medio del proyecto Google Books Ngram Viewer, divulgado en Science (Lieberman & Erez, 2010). Colocaron 10 archivos que contienen las palabras de varios millones de libros en español digitalizados hasta la fecha. También lo hicieron para varias versiones del inglés, alemán, francés, chino simplificado, ruso y español. Por lo tanto, este mismo análisis se puede hacer para esos idiomas.

Descargamos los 10 archivos que contienen las palabras (una palabra por registro), que ellos llaman: 1-grams. También construyeron 2-grams, 3-grams, 4 grams y 5-grams, que no nos interesan para el *corpus* de esta investigación. A pesar de la inmensa cantidad de libros incluidos en la base de datos, tanto solo representan el 4% de los publicados en la historia. Luego, prometen, actualizarlos al avanzar la digitalización. El nombre de los archivos es: googlebooks-spa-all-1gram-20090715-X.csv.zip, *spa* identifica al idioma español, *1gram* indica que viene una palabra por registro, 20090715 es la fecha de corte: 15 de julio de 2009, y X varia de 0 a 9. Cada archivo contiene alrededor de 16 millones de registros. El formato es:

ngram TAB year TAB match_count TAB page_count TAB volume_count NEWLINE

El TAB corresponde a '\t' y el NEWLINE a '\n'. (Ver la tabla 4).

Con la siguiente sentencia SQL (lenguaje que permite ejecutar operaciones en una base de datos) cargamos el archivo googlebooks-spa-all-1gram-20090715-X.csv, luego de desempacarlo del zip, en una tabla llamada corpusGoogleX. Esta carga la repetimos para X de 0 a 9 hasta completar los 10 archivos. El ejemplo corresponde al primer archivo.

```
LOAD DATA LOCAL INFILE 'google-
books-spa-all-1gram-20090715-0.csv'
REPLACE INTO TABLE corpusGoogle0
FIELDS TERMINATED BY '\t'
LINES TERMINATED BY '\n';
```

Al finalizar el proceso tenemos 10 tablas con aproximadamente 16 millones de registros en cada una. Para cada palabra (ngram), los investigadores de Google y Harvard, registraron el año (year), la cantidad de veces que se usó la palabra(match_count), la cantidad de páginas donde se usó (page_count) y la cantidad de libros (volume_count). A continuación, se muestra el ejemplo para la palabra almohada de los cuatro primeros años.

Ngram	Year	match_ count	page_ count	volume_ count
almohada	1792	1	1	1
almohada	1795	1	1	1
almohada	1833	1	1	1
almohada	1841	2	2	2

Tabla 4. Muestra de los registros de la palabra almohada.

"Almohada" está registrada en 83 años diferentes a partir de su primera aparición en 1792 hasta el año 2009. Así, tenemos 83 registros tan solo en esta tabla. Para el *corpus* nos interesa un registro único para cada palabra. Al finalizar la carga de las 10 tablas, efectuamos otros procesos para agrupar las palabras. Con este agrupamiento los 10 archivos los pasamos a otras 10 tablas cuya cantidad pasa de los 16 millones por archivo a alrededor de 217 000 registros en cada una. Luego, juntamos las 10 tablas en una llamada *corpusEsp* (corpus Google) que contenía unos 2 170 000 registros, de los cuales muchos eran hileras con caracte-

res no alfabéticos, que no nos interesan para nuestro análisis. Efectuamos la limpieza para obtener un *corpus* final con 1 200 000 palabras. Aunque sigue teniendo inconsistencias, por ejemplo, palabras inexistentes en el español, decidimos trabajar con este *corpus*.

COMPARACIÓN DE LOS ÍNDICES

Programamos una serie de algoritmos para analizar el texto y calcular el Índice de Niebla original de Gunning (I.N.) y el modificado (I.N.+). Con estos resultados comparamos los índices y determinamos si existe una variación notable. Observamos en el análisis que al pasar la constante de "palabras desconocidas" de 10 000 a 20 000 el resultado, en promedio, sube un año. Aumentando el valor de esta constante podemos acercarlo tanto como se quiera al valor del I.N. original. Lo hicimos. Para una constante de 600 000 (la mitad de las palabras de la base de datos) los resultados se acercan bastante en la mayoría de los casos analizados. La constante se acomoda al gusto del analista al definir su lector meta.

En la siguiente tabla efectuamos una comparación los resultados de varios textos. Para el I.N. + utilizamos una constante de 10 000.

El Quijote tiene frases muy largas, en el capítulo I de la parte I el promedio es de 55 palabras y en el capítulo XLV de la segunda parte es 41,37. Las traducciones del inglés al español tienen frases cortas, lo observamos en el cuento de Agatha Christie *La esmeralda del Rajá* con 13,05. En general notamos que las frases fueron acortándose desde la edad media hasta nuestros días.

DIVISIÓN SILÁBICA

Para el cálculo de la fórmula original del Índice de Niebla de Gunning programamos varios algoritmos que nos permitieran dividir en sílabas las palabras obtenidas de los textos. Es un ejercicio interesante para un curso de Estructuras de Datos o de Lingüística Computacional, pues requiere el manejo detallado de varias funciones para el tratamiento de hileras y caracteres, además del conocimiento de las reglas para la división silábica, (Ríos, 1999) y (Cuayáhuitl, 2004).

ERRORES DE DIGITALIZACIÓN

La digitalización de libros genera errores, principalmente, cuando los libros son antiguos, con tipografías dañadas, hojas

Fuente	Total de palabras	Total de palabras complejas	Total de palabras desconocidas	Cantidad de oraciones	Promedio de palabras por oración	I.N.	I.N. + 10 K
Historia de la vida del Buscón Francisco de Quevedo (Completo)	42 865	2 998	2 985	1 807	23,72	18,3	12,3
El Quijote Capítulo I Parte I	1 893	441	110	34	55	31,6	24,6
El Quijote Capítulo XLV Parte II	2 606	538	144	63	41,37	24,8	18,8
La Biblioteca de Babel Jorge Luis Borges	2 545	884	246	114	22,3	22,8	11,8
La esmeralda del Rajá Agatha Christie	5 598	1 558	301	429	13,05	16,4	7,4

Tabla 5. Análisis de varios textos con ambos Índices

manchadas. A continuación, mostramos una lista con algunos de los errores encontrados en el *corpus* obtenido por medio de Google Books Ngram Viewer, en la columna de la izquierda tenemos la palabra registrada en el *corpus* y a la derecha la palabra correcta:

Palabra registrada	Palabra correcta
aaestro	maestro
aalud	salud
clegrement	alegremente
cnodadados	anonadados
cxcmplo	ejemplo
cxiste	existe
oeneral	general

Tabla 6. Muestra de algunos errores

PALABRAS MÁS USADAS SEGÚN EL *CORPUS*

En la siguiente tabla mostramos las 30 palabras más contabilizadas en el *corpus* obtenido, está compuesta, en su mayoría, por artículos, conjunciones, preposiciones y adverbios.

Nº.	ngram	match_count
1	de	2 147 483 647
2	la	1 597 505 555
3	que	1 089 514 318
4	en	1 089 030 314
5	el	1 070 446 726
6	y	1 054 154 430
7	los	692 014 843
8	a	654 217 920
9	del	509 192 465
10	se	485 497 642
11	las	462 860 464
12	por	401 218 947
13	con	322 459 267
14	no	306 204 223
15	un	292 551 072
16	su	276 709 946
17	una	265 048 294
18	para	253 343 949
19	al	235 152 222
20	es	223 729 342
21	lo	188 914 993
22	como	180 679 049
23	o	134 572 812
24	sus	126 930 239
25	pero	79 313 168
26	este	77 896 456
27	ha	77 151 482
28	sobre	77 039 305
29	esta	75 545 841
30	sin	72 276 844

Tabla7. Las 30 palabras más usadas

CONCLUSIONES

Aunque se debe lidiar con errores que contiene, el *corpus* obtenido, a partir de los libros digitalizados por Google, representa una fuente importante para el análisis efectuado. Es factible construir un *corpus* con un millón de palabras obteniendo libros ya digitalizados, por ejemplo, del proyecto Gutenberg, prácticamente sin errores. Cada interesado puede tener su propio *corpus* construido según sus características. Un periódico puede hacerlo partiendo de la digitalización de sus propios ejemplares, una universidad de sus publicaciones.

El problema de apegarse estrictamente a un *corpus*, cambiando las palabras desconocidas por sinónimos familiares, limitaría el idioma.

Los editores de texto podrían tener un *corpus* para revisar que las palabras escritas sean familiares para el lector meta definido. El editor las marcaría y el usuario toma la decisión si la cambia.

BIBLIOGRAFÍA

Álvarez, Carlos J. y Carreiras, Manuel. (1992). Estudio Estadístico de la Ortografía Castellana: (1) La frecuencia silábica. *Cognitiva. Vol.4, Nº. 1.* Pp. 75-105.

Butcher, Tony & Maslakowski, Mark. (2001). *Aprendiendo MySQL en 21 días*. Prentice Hall.

Cantos, P. (2001). *Spanish corpus*. Corpora List. October 1.

Cervantes, Miguel. (2000). *El ingenioso caballero don Quijote de la Mancha*.

Cuayáhuitl, H. (2004). *A Syllabification Algorithm for Spanish*. A. Gelbukh (Ed.): CICLing 2004, LNCS 2945, pp. 412-415.

Gunning, Robert. (1973). *The Technique of Clear Writing*. Rev. ed. New York: McGraw-Hill Book Company.

King, Stephen. (2000). *On Writing: A Memoir of the Craft*.

Kucera, H. y Francis, N. (1967). *Computational Analysis of Present Day American English*. Brown University Press.

Michel, Jean-Baptiste, Lieberman Aiden, Erez. (2010). Quantitative Analysis of Culture Using Millions

of Digitized Books. *Science DOI: 10.1126/ science.1199644*. Published Online 16 December 2010.

Ovares, Felipe et al. (2010). *Antología de Haikus costarricenses*. Nueva Acrópolis y Embajada de Japón.

Ríos Mestre, Antonio. (1999). *La transcripción fonética automática del diccionario electrónico de formas simples flexivas del español: Estudio fonológico en el léxico*. Volumen 4. Barcelona.

Sánchez, A., Sarmiento, R., Cantos, P. y Simón, J. (1995). *Cumbre. corpus lingüístico del español contemporáneo. Fundamentos, metodología y aplicaciones*. Madrid: SGEL.

Sebastián, N., Cuetos, F., Martí, M.A., y Carreiras, M.F. (2000). *LEXESP: Léxico informatizado del español*. Edición en CD-ROM. Barcelona: Edicions de la Universitat de Barcelona (Colleccions Vàries, 14).