Un método para la eliminación de la redundancia en la transmi-- sión de textos escritos en castellano

Por Félix Ares de Blas

A lo largo del presente trabajo se han conseguido algunas modestas aportaciones en el campo de la eliminación de la redundancia en transmisión de textos escritos en castellano. Las podríamos resumir así:

1º.- Hemos calculado unos límites superior e inferior a la cantidad de información que proporciona una letra del caste-- llano impreso, siguiendo el método que utilizó Shannon para el idioma inglés. En principio dicho conocimiento permite a los investigadores de este tema saber dónde está su - límite teórico.

2º.- Hemos constatado que refiriéndonos exclusivamente a la cantidad de información de una letra o de un conjunto de le-- tras, el inglés y el castellano tienen valores muy similares, por lo que los datos existentes para el idioma inglés pueden extrapolarse, con las debidas precauciones al cas-- tellano.

3º.- Hemos constatado, aunque a un nivel todavía sujeto a discusiones, que cada letra de un texto traducido da menos in-- formación que la del texto original.

4º.- Hemos demostrado que hay unas limitaciones de orden teórico a la codificación de la extensión de orden n de una -- fuente. n no puede crecer indefinidamente. Hemos probado - que, dependiendo de la longitud del texto a codificar y de los diccionarios utilizados en su codificación, hay un n - óptimo. Con ello hemos clarificado la gran influencia que tiene la inestabilidad de las frecuencias relativas de aparición de n - gramas.

5º.- Hemos constatado que la transmisión de textos rompiéndolos en palabras sólo es adecuado para fuentes y diccionarios - muy grandes, que para fuentes o diccionarios de pequeño tamaño, codificar a nivel de palabra puede ser peor que codificar a nivel de 3 - gramas.

6º.- Hemos visto que la sílaba es un buen bloque para conseguir un ahorro de binits en transmisión de textos escritos en castellano. Se trata de un bloque grafémico multiletra frecuente y estable. Como consecuencia del estudio silábico hemos desarrollado unos programas que rompen las palabras en sílabas.

7º.- Hemos visto que la información que proporciona una letra depende de su posición dentro de la palabra, y hemos desarrollado algunos métodos para obtener una ventaja de este hecho.

8º.- Hemos demostrado que la información que proporciona una sílaba depende de su posición dentro de la palabra, y hemos desarrollado algunos métodos para obtener un cierto ahorro de binits a partir de este hecho. Concretamente el método de distinguir entre tres tipos de sílabas posicionales, suponiendo que el blanco de separación entre palabras forma parte de la sílaba final, ha demostrado ser el mejor para casi todos los casos estudiados, llegándose a los 2'91 --- binits/letra. Solamente ha sido superado en el caso de textos fuente y diccionarios muy grandes por la codificación en palabras.

9º.- Como conclusión podemos decir que hemos desarrollado un método de eliminación de la redundancia para textos escritos en castellano totalmente original, muy adaptado a dicho -- idioma, y con el que se consiguen resultados similares, ligeramente mejores que los reportados para el inglés. (Codificado y transmitido: 2'91 binits por letra. Teniendo en cuenta sólo la redundancia, sin llegar a codificar: 2'61 - bits de información por letra).

10º.- Marginalmente, como subproductos de la investigación, hemos obtenido ciertos datos estadísticos sobre textos escritos en castellano. Por ejemplo, frecuencias relativas --- (f.r.) de letras, f.r. de letras según su posición en la palabra tanto empezando a numerar por la izquierda como -- por la derecha, palabras diferentes que constituyen el 25% y el 50 % de un texto, f.r. de palabras, f.r. de sílabas, f.r. de las sílabas de acuerdo con su posición dentro de la palabra, f.r. de las letras dentro de las sílabas te---

niendo en cuenta su posición en ellas, f.r. de las letras dentro de las sílabas teniendo en cuenta su posición en -- ellas y considerando tres tipos diferentes de sílabas se-- gún su posición dentro de la palabra, sílabas que forman - 25 % y el 50 % de un texto separando por los diferentes ti pos posicionales de sílabas, sílabas más frecuentes tenien do en cuenta sus diversos tipos, cálculo de las entropías de todos los casos en que había f.r., longitudes medias de palabras, de sílabas, de sílabas posicionales, f.r. de apa rición de letras dentro de n-gramas (Para n = 3,4,5 y 6), f.r. de n-gramas, costos relativos de transmisión en binits (dígitos binarios) y en bits (unidades de información) de n-gramas, palabras, sílabas con el blanco considerado síla ba aparte, sílabas con el blanco integrado en la última sí laba, sílabas posicionales con el blanco como última síla- ba, sílabas posicionales con el blanco integrado en la úl- tima sílaba, costo relativo de inestabilidad de los mismos casos que los ya señalados para costos relativos de trans- misión. Además, frecuencias relativas de palabras de n le- tras, frecuencias relativas de sílabas de n letras, f.r. de palabras de n sílabas. Codificación Huffman de diversos alfabetos castellanos (con 27 letras, con 28, etc.), codi- ficación de Huffman de las longitudes de las palabras me-- didas en letras y en sílabas, etc, etc.

TEXTOS BASE USADOS

Para realizar la investigación hemos partido de la -- elección aleatoria de una serie de autores, para los que hemos perforado parte de su obra. La perforación se ha hecho en código Fieldata, sólo se han tenido en cuenta las mayúsculas, se -- tienen en cuenta varios signos de puntuación que se intercalan entre las palabras: punto, punto y coma, punto y aparte, dos puntos, abrir admiración, cerrar admiración, abrir y cerrar admiración, paréntesis, etc., pero no se han tenido en cuenta los signos de puntuación que van sobre las letras: acentos, diéresis, etc.

Las obras han sido (en el orden de elección aleato -- rio):

José Mª Gironella.- "Un millón de muertos". Editorial Planeta. Colección Omnibus. Barcelona 1966. 139 Edición. (Cap. 1).

Ramón Menéndez Pidal.- "Los españoles en la historia". Ed. Espasa Calpe (Argentina) S.A. Colección Austral. (Vol. Nº 1.260). Buenos Aires 1959. (1ª Edición). (Cap. 1)

Pío Baroja.- "La Trapera" y la "La sima" Ed. Alianza Editorial. Colección el libro de bolsillo. Madrid 1967. 2ª edición (Primeras 12.628 letras que comprenden íntegramente "La trapera" y parte de "La sima".

Miguel de Unamuno.- "Amor y Pedagogía". Ed. Emesa. Colección Noveles y Cuentos. Madrid 1967. 2ª edición. (Primeras 8.426 letras del capítulo 2).

Alvaro de Laiglesia.- "Sólo se mueren los tontos". Ed. Planeta. Colección novela. Barcelona 1967. (10ª Edición). (Pedazo IV).

Benito Pérez Galdós.- "Marianela". Ed. Hernando. S.A. Colección Novelas españolas contemporáneas. 1ª época Madrid 1943. (Capítulo 1)=

Camilo José Cela.- "La familia de Pascual Duarte".
Ed. Destino. Colección: Destinolibro. Vol 4.
Barcelona 1977. (6ª edición). (Capítulo 3)

Además añadimos un texto periodístico al azar. La --- elección recayó sobre el País. Editorial "Voto de conciencia" - de Marzo de 1981 y el artículo "Prudencia y jurisprudencia" del 15/7/81.

Para tener una referencia clásica utilizamos "Don Quijote de La Mancha" de Miguel de Cervantes. Editorial Bruguera. Libro clásico. Barcelona 1981. La obra consta de dos volúmenes se perforó el primero de ellos completo, incluyendo el prólogo de Juán Alcina Franch.