

Así es la frecuencia de la posición de las letras en inglés (INFOGRAFÍA)



¿Sabías que muchísimas palabras en inglés terminan por la letra E pero menos de una quinta parte empiezan por ella? ¿O que la letra B termina una milésima fracción de las palabras que empiezan por dicha letra? Como en todos los idiomas, el inglés tiene palabras

y letras más comunes que otras palabras y letras respectivamente, pero llama la atención cómo las letras tienen una distribución sólida dentro de las palabras. Letras que aparecen más al principio, al final, o en medio de las palabras.

Esto lo demostraron en ProffReader en 2014, pero lo rescato hoy día porque me topé con ello la semana pasada y me pareció fantástico que alguien pierda invierta el tiempo en analizar este tipo de datos. Por ejemplo, tomemos las letras D y Q del gráfico de abajo, en el que el color (de blanco a carmesí) indica la frecuencia de la letra en el idioma. Es decir, que la D es más frecuente que la Q. Esto último se ve mejor en la infografía de abajo.



El gráfico se lee de la siguiente manera: muy pocas palabras empiezan con la letra D (dad, damage, dance); menos aún la tienen en el centro de la palabra (adz, ado, ads); pero a medida que nos desplazamos al final de las palabras, la letra

D se vuelve muy frecuente (bad, had, mud). Algo parecido aunque inverso ocurre con la letra Q, muy frecuente al inicio de las palabras (queen, quite, quality) pero casi imposible de localizar al final de la palabra. ¿Por qué?

Esto recuerda a cómo grandes marcas como Google, Facebook, Apple, Nike, Amazon, etc conservan su sílaba tónica a comienzo de la palabra y luego van perdiendo fuerza. La economía del lenguaje favorece ciertos sonidos en determinadas posiciones de las palabras, incitando al uso, mientras que relega otros al olvido. En La Piedra de Sísifo somos adictos a este tipo de estudios.

Por ejemplo, en castellano hay muy pocas palabras que empiecen por Z. Concretamente 49, mientras que existen al menos 200 que terminan con esa letra. La RAE solo permite consultar hasta 200 términos, así que deducimos que hay alguna más. La relación es, por tanto, 1:4, 1:5 o más.

El estudio realizado en inglés por Prooff Reader contaba con una curiosa metodología de análisis:

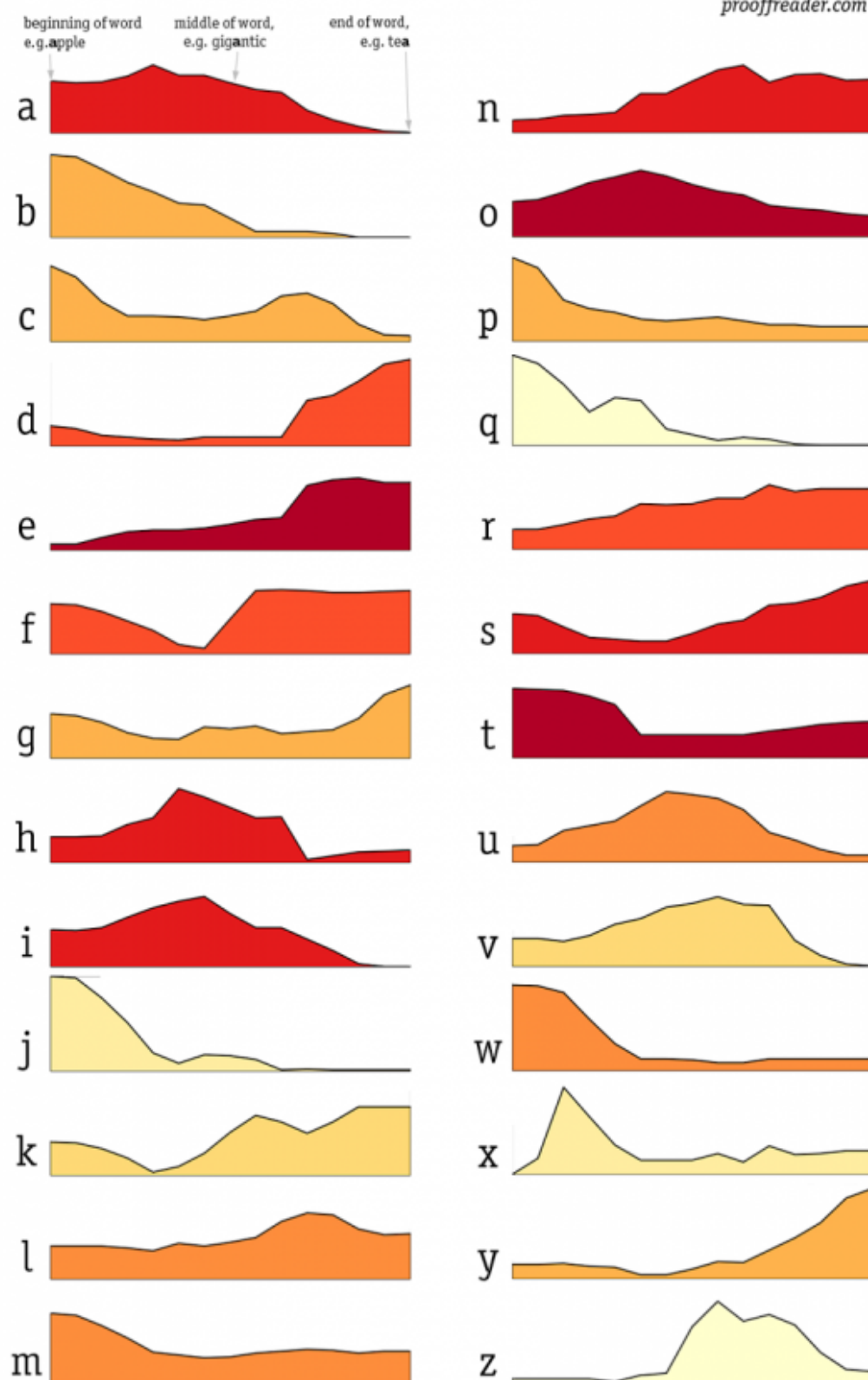
1. Los datos para el análisis se sacaron del Natural Language Toolkit, que viene a ser una suerte de biblia del léxico americano para hacer estudios. Como ellos mismos dicen, es limitado y ni siquiera está actualizado, pero sirve para hacer una aproximación estadística interesante para frikis estudiosos del lenguaje.
2. Han ponderado las palabras en base a su uso. Por ejemplo, el artículo *the* pesa más que la conjunción o preposición *than* porque estas son menos frecuentes. Eso significa que no todas las palabras valen lo mismo para el estudio.
3. Hacen uso de colores para matizar la frecuencia con la que una letra determinada aparece en el idioma, y ponen el ejemplo entre la E y la Z. La primera se usa entre 100 y 200 veces más que la segunda, por lo que un

gráfico a escala entre ambas haría que la Z se viese plana (cuando en realidad tiene una variabilidad de frecuencia a lo largo de la longitud de las palabras que ha de ser representada).

4. Como no todas las palabras tienen la misma longitud, en realidad no se puede comparar tal cual una letra que aparezca mucho en la segunda posición de una palabra de cinco letras con esa misma letra apareciendo en la segunda posición de una palabra que tenga siete. Aquí entra un poco con calzador, y el estudio es más cualitativo que cuantitativo, aunque se ha usado la estadística para agrupar frecuencias.

Distribution of English letters toward beginning, middle and end of words

proofreader.com



frequency of letter:

0% 0.1% 0.5% 1% 2% 3% 5% 9% 13%

Weighted for word frequency,
e.g. the t in "the" contributes
more than the t in "thee".

Source: Brown corpus
Methodology: www.prooffreader.com/explain.html

Fuente: lapiedradesisisifo.