Una aplicación para explorar la frecuencia léxica a partir de corpus de referencia

Mario Casado-Mancebo
Universidad Nacional de Educación a Distancia

Lexical frequency or how often a word is used as compared to others in a language is a fundamental factor in reading and processing tasks. This has been proved by experimental research both in adults and children showing the close relationship between reading comprehension, lexical skill, and lexical decoding. In this work we present an online application for exploring lexical frequency in Spanish texts compared to reference corpora. The current version includes all three corpora from Real Academia Española (CORDE, CREA and CORPES XXI), which allows both diachronic and synchronic research. This application let users send texts (or words) to be processed and, as a result, it retrieves a table with each word's reference frequencies in a selected corpus. Frequency information includes order in the list of frequencies, and absolute and normalized frequency. Results are easily downloadable for external processing.

Keywords: Corpus linguistics, Reading tasks, Lexical frequency, Online application

1. Introducción

Un factor de peso en el diseño de corpus para la investigación en ámbitos de la lingüística como la fonología, la lingüística clínica o la enseñanza es la frecuencia léxica o la tasa de uso de una palabra en la lengua. Esto se remonta la Ley de Zipf, que define esta frecuencia como una función lineal en escala: una palabra se vuelve frecuente por su uso frente a otro conjunto de uso menor o mínimo (Bargetto Fernández & Riffo Ocares 2019). Como señala Graña López (1997), este parámetro es un condicionante fundamental en las tareas de procesamiento y producción lingüísticos. Concretamente, en la producción del habla, Stemberger & MacWhinney (1988) muestran la relación inversa que existe entre la frecuencia léxica y el nivel de error y la relacionan con la posibilidad de que estén presentes en el lexicón mental, de modo que el interlocutor únicamente debe recuperarlos y no reconstruirlos (o derivarlos) a partir de reglas lingüísticas (Graña López 1997). En esta línea, Bybee (1988) explica cómo cuanto más se procesa una palabra más fuerza léxica acumula, lo que explica ciertos patrones de organización morfológica (Graña López 1997).

CHIMERA. Romance Corpora and Linguistic Studies 10 (2023), 89-95. ISSN 2386-2629

En las tareas de percepción, los estudios también han mostrado el peso de la frecuencia. El modelo Logogén introduce la noción de activación, relacionada con el momento en el que la información procesada por el sujeto es suficiente como para identificar elementos específicos del léxico (Bargetto Fernández & Riffo Ocares 2019). Es importante recordar que la activación no implica un solo elemento sino toda una serie de ellos (en el modelo mencionado, aquellos que comparten logogenes). Esta forma de procesamiento trae consigo una serie de efectos. Destaca, por un lado, el efecto de lexicalidad, por el que las pseudopalabras o palabras inventadas se reconocen más lento; es decir, casa se identificaría más rápidamente que casu (Baquero Castellanos 2005). Por otro, el de frecuencia léxica: reconocemos más rápido las palabras con las que tenemos experiencia previa (Balota & Chumbley 1984). Esto se ha comprobado empíricamente mediante, entre otras, tareas de decisión léxica, nombrado y rastreo de movimientos oculares (Bargetto Fernández & Riffo Ocares 2019). Otros efectos de sobra conocidos son el de longitud o el de priming (Baquero Castellanos 2005; Bargetto Fernández & Riffo Ocares 2019).

Todo lo mencionado hace patente la importancia de la frecuencia léxica en las tareas de lectura y, por consiguiente, en la elaboración de corpus destinados a tareas de esta índole. Concretamente, Perfetti & Adlof (2012) señalan el valor fundamental de la identificación de palabras en la lectura para la comprensión: la debilidad en el control explícito de palabras conlleva a errores de asignación de significados. Y Perfetti (2010) establece la triple relación entre habilidad léxica, de comprensión y de descodificación. Cabe destacar que la investigación experimental ha mostrado que estos efectos estudiados principalmente en adultos también se dan cuando las tareas de lectura son efectuadas por niños: Rodrigo López (1994) encuentra un efecto principal de la lexicalidad y de la frecuencia léxica.

2. Una aplicación para la exploración de la frecuencia léxica a partir de corpus de referencia

La aplicación, disponible en https://frecuencias.mcasado.org, consiste en una interfaz web que permite insertar palabras, secuencias de palabras o textos para que sus formas sean cotejadas con corpus de referencia y así recuperar diferentes informaciones relacionadas con la frecuencia.

2.1 Los corpus de referencia

La aplicación incorpora los tres corpus del banco de datos de la Real Academia Española. Dos de carácter sincrónico: el Corpus de Referencia del Español Actual (CREA) y el Corpus del Español del Siglo XXI (CORPES XXI); y uno de carácter diacrónico: el Corpus Diacrónico del Español (CORDE). Se eligieron estos tres por estar sus listados de frecuencia en acceso público y documentados. En el caso de otros corpus de referencia del español, como *News on the Web* (Davies 2016), la información de frecuencia no es accesible de forma pública, por lo que no se puede automatizar la consulta.

Estos tres corpus, sin embargo, ofrecen una buena panorámica del español a través del tiempo. El CORDE abarca "todas las épocas y lugares en que se habló español, desde los inicios del idioma hasta el año 1974" (Real Academia Española s. f.-a) y cuenta con 250 millones de registros de textos de diferentes géneros (narrativos, líricos, dramáticos, científico-técnicos, históricos, jurídicos, religiosos, periodísticos, etc.) en prosa y verso. Para ser suficientemente representativo, en él "se pretende recoger todas las variedades geográficas, históricas y genéricas" (Real Academia Española s. f.-a). Por su parte, el CREA en su última versión (junio de 2008) cuenta con algo más de ciento sesenta millones de formas. La Real Academia Española destaca que "se compone de una amplia variedad de textos escritos y orales, producidos en todos los países de habla hispana desde 1975 hasta 2004." (Real Academia Española s. f.-c). Supone, por lo tanto, una continuación a la información que aporta el CORDE con textos escritos procedentes tanto de libros como de periódicos y revistas que abarcan más de cien materias distintas. En el CREA encuentra también representación la lengua hablada, representada por "transcripciones de documentos sonoros, obtenidos, en su mayor parte, de la radio y la televisión." (Real Academia Española s. f.-c).

Por último, el CORPES XXI recoge textos orales y escritos procedentes de España, América, Filipinas y Guinea Ecuatorial desde 2001 hasta la actualidad. La versión actual (julio de 2021) "cuenta con más de 327 mil documentos que suman ya unos 350 millones de formas ortográficas" (Real Academia Española s. f.-b). También incluye en su base de datos multitud de géneros: "Por lo que respecta al bloque de ficción (novelas, guiones de cine, relatos, obras de teatro) las formas de CORPES sobrepasan los 95 millones, mientras que las contenidas en textos de libros de no ficción y en publicaciones periódicas (ciencias sociales, salud, política, artes, tecnología...) se acercan a los 250 millones. Los textos procedentes de libros suponen casi 172 millones de formas; las publicaciones periódicas están representadas con unos 167 millones. Seis millones y medio

más provienen de blogs, entrevistas digitales y miscelánea." (Real Academia Española s. f.-b). Estos tres corpus, en conclusión, suponen una base de datos suficiente para poder cotejar textos de investigación tanto de índole diacrónica como sincrónica y obtener una referencia sobre la frecuencia de sus formas en la lengua española.

2.2 Objetivo de la aplicación

Si bien la información que incorpora la aplicación es de acceso público, hasta este momento, para realizar la tarea de exploración de frecuencias de las formas de un texto o de una secuencia de palabras, los investigadores debían hacerlo manualmente cotejando inmensos listados de frecuencias diferentes o contar con sólidos conocimientos de programación para automatizar la tarea. Esta aplicación pretende, por lo tanto, dotar a la comunidad investigadora de una herramienta que permita, de una manera accesible y cómoda, explorar la información de frecuencia léxica de diferentes formatos: palabras sueltas, listas de palabras, listas de oraciones, textos, etc.

2.3 Objetivo de la aplicación

En pro de la accesibilidad, la aplicación se presenta en una interfaz web muy sencilla cuya página principal permite seleccionar el corpus de referencia en el que realizar el procesamiento (Ilustración 1). Una vez seleccionado, se accede a la página de exploración, que es la misma para todos los corpus.

Ilustración 1. Página principal de la aplicación

La página de exploración (Ilustración 2) consta de tres partes. En primer lugar, unas breves orientaciones para utilizar la herramienta de búsqueda. En segundo lugar, un formulario que permite insertar textos (o secuencias palabras) de hasta

500 formas. Como se indica en las instrucciones, los textos deben estar libres de signos de puntuación, ortográficos, elementos diacríticos, etc. Solo se debe incluir la secuencia de formas que se desea explorar. Una vez se hace el envío del formulario, la aplicación activa una serie de operaciones de procesamiento de texto para cotejar los elementos léxicos con los corpus y recuperar las frecuencias correspondientes a sus formas.

En el procesamiento del texto enviado, una vez confirmado el límite de 500 palabras para evitar posibles sobrecargas, se extraen las formas del texto y se seleccionan los elementos únicos; es decir, se hace un cribado de elementos repetidos que podrían aumentar el tiempo de procesamiento de manera innecesaria. Con todo ello, se cotejan las formas con las del corpus de referencia para recuperar la información de frecuencia de cada una ellas de manera estructurada y presentarla en la sección de resultados.

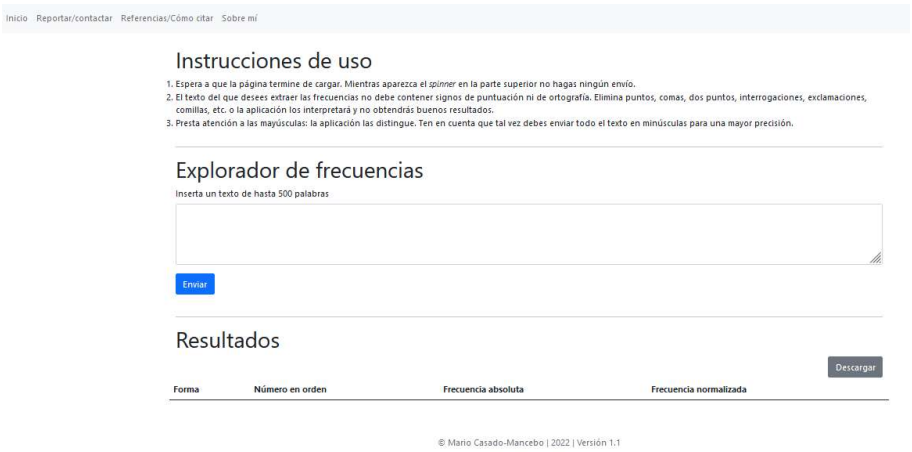

Ilustración 2. Página de exploración de la aplicación

Finalmente, en la parte inferior está la sección de resultados. Cuando se completa el procesamiento, se carga una tabla con diferentes informaciones de frecuencia (**Error! Reference source not found.**). En la versión 1.1 la tabla de resultados incluye cuatro columnas: forma, número en orden, frecuencia absoluta y frecuencia normalizada. En la columna forma, se muestra el elemento literal que se ha procesado. Como algunos corpus incluyen la distinción de mayúsculas y minúsculas, en esta columna se puede comprobar si el elemento enviado en el formulario de búsqueda incluye alguna mayúscula que haya podido resultar significativa en los datos de frecuencia. La columna número en

orden devuelve el índice de posición de la forma dentro de la lista de formas del corpus correspondiente ordenada según frecuencia. Con esta información se puede concluir que una forma está dentro de las N más frecuentes según el corpus que se haya consultado. La columna de frecuencia absoluta muestra el número de veces que aparece la forma procesada en el corpus y la de frecuencia normalizada, el número de casos por cada millón de elementos. Todos estos resultados se pueden descargar fácilmente en formato CSV mediante el botón Descargar para procesarlos externamente.

Ilustración 3. Un ejemplo de tabla de resultados devuelta por la aplicación tras una búsqueda

Esta aplicación está diseñada en Javascript, lo que permite hacerla depender únicamente del navegador y no de un servidor. Esto ahorra costes de mantenimiento y de infraestructura, pudiendo alojarse en cualquier servicio de despliegue web estático gratuito. Por otro lado, la desventaja de no contar con un servidor detrás de la aplicación es que el tiempo de carga de los corpus de referencia es mayor, puesto que el navegador debe almacenar todo su contenido, y se debe hacer cada vez que el usuario carga la página.

Referencias

Balota, D. A., & Chumbley, J. I. 1984. Are lexical decisions a good measure of lexical access? The role of word frequency in the neglected decision stage. *Journal of Experimental Psychology. Human Perception and Performance*, 10(3), 340–357.

Baquero Castellanos, S. 2005. Procesamiento léxico del castellano por parte de niños y adultos. *Forma y Función*, 18, 45–73.

Bargetto Fernández, M. Á., & Riffo Ocares, B. 2019. El reconocimiento de palabras y el acceso léxico: Revisión de modelos y pruebas experimentales. *Boletín de filología*, 54(1), 341–361. https://doi.org/10.4067/S0718-93032019000100341

Bybee, J. 1988. *Morphology as lexical organization*. En M. Hammond & M. Noonan (Eds.), *Theoretical morpho*logy. Academic Press. 119–141.

Davies, M. 2016. *News on the Web* [En línea]. https://www.corpusdelespanol.org/now/

Graña López, B. 1997. Frecuencia y procesamiento léxico. *Revista española de lingüística aplicada*, 12, 27–42.

Perfetti, C. (2010). *Decoding, vocabulary, and comprehension*. En M. McKeown & L. Kuncan (Eds.), *Bringing reading research to life.* The Guiford Press. 291–303.

Perfetti, C., & Adlof, S. (2012). One Reading Comprehension: A Conceptual Framework from Word Meaning to Text Meaning. *Psychology*.

Real Academia Española. s. f.-a. *Banco de datos: CORDE* [Real Academia Española]. Recuperado 20 de abril de 2022, de https://www.rae.es/banco-de-datos/corde

Real Academia Española. s. f.-b. *Banco de datos: CORPES XXI* [Real Academia Española]. Recuperado 20 de abril de 2022, de https://www.rae.es/banco-de-datos/corpes-xxi

Real Academia Española. s. f.-c. *Banco de datos: CREA* [Real Academia Española]. Recuperado 20 de abril de 2022, de https://www.rae.es/banco-de-datos/crea

Rodrigo López, M. 1994. *Acceso al léxico en buenos y malos lectores con diferente cociente intelectual (C.I.) En un sistema ortográfico transparente*. Universidad de La Laguna.

Stemberger, J. P., & MacWhinney, B. 1988. *Are Inflected Forms Stored In The Lexicon?* En M. Hammond & M. Noonan (Eds.), *Theoretical Morphology*. Academic Press. 101–116.