ISSN: 2591-541X | DOI: 10.5281/zenodo.14051547 | *Quintú Quimün. Revista de lingüística.* Nº 8 (2) 2024: Q105
Dossier *Lenguaje y cognición: la interdisciplina en los estudios sobre procesamiento de lenguaje*

Diccionario de frecuencia léxica infantil para el español rioplatense: el *Cuenta Palabras*

Children's lexical Frequency Dictionary for *Rioplantense* Spanish*:* the *Cuenta Palabras*

Julieta Fumagalli[1(a)(b)] **María Elina Sánchez**[2(a)(b)] **Bruno Bianchi**[3(b)(c)]
Matías Cancino[4(a)] **César Cocaro**[5(c)] **Martín Melman**[6(a)] **Laura Zabala**[7(a)]

[(a)] Universidad de Buenos Aires, Facultad de Filosofía y Letras, Instituto de Lingüística
[(b)] CONICET
[(c)] Universidad de Buenos Aires, Facultad de Ciencias Exactas y Naturales, Departamento de Computación, Laboratorio de Inteligencia Artificial Aplicada

Email
[1]julietafumagalli@filo.uba.ar
[2]mesanchez@filo.uba.ar
[3]brunobian@gmail.com
[4]cancinomatias@gmail.com
[5]cesar.cocaro@gmail.com
[6]melmanmartin@gmail.com
[7]laura.zabala@gmail.com

ORCID
[1]https://orcid.org/0000-0002-9532-5777
[2]https://orcid.org/0000-0001-6159-8366
[3]https://orcid.org/0000-0001-5252-4840
[4]https://orcid.org/0009-0001-8071-6257
[5]https://orcid.org/0009-0002-0573-0259
[6]https://orcid.org/0009-0001-4623-8845
[7]https://orcid.org/0009-0003-6061-3175

RESUMEN. La frecuencia léxica es una de las variables psicolingüísticas más importantes en el estudio de la comprensión y la producción del lenguaje oral y escrito. Esta variable explica por qué las palabras que nos encontramos más frecuentemente en la vida diaria se procesan de manera más rápida y efectiva que aquellas que rara vez leemos o escuchamos. Cualquier investigación que se realice en el marco de la psicolingüística, neurolingüística, psicología cognitiva y neurociencias, ya sea con niños o adultos, en la que se evalúe la comprensión o la producción de ítems léxicos debería considerar la frecuencia léxica para la selección de estímulos con los que se trabaje. El objetivo de este trabajo es presentar la elaboración del *Cuenta palabras*, un diccionario de frecuencia léxica infantil para el español en su variedad rioplatense de Argentina. Para tal fin, se construyó un corpus a partir de textos escolares de distintas áreas disciplinares acorde a los niveles de escolaridad (1° a 7° año de educación primaria).
Palabras clave: Psicolingüística, diccionario, instrumento, frecuencia léxica, español rioplatense

ABSTRACT. Lexical frequency is one of the most important psycholinguistic variables in the study of the oral and written language´s comprehension and production. This variable explains why words that we encounter more frequently in daily life are processed more quickly and effectively than those that we rarely read or hear. Any research carried out within the framework of psycholinguistics, neurolinguistics, cognitive psychology and neurosciences, whether with children or adults, in which the comprehension or production of lexical items is evaluated, should consider lexical frequency for the selection of stimuli to work. The objective of this work is to present the development of the *Cuenta Palabras*, a children's lexical frequency dictionary for Argentinian Spanish in *rioplatense* variety. For this purpose, a corpus was built from school texts from different disciplinary areas according to the levels of schooling (1st to 7th of primary education).
Keywords: Psycholinguistics, dictionary, instrument, lexical frequency, *Rioplatense* Spanish

Recibido: 3 mayo 2024 | Aceptado: 8 julio 2024

1 | INTRODUCCIÓN

¿Por qué algunas palabras se leen más rápido? ¿Por qué en una tarea de decisión léxica determinados estímulos son reconocidos como palabras con mayor velocidad que otros? Incluso, cabe preguntarse por qué accedemos y producimos en voz alta de manera más veloz algunas palabras cuando nos enfrentamos a un dibujo que las representa.

Cuando una persona habla, escucha o lee necesita acceder al léxico mental. Este es un sistema de memorias de largo plazo en el cual están almacenadas las representaciones fonológicas, morfológicas, sintácticas y semánticas, así como también ortográficas (una vez que se aprendió a leer y a escribir) de todas las palabras conocidas (Foster 1976; Libben & Jarema 2002). El proceso de activación y recuperación de las distintas formas se denomina *acceso léxico* y durante las últimas décadas los estudios experimentales que investigaron el reconocimiento y la producción de palabras en cualquiera de sus modalidades, oral o escrita, ha tratado de identificar los procesos involucrados y determinar las variables que se ven implicadas en ellos.

Las investigaciones sobre procesamiento del lenguaje en el marco de la psicolingüística y la psicología cognitiva han utilizado diversas tareas para indagar acerca de la organización del léxico mental. Entre las más utilizadas podemos mencionar aquellas que se apoyan en métodos cronométricos que, con la ayuda de un software específico, miden el tiempo que demora un sujeto en dar una respuesta y/o la cantidad de aciertos. Ejemplos de estas tareas son: la decisión léxica donde los sujetos tienen que determinar si una secuencia de letras o fonemas se corresponde o no con una palabra; la denominación oral o escrita en la cual se debe dar el nombre que designa a una imagen presentada; o la lectura en voz alta de palabras y pseudopalabras, entre otras (Hendrix & Sun 2021; Perea & Rosa 1999; Vergara, Gómez & Perea 2020). La evidencia empírica obtenida a partir de la aplicación de estas técnicas ha permitido detectar diversas variables que explican ciertos patrones de comportamiento; es decir, por qué algunas palabras requieren más tiempo y/ o presentan más errores durante el procesamiento del lenguaje. Así, se identifican variables lingüísticas y psicolingüísticas. Entre las primeras, se pueden mencionar dos variables muy significativas. En primer lugar, la longitud de la palabra. Esta hace referencia a la cantidad de letras o sílabas que la componen, lo cual tiene incidencia en el procesamiento léxico, es decir, las palabras de mayor longitud presentan mayores latencias y errores que las palabras cortas (Ashby & Rayner 2004; Carreiras & Grainger 2004; Taft & Forster 1975; Bijeljac-Babic, Millogo, Farioli & Grainger 2004). En segundo lugar, la complejidad silábica u ortosilábica también cumple un rol, ya que se observa una distinción en el procesamiento de aquellas palabras que presentan sílabas simples o complejas, siendo estas últimas más difíciles (Fenk-Oczlon & Pilz 2021; Fumagalli *et al*. 2014; Riecker et al. 2008). Entre las variables psicolingüísticas podemos mencionar, por un lado, aquellas que se relacionan con la experiencia del sujeto. Por ejemplo, la familiaridad, entendida como el grado de contacto diario que una persona tiene con un concepto determinado, apunta a que aquellos ítems que son más familiares se procesarán de manera más rápida y eficaz (Colombo, Pasini & Balota 2006; Leroy & Kauchak 2014) que los que no. Otro ejemplo de variable psicolingüística que toma en cuenta al sujeto es la edad de adquisición, que atiende al momento en que una persona adquirió determinado ítem léxico y que se obtiene mediante encuestas a poblaciones adultas o analizando producciones infantiles. Esta variable explica que cuanto más temprano se adquiere una palabra esto tiene consecuencia en la velocidad para procesarla lingüísticamente (Segui & Grainger 1990; Morrison *et al*. 2002). Por el otro, están aquellas variables que caracterizan a los ítems particularmente y que son consideradas en los estudios experimentales ya que pueden facilitar o no el reconocimiento o producción de una palabra. Entre ellas, se destaca la vecindad ortográfica y fonológica, que se describe según la cantidad de palabras que comparten rasgos ortográficos y/o fonológicos con otras palabras (Huntsman & Lima 2002; Adelman *et al*. 2013; Perea 2015; Ballot, Mathey & Robert 2021); la concretud e imaginabilidad, variables que explican diferencias

en el reconocimiento según los ítems sean abstractos (más difíciles) o concretos (Khanna & Cortese 2021); y, por último, la frecuencia léxica, que se calcula en función de la cantidad de veces que aparece una palabra en un corpus de textos escritos u orales (Kliegl *et al*. 2004; Grainger 1990, Yap & Balota 2015).

Específicamente, la frecuencia léxica es una de las variables psicolingüísticas más importantes en el estudio de la comprensión y la producción del lenguaje oral y escrito. Esta variable explica por qué las palabras con las cuales nos encontramos más frecuentemente en la vida diaria se procesan de manera más rápida y efectiva que aquellas que rara vez leemos o escuchamos. Toda investigación que se realice en el marco de la psicolingüística, neurolingüística, psicología cognitiva y neurociencias, ya sea con niños y niñas o adultos sanos o sujetos con alguna patología del lenguaje (dislexia, afasia, trastorno del desarrollo del lenguaje, entre otras) en la cual se evalúe la percepción o la producción de ítems léxicos es fundamental que considere la frecuencia léxica para la selección de ítems.

Existen dos formas de obtener datos de la frecuencia de los ítems léxicos. Por un lado, la frecuencia léxica puede medirse de manera objetiva, fundamentalmente a partir de corpus que recopilan materiales escritos como diarios, revistas, libros de texto, etc. Estos se procesan y permiten obtener información acerca de la cantidad de apariciones que tienen los distintos ítems léxicos que se registran, lo cual se traduce o interpreta como la frecuencia de exposición a estas palabras. Esta forma de cuantificar la frecuencia léxica ha recibido críticas porque se basa fundamentalmente en materiales escritos (Gilhooly & Logie 1980; Gernsbasher 1984; Balota, Pilotii & Cortese 2001; Baayen, Milin & Ramscar 2016). El otro modo de considerar la frecuencia, es de manera subjetiva a partir de escalas de percepción personal. Este modo tampoco está exento de discusiones teóricas, ya que algunos autores plantean que las consignas que se administran para estimar la frecuencia subjetiva pueden superponerse tanto con diferentes variables como la familiaridad, la edad de adquisición, la edad de los participantes, así como también con variables semánticas y aspectos relativos a la información fonológica y ortográfica (Balota, Pilotti & Cortese 2001) y además, las normas disponibles suelen ser escasas y con una cantidad muy reducida de ítems (Ferrand 2008). Sin embargo, suele utilizarse en estudios de adultos o de niños cuando no existen herramientas objetivas.

Resulta importante que los investigadores cuenten con herramientas fiables que les permitan seleccionar los estímulos de acuerdo a esta variable. Por ello, hasta el momento la mejor manera de obtenerla es de manera objetiva. En el caso del español rioplatense, no se cuenta con un diccionario de frecuencia para la región para ninguna población. *et al*.En este contexto, las diversas investigaciones y herramientas de evaluación se realizan recurriendo a bases generadas en otros dialectos del español, generalmente el español ibérico.

1.1 | Los diccionarios de frecuencia

Las investigaciones en el marco de la psicolingüística realizadas en diferentes lenguas utilizan bases de datos de palabras para obtener datos sobre frecuencia léxica. Estos corpus pueden elaborarse a partir de un número representativo de diversos tipos textuales en soporte papel, digitalizados o provenientes de sitios web (diarios, revistas, textos literarios, ensayos, etc.) o al recopilar subtítulos de películas o series en sitios web especializados.

Los diccionarios para población adulta que se destacan para el inglés son los trabajos de Thorndike & Lorge (1944), de Kučera & Francis (1967), de Carroll, Davies & Richman (1971), el programa *N-Watch* de Davis (2005) que permite obtener información de distintas variables como frecuencia léxica, edad de adquisición, vecindad ortográfica e imaginabilidad, entre otras y el *Corpus of Contemporary American English* (COCA) (Davies 2008) y su ampliación (Gardner & Davies 2014). En el caso del francés las herramientas más citadas

son las de Imbs (1971), *Brulex* (Content, Mousty & Radeau 1990) y *Lexique* (New *et al*. 2004). Para el ruso se ha presentado recientemente una herramienta denominada *StimulStat* que brinda información sobre distintas variables lingüísticas y psicolingüísticas como longitud, propiedades gramaticales, la frecuencia léxica, vecindad ortográfica y fonológica, entre otras (Alexeeva, Slioussar & Chernova 2018). Por último, existen otras herramientas que ofrecen datos de frecuencia para distintas lenguas de manera conjunta, como es el caso de *Celex*, un diccionario de frecuencia para el inglés, el holandés y el alemán realizado por Baayen, Piepenbrock & Gulikers (1995).

Para el español, los diccionarios disponibles están enfocados en su mayoría para la población adulta y se basan en corpus recolectados para la variedad ibérica. Hasta mediados de los años 90 del siglo pasado se utilizaba la base compilada por Juilland y Chang-Rodriguez (1964) basada en un corpus de prosa escrita (teatro, novela, ensayos, periódicos y obras técnicas) publicada en España entre 1920 y 1940. Este diccionario de frecuencia ofrece información sobre el coeficiente de uso, frecuencia y dispersión de 5024 ítems léxicos.

Dada las limitaciones de esta base de datos, Alameda y Cuetos (1995) publicaron el *Diccionario de frecuencia de las unidades lingüísticas del español*. Este corpus se realizó en base a un corpus de 2 millones de palabras provenientes de diversos tipos de textos escritos producidos entre 1978 y 1993. El mismo ofrece información sobre la frecuencia de aparición escrita para 81.323 palabras así como la frecuencia para sílabas, de bigramas y letras.

En el año 2000 se publicó otra herramienta muy difundida para el español ibérico, el *Lexesp*, compilado por Sebastián-Gallés *et al.* (2000) que no solo amplía la base de Alameda y Cuetos sino también incorpora índices relevantes para la selección de estímulos como imaginabilidad, concretud y familiaridad, número de sílabas, posición del acento y pronunciación. Este diccionario se presenta con una aplicación llamada *Corco* que facilita el acceso a las distintas variables mencionadas. Sin embargo, este software no permite obtener información sobre variables de suma relevancia para la investigación psicolingüística como son medidas de vecindad fonológica y ortográfica, edad de adquisición, y frecuencia silábica, entre otras. Para superar estas limitaciones se creó una aplicación denominada *Buscapalabras* (B-pal) (Davis & Perea 2005) que es la versión en español de la aplicación *N-watch* para el inglés diseñada por Davis (2005).

El 2005 también se publicó *Frecuencias del español. Diccionario de estudios léxicos y morfológicos* de Almela *et al.* (2005). Este diccionario está basado en el *Corpus Cumbre* que contempla 20 millones de palabras presentes en fragmentos de textos orales y escritos recientes de España e Hispanoamérica. Otra herramienta confeccionada por autores españoles es el *EsPal* (Duchon *et al.* 2013) que se construyó a partir de textos escritos presentes en la web así como textos gubernamentales, diarios, textos literarios y subtítulos de series y películas. Este corpus brinda datos para el español ibérico y de Latinoamérica de múltiples variables involucradas en el reconocimiento de las palabras: frecuencia léxica, características ortográficas, vecinos ortográficos y fonológicas, imaginabilidad, propiedades de la estructura subléxica de como bigramas, trigramas, entre otros datos.

Para finalizar, otros corpus realizados para otras variedades del español que se pueden mencionar son los siguientes: *ACTIV-ES* de Francom, Hulden & Ussishkin (2014) que propone un corpus basado en subtítulos de películas para el español de Argentina, México y España, y *Diccionario de frecuencia de español de México, LexMex*, publicado por Silva-Pereyra *et al.* (2014), un corpus confeccionado a partir de textos extraídos de 32 publicaciones periódicas digitales.

En el caso de la población infantil, la cantidad de herramientas disponibles es más escasa tanto en español como en otras lenguas. Los diccionarios más citados para el inglés son *The American Heritage Word Frequency Book* (Carroll, Davies & Richman 1971), *The Educator´s Word Frequency Guide* (Zeno *et al.* 1995), y el

diccionario de frecuencia para niños de 5 a 7 años de Stuart *et al.* (2003) y su revisión y ampliación (Masterson *et al.* 2010). El italiano y el francés también cuentan con una serie de diccionarios de frecuencia infantil para alumnos de nivel primario como el de Marconi *et al.* (1994) y las bases *Novlex* (Lambert & Chesnet 2001) y *Manulex* (Lété, Sprenger-Charolles & Colé 2004) respectivamente.

En español son muy pocos los trabajos que se proponen la elaboración de diccionarios de frecuencia infantil. A continuación se nombran diferentes diccionarios diseñados en español en su variedad ibérica. Un trabajo pionero en esta temática es el de Justicia (1995) que propone un diccionario de frecuencia de vocabulario productivo escrito. Esta base fue desarrollada a partir de una muestra de 225.711 palabras donde los ítems léxicos fueron extraídos de producciones escritas espontáneas de tema libre y de un test asociación libre de palabras de niños y niñas de entre 6 y 10 años de edad que concurrían de 1° a 5° de la Educación General Básico (EGB) de escuelas de zona rural y urbana de distinto nivel socioeconómico de las provincias de Almería, Granada, Jaén y Málaga. El corpus está conformado por un total de 5.750 palabras cuya frecuencia media es de 44,47. Este diccionario de frecuencia funcionó como un insumo para que Justicia *et al.* (1996) confeccionaran el *Diccionario de frecuencia silábica infantil*. Luego, podemos citar el *Diccionario de Frecuencia del Castellano Escrito* en niños y niñas de 6 a 12 años de Martínez-Martín y García-Pérez (2004) que es la herramienta más difundida y utilizada para la selección de estímulos para test estandarizados en español tanto en su variedad ibérica como en la rioplatense. Esta base de datos está conformada a partir de libros de texto y de lectura utilizados por niños de 1° a 6° grado de la escuela primaria española. Para elaborar el corpus, se contabilizó la cantidad de libros leídos por la totalidad de niños y niñas de los diferentes grados y se estableció el porcentaje de textos leídos por ocho estudiantes de cada curso. Este corpus de datos brinda información sobre la frecuencia de cada palabra según nivel escolar y la frecuencia acumulada de cada ítem léxico para todos los grados. Martínez-Martín y García-Pérez (2008) utilizaron el Diccionario de Frecuencia del Castellano Escrito en niños de 6 a 12 años para elaborar una base de vecinos ortográficos de palabras que los niños españoles leen durante la escuela primaria. Por último, se puede nombrar la base de datos léxicos de niños españoles de jardín de infantes y primer grado, *Lexin*, elaborada por Corral, Ferrero y Goikoetxea (2009). Esta base de datos se compiló a partir de 134 libros de lectura y escritura españoles para niños que comienzan el proceso de alfabetización. Para calcular la frecuencia de los ítems léxicos presentes en estos materiales, se realizó la transcripción de los distintos libros y se les asignó la categoría gramatical en base al *Lexesp* (Sebastián-Gallés *et al.* 2000).

A pesar de contar con diccionarios, al igual que ocurre con otras bases de datos, los resultados obtenidos en una lengua no pueden ser extrapolados a otras. Tampoco es suficiente contar con datos para una única variedad, ya que las diferencias dialectales implican usos distintos de las mismas formas e incluso el uso de formas léxicas diferentes para referirse a los mismos objetos o conceptos. Asimismo, se debe tener en cuenta que la información que se obtiene sobre los distintos dialectos debe actualizarse y ampliarse dado que el paso del tiempo produce cambios que deben ser tenidos en cuenta. Por otra parte, a partir de los datos presentados se puede observar que la mayoría de los diccionarios y corpus están pensados para la población adulta y son muy escasas las herramientas que permiten controlar variables psicolingüísticas para la población infantil.

El objetivo de este trabajo es mostrar la elaboración de una herramienta, el *Cuenta palabras*, que permite consultar principalmente la frecuencia léxica infantil a partir de un corpus de textos producidos en la variedad rioplatense. Esta herramienta tiene transferencia fundamentalmente en el área de la investigación psicolingüística, pero también resulta de central importancia para la elaboración de materiales destinados a la clínica de los trastornos del lenguaje y la educación.

2 | METODOLOGÍA

2.1 | Materiales

El corpus de *Cuenta palabras* fue elaborado a partir de 57 textos escolares de 1º a 7º grado de nivel primario de diversas asignaturas (lengua, ciencias sociales y ciencias naturales). Los manuales escolares eran material utilizado en Argentina, fundamentalmente en la Ciudad Autónoma de Buenos Aires y en la Provincia de Buenos Aires. En la Tabla 1 figura la información referente a cantidad de manuales por años escolares y por área.

TABLA 1 Distribución de materiales utilizados para la confección del diccionario por área y por grado escolar

Grado	Prácticas del lenguaje	Ciencias Sociales	Ciencias Naturales	Biciencias	Áreas integradas
1°	1			2	2
2°	3			2	4
3°				2	5
4°	6	3	3	1	
5°	5	4	2	1	
6°	4	8	6		
7°	7	1		2	

2.2 | Preparación del corpus

Para la elaboración del corpus los textos en formato físico (N=23) fueron digitalizados mediante escáner óptico y procesados con sistemas de Reconocimiento Óptico de Caracteres (OCR) provisto directamente por el sistema de escaneo. Por otro lado, los textos obtenidos directamente en formato digital (archivos en formato PDF) pero provenientes de escaneos (N=7) fueron procesados con un sistema de OCR implementado en la librería *pytesseract* de Python3. El resto de los textos fueron obtenidos originalmente en formato digital con acceso directo al texto (N=27). Para cada uno de los documentos analizados se eliminaron manualmente aquellas páginas que contaban con texto no dirigido a los estudiantes.

2.3 | Cálculo de frecuencias

Una vez preprocesados todos los documentos, los textos fueron cargados, segmentados (*parseados*) y normalizados para calcular la frecuencia total (cantidad total de palabras) y la frecuencia de cada *token* (i.e. de cada cadena de caracteres) mediante la función FreqDist de la librería NLTK. Todo el procesamiento se realizó mediante scripts propios, implementados en Python3, con las librerías de Procesamiento de Lenguaje Natural NLTK y SpaCy. El *parseo* se realizó separando los textos en espacios y eliminando los distintos signos de puntuación (comas, puntos, paréntesis, interrogativos, entre otros), de forma de obtener sólo las palabras compuestas por caracteres alfabéticos. Debido a posibles errores de reconocimiento de caracteres por parte de los sistemas de OCR, aquellas palabras compuestas por caracteres numéricos no fueron eliminadas de forma automática (ver sección 2.3 Corrección manual).

2.4 | Corrección manual

Luego de generar una primera versión del diccionario de frecuencias se notó que existían muchos *tokens* que no correspondían a palabras del español. Por esta razón, se realizó una inspección manual, con especial foco en los *tokens* de frecuencia absoluta 1. En esta revisión se corrigieron errores provenientes del OCR (por ejemplo: donde el OCR reconoció *surgimieh0* se corrigió por *surgimiento*), errores provenientes de la segmentación, en general por falta espacios entre palabras (por ejemplo: se cambió *hormigascomen* por *hormigas* y *comen*) y otros casos puntuales. En todos estos, la frecuencia original del *token* erróneo fue sumada a la frecuencia del o de los *tokens* corregidos. Durante la corrección manual también fueron eliminados nombres propios y *tokens* que no pudieron ser identificados como palabras del español.

2.5 | Análisis de frecuencias y disponibilidad de los datos

Todos los análisis que serán presentados en el presente trabajo fueron realizados mediante códigos propios en Python3. Todo el código, en conjunto con la base de datos, estará disponible para acceso público en la página web[1] del Instituto de Lingüística, de la Facultad de Filosofía y Letras de la Universidad de Buenos Aires.

3 | RESULTADOS

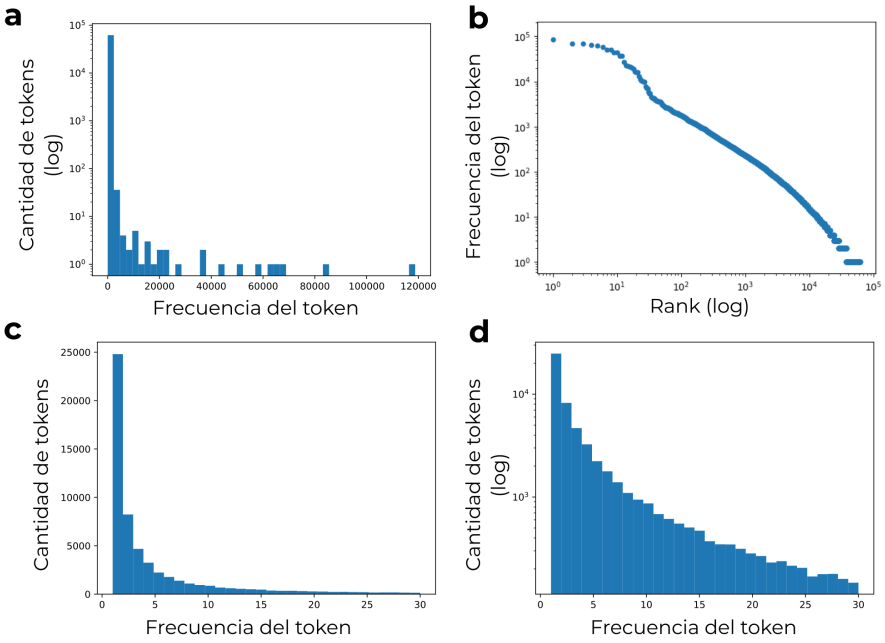

FIGURA 1 **(a)** Histograma de frecuencia de tokens. La Frecuencia está expresada en unidades logarítmicas. **(b)** Relación entre (el logaritmo de) la frecuencia de los tokens y (el logaritmo de) la posición de cada token en el ranking. **(c)** Histograma de frecuencias para los primeros 9 deciles de la distribución. **(d)** Histograma de frecuencias (en base logarítmica) para los primeros 9 deciles de la distribución.

[1] (http://il.institutos.filo.uba.ar/)

Luego de realizado el pre-procesamiento y la corrección manual el diccionario cuenta con un total de 2.203.703 palabras, correspondientes a 61.967 *tokens* diferentes. La distribución de frecuencias muestra un gran porcentaje de *tokens* que aparecen una cantidad muy baja de veces (ver FIGURA 1a). Se puede notar que el eje vertical del histograma principal está presentado en unidades logarítmicas. De acuerdo a la literatura se espera que la distribución de frecuencias encontradas siga una distribución cercana a la distribución de Zipf (Piantadosi 2014). Esta distribución se caracteriza por una tendencia lineal al analizar, en base logarítmica, la distribución de frecuencias en función del *ranking* de las palabras según su frecuencia (ver FIGURA 1b).

Profundizando el análisis sobre la distribución de frecuencias se dividió el diccionario en deciles. De acuerdo a esta división, casi el 40 % del corpus está compuesto por palabras de frecuencia 1. Es decir, del total de 61.967, 24.797 *tokens* aparecen una sola vez en todos los documentos analizados (ver Tabla 2). Al analizar la distribución de frecuencias para el 90 % de palabras menos frecuentes (FIGURA 1d) se puede observar mejor la diferencia en la cantidad de palabras con cada frecuencia. Inclusive, es posible observar este efecto de forma más clara, al analizar la escala lineal (FIGURA 1c).

TABLA 2 Cantidad de palabras segmentadas por frecuencia o rango de frecuencia y ejemplos

Frecuencia	Cantidad de palabras	Decil	Ejemplos
1	24.797	1-5	['utilero', 'amputada', 'mascullar', 'fusobacterias', 'defendidas']
2	8.230	6	['peregrinación', 'temporalidad', 'arterias', 'lustró', 'incumplir']
3	4.678	7	['desarmado', 'analogíassol', 'explicados', 'migratoria', 'navales']
4	3.253	7	['sarcófago', 'forja', 'solapa', 'enfatizar', 'reprimidos']
5	2.232	7	['compacta', 'asombrados', 'graficar', 'desapariciones', 'maza']
6-10	6.077	8	6: ['bucal']; 7: ['elipsis']; 8: ['planetarias']; 9: ['comarca']; 10: ['peperina']
11-30	6.475	9	11: ['azucarera']; 13: ['calabaza']; 15: ['porotos']; 17: ['cordones']; 19: ['hornero']; 21: ['carpa']; 23: ['violenta'] 25: ['obligados']; 27: ['manada']; 29: ['mentira']; 30: ['bomberos']
31-100	3.929	10	31: ['manifestaciones']; 40: ['crónicas']; 50: ['órbitas']; 60: ['prudencia']; 70: ['bailar']; 80: ['verduras']; 90: ['cosecha']; 100: ['interesante']
+100	2.296	10	150: ['ayer']; 250: ['distancia']; 1000 : ['antes']; 2035 ['animales']; 3243: ['otros']; 4290 ['pero']; 5395 ['cada']; 6885: ['sus']; 12815: ['no']; 84460: ['la']
TOTAL	61.967		

En *Cuenta palabras*, la palabra que presenta mayor frecuencia es “de” cuya cantidad de apariciones es 118.730 y, como se aprecia en la Tabla 2, una gran cantidad de palabras (24.797) aparece sólo una vez y, por lo tanto, su frecuencia es 1.

Al analizar a los *tokens* en base a su categoría gramatical más frecuente (de acuerdo a la base de datos EsPal [Duchon *et al.* 2013]) podemos observar que la palabra de contenido (sustantivo, adjetivo o verbo) más frecuente es “es” (16.233 apariciones), y en particular, el sustantivo más frecuente es “texto” (3.550 apariciones). Las distribuciones de frecuencias de las dos categorías (palabras de clase abierta y clase cerrada) muestran una clara diferencia. Por un lado, las palabras de clase abierta o contenido aparecen, en promedio 19.2 veces cada una en los textos, mientras que en las de clase cerrada o de función ese número asciende a 3059.9 (ver FIGURA 2).

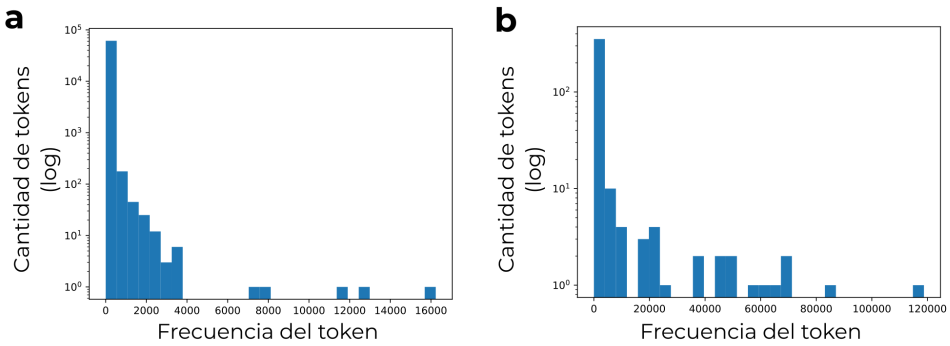

FIGURA 2 Distribuciones de frecuencias para **(a)** palabras de clase abierta o contenido (sustantivos, adjetivos, verbos y adverbios) y **(b)** palabras de clase cerrada o funcionales (conjunciones, pronombres, preposiciones, determinantes, entre otras). Las frecuencias están expresadas en unidades logarítmicas.

4 | DISCUSIÓN

El diccionario de frecuencia infantil presentado en este trabajo configura una herramienta relevante para lingüistas, psicólogos y otros científicos que lleven adelante estudios experimentales sobre el procesamiento del lenguaje en español rioplatense. También es un recurso destacable para la elaboración de materiales didácticos para trabajar en el aula con niños y niñas de nivel primario y para la enseñanza del español como segunda lengua. Asimismo, los datos arrojados por esta herramienta permiten un mayor control para la elaboración de materiales de evaluación y tratamiento en la práctica clínica de las alteraciones del lenguaje.

Cuenta palabras es una herramienta que permite acceder a la frecuencia léxica escrita de 61.967 palabras únicas, distribuidas en frecuencias que van de 1 a 118.730. El corpus contempló materiales didácticos utilizados en el nivel de escolaridad primario fundamentalmente de la Ciudad de Buenos Aires y de la Provincia de Buenos Aires, por lo que comprende un rango de edad de niños y niñas de 6 a 13 años.

Es importante destacar que la mayoría de las palabras de alta frecuencia son palabras funcionales o de clase cerrada (artículos, pronombres, conjunciones, preposiciones, entre otras) en detrimento de las palabras de clase abierta o de contenido (sustantivos, adjetivos, verbos, adverbios). Algunos estudios han mostrado un procesamiento diferencial entre las palabras funcionales y las de contenido tanto a nivel conductual como cerebral en población neurotípica y con alteraciones del lenguaje (Herron & Bates 1997; Caramazza & Berndt 1985). Por ejemplo, en un estudio con una tarea de denominación, Segalowitz & Lane (2000) han reportado que el acceso léxico para las palabras funcionales es más rápido que para las palabras de clase abierta y esta diferencia puede atribuirse a la previsibilidad y a la familiaridad de las palabras de clase cerrada. En consonancia

con este efecto, el corpus analizado para este diccionario también permite ver esta diferencia, ya que a pesar de contar con un repertorio mucho más acotado, indefectiblemente están presentes de manera repetida en las diversas construcciones sintácticas y en los textos, con una fuerte función gramatical.

El corpus recolectado en *Cuenta palabras* es equivalente en cantidad de ítems al *Diccionario de Frecuencia del Castellano Escrito* de Martínez-Martín y García-Pérez (2004), por lo tanto a futuro se espera realizar un estudio comparativo entre ambos corpus. *Cuenta palabras* permite la actualización de materiales, por lo que está previsto que se continúe ampliando para, de esta manera, equiparar la distribución de los materiales por áreas y grados. Por último, se complementará este diccionario con valores de familiaridad y edad de adquisición para las palabras presentes en el corpus. Para ello, se diseñarán escalas tipo *likert* para valorar estas variables en la totalidad de las palabras únicas del corpus. Las distintas encuestas presentarán una serie de ítems aleatorizados de baja, media y alta frecuencia que serán suministradas a 30 niños de cada uno de los grados contemplados en la recolección del corpus y 30 docentes de nivel primario.

Fuente de financiación

FILO CyT FC19-049

Referencias

Adelman, James, Suzanne Marquis, Maura Sabatos-DeVito & Zachary Estes (2013). The unexplained nature of reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 39/4: 1037. DOI: 10.1037/a0031829.

Alameda, José Ramón & Fernando Cuetos (1995). *Diccionario de frecuencias de las unidades lingüísticas del castellano*. Oviedo: Servicio de Publicaciones de la Universidad de Oviedo.

Alexeeva, Svetlana, Natalia Slioussar & Daria Chernova (2018). StimulStat: A lexical database for Russian. *Behavior Research Methods* 50: 2305–2315.

Almela, Ramón, Pascual Cantos, Aquilino Sánchez, Ramón Sarmiento & Moisés Almela (2005). *Frecuencias del español. Diccionario de estudios léxicos y morfológicos*. Madrid: Editorial Universitas.

Ashby, Jane & Keith Rayner (2004). Representing syllable information during silent reading: Evidence from eye movements. *Language and Cognitive Processes* 19/3: 391–426. DOI: 10.1080/0169096034400023

Baayen, Harald, Richard Piepenbrock & Leon Gulikers (1995). *CELEX2 LDC96L14*. Philadelphia: Linguistic Data Consortium.

Baayen, Harald, Petar Milin & Micheal Ramscar (2016). Frequency in lexical processing. *Aphasiology* 30/11: 1174–1220.

Balota, David, Maura Pilotti & Micheal Cortese (2001). Subjective frequency estimates for 2,938 monosyllabic words. *Memory & Cognition* 29/4: 639–647.

Ballot, Claire, Stephanie Mathey & Christelle Robert (2021). Word imageability and orthographic neighbourhood effects on memory: a study in free recall and recognition. *Memory* 29/6: 829–834. DOI: 10.1080/09658211.2021.1921216.Bijeljac-Babic, Ranka, Victor Millogo, Fernand Farioli & Jonathan

Grainger (2004). A developmental investigation of word length effects in reading using a new on-line word identification paradigm. *Reading and Writing* 17: 411–431. 10.1023/B:READ.0000032664.20755.af

Caramazza, Alfonso & Rita Berndt (1985). A multicomponent deficit view of agrammatic Broca's aphasia. En Kean, M.-L. (Ed.), *Agrammatism*. Orlando: Academic Press, 27–63.

Carreiras, Manuel & Jonathan Grainger (2004). Sublexical Representations in Visual Word Recognition. A special issue of language and cognitive processes. *Language and Cognitive Processes* 19/3.

Carroll, John B., Peter Davies & Barry Richman. (1971). *The American Heritage Word Frequency Book*. New York: American Heritage.

Colombo, Lucia, Margherita Pasini & David Balota (2006). Dissociating the influence of familiarity and meaningfulness from word frequency in naming and lexical decision performance. *Memory & Cognition* 34/6: 1312–1324. DOI: 10.3758/bf03193274

Content, Alan, Phillip Mousty & Monique Radeau (1990). *Brulex.* Une base de données lexicales informatisée pour le français écrit et parlé. *L'année psychologique* 90/4: 551–566.

Corral, Silvia, Marta Ferrero & Edurne Goikoetxea (2009). LEXIN: A lexical database from Spanish kindergarten and first-grade readers. *Behavior Research Methods* 41: 1009–1017. DOI: 10.3758/BRM.41.4.1009

Davies, Mark (2008). *Corpus of Contemporary American English (1990–2019)*. Disponible en: http://corpus.byu.edu/coca/

Davis, Colin (2005). N-Watch: A program for deriving neighborhood size and other psycholinguistic statistics. *Behavior Research Methods* 37: 65–70.

Davis, Colin & Manuel Perea (2005). BuscaPalabras: A program for deriving orthographic and phonological neighborhood statistics and other psycholinguistic indices in Spanish. *Behavior Research Methods* 37: 665–671.

Duchon, Andrew, Manuel Perea, Nuria Sebastián-Gallés, Antonia Martí & Manuel Carreiras (2013). EsPal: one-stop shopping for Spanish word properties. *Behavior Research Methods* 45: 1246.

Fenk-Oczlon, Gertraud & Jürgen Pilz (2021). Linguistic complexity: relationships between phoneme inventory size, syllable complexity, word and clause length, and population size. *Frontiers in Communication* 6.

Ferrand, Ludovic, Patrick Bonin, Alain Méot, Maria Augustinova, Boris New, Christoph Pallier & Mark Brysbaert (2008). Age-of-acquisition and subjective frequency estimates for all generally known monosyllabic French words and their relation with other psycholinguistic variables. *Behavioral Reserach Methods* 40/4: 1049–1054. doi: 10.3758/BRM.40.4.1049. PMID: 19001395.

Foster, Kenneth (1976). Accessing the mental lexicon. En: R. J. Wales & E. Walker (Eds.). *New approaches to language mechanisms* (pp. 257-287). Amsterdam: North Holland.

Francom, Jerid, Mans Hulden, & Adam Ussishkin (2014). ACTIV-ES: a comparable, cross-dialect corpus of 'everyday' Spanish from Argentina, Mexico, and Spain. En Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). Reykjavik, Iceland. European Language Resources Association (ELRA): 1733–1737.

Fumagalli, Julieta, Juan Pablo Barreyro, Ana María Borzone & Virginia Jaichenco (2014). Incidencia del tipo de unidad y la complejidad silábica en una tarea de conciencia fonológica. *Estudios de Lingüística Aplicada*

32/60: 35–55.

Gardner, Dee & Mark Davies (2014). A New Academic Vocabulary List. *Applied Linguistics* 35/3: 305–327. DOI: doi.org/10.1093/applin/amt015

Gernsbacher, Morton (1984). Resolving 20 years of inconsistent interactions between lexical familiarity and orthography, concreteness, and polysemy. *Journal of Experimental Psychology: General* 113/2: 256–281. https://doi.org/10.1037/0096-3445.113.2.256

Gilhooly, Kent & Robert Logie (1980). Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words. *Behavior Research Methods & Instrumentation* 12/4: 395–427. https://doi.org/10.3758/BF03201693

Grainger, Jonathan (1990). Word frequency and neighborhood frequency effects in lexical decision and naming. *Journal of Memory and Language* 29/2: 228–244. DOI: 10.1016/0749-596X(90)90074-AHerron, Daniel & Elizabeth Bates (1997). Sentential and acoustic factors in the recognition of open- and closed-class words. *Journal of Memory and Language* 37: 217–239.

Hendrix, Peter & Chu Sun Ching (2021). A word or two about nonwords: frequency, semantic neighborhood density, and orthography-to-semantics consistency effects for nonwords in the lexical decision task. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 47/1: 157.

Huntsman, Laree & Susan Lima (2002). Orthographic neighbors and visual word recognition. *Journal of Psycholinguistic Research* 31/3: 289–306. DOI: 10.1023/a:1015544213366

Imbs, Paul (1971). *Dictionnaire des fréquences: Vocabulaire littéraire des XIXe et XXe siècles. I: Table alphabétique. II: Table des fréquences décroissantes*. Nancy, Paris: CNRS, Didier.

Juilland, Alphonse & Eugenio Chang-Rodríguez (1964). *Frequency dictionary of Spanish words*. La Haya: Mouton.

Justicia, Fernando (1995). *El desarrollo del vocabulario: Diccionario de frecuencias*. Granada: Universidad de Granada.

Justicia, Fernando, Julio Santiago, Alfonso Palma, Dolores Huertas & Nicolás Gutiérrez (1996). La frecuencia silábica del español escrito por niños: Estudio estadístico. *Cognitiva* 8: 131–168.

Khanna, Maya & Michael Cortese (2021). How well imageability, concreteness, perceptual strength, and action strength predict recognition memory, lexical decision, and reading aloud performance. *Memory* 29: 1–15. DOI: 10.1080/09658211.2021.1924789.

Kliegl, Reinhold, Ellen Grabner, Martin Rolfs & Ralf Engbert (2004). Length, frequency, and predictability effects of words on eye movements in reading. *European Journal of Cognitive Psychology* 16/1-2: 262–284. DOI: 10.1080/09541440340000213

Kučera, Henry, & Nelson Francis. (1967). *Computational Analysis of Present-Day American English*. Providence: Brown University Press.

Lambert, Eric & David Chesnet (2001). NOVLEX. A lexical database for elementary school students. *Annee Psychologique* 101: 277–288.

Libben, Gary & Gonia Jarema (2002). Mental Lexicon Research in the New Millennium. *Brain and Language*, 81/1–3: 2–11.

Leroy, Gondy & David Kauchak (2014). The effect of word familiarity on actual and perceived text difficulty. *Journal of the American Medical Informatics Association*: *JAMIA* 21: 169–172.

Lété, Bernard, Liliane Sprenger-Charolles & Pascale Colé (2004). MANULEX: a grade-level lexical database from French elementary school readers. *Behavior Research Methods Instruments and Computers* 36/1: 156–66.

Marconi, Lucia, Michaela Ott, Elia Pesenti, Daniela Ratti & Mauro Tavella (1994). *Lessico elementare: Dati statistici sull'italiano scritto e letto dai bambini delle elementary*. Bologna: Zanichelli

Martínez-Martín, Jesús, & Emma García Perez (2004). *Diccionario de frecuencias del castellano escrito en niños de 6 a 12 años*. Salamanca: Universidad Pontificia de Salamanca.

Martínez-Martín, Jesús & Emma García Pérez (2008). ONESC: a database of orthographic neighbors for Spanish read by children. *Behavior Research Methods* 40/1: 191–197. DOI: 10.3758/brm.40.1.191

Masterson, Jackie, Morag Stuart, Maureen Dixon & Sophie Lovejoy (2010). Children's printed word database: Continuities and changes over time in children's early reading vocabulary. *British Journal of Psychology* 101/2: 221–242. DOI: 10.1348/000712608x371744

Morrison, Catriona M., Katherine Hirsh, Tameron Chappell & Andrew W. Ellis (2002). Age and age of acquisition: An evaluation of the cumulative frequency hypothesis. *European Journal of Cognitive Psychology* 14/4: 435–459.

New, Boris, Christophe Pallier, Marc Brysbaert & Ludovic Ferrand (2004). Lexique 2: A new French lexical database. *Behavior Research Methods, Instruments, & Computers* 36/3: 516–524. DOI: 10.3758/BF03195598

Perea, Manuel (2015). Neighborhood effects in visual word recognition and reading. En A. Pollatsek & R. Treiman (Eds.), *The Oxford Handbook of Reading*. Nueva York: Oxford University Press, 76–87.

Perea, Manuel & Eva Rosa (1999). Psicología de la lectura y procesamiento léxico visual: una revisión de técnicas experimentales y procedimientos de análisis. *Psicológica* 20: 65–90.

Piantadosi, Steven. (2014). Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic bulletin & review* 21: 1112–1130.

Riecker, Axel, Bettina Brendel, Wolfram Ziegler, Michael Erb & Hermann Ackermann (2008). The influence of syllable onset complexity and syllable frequency on speech motor control. *Brain and Language* 107/2: 102–113. DOI: 10.1016/j.bandl.2008.01.008

Sebastián-Gallés, Nuria, María Antonia Martí, Manuel Carreiras & Fernando Cuetos (2000). *LEXESP. Léxico informatizado del español*. Barcelona: Edicions Universitat de Barcelona.

Segalowitz, Sid J. & Korri C. Lane (2000). Lexical access of function versus content words. *Brain and Language* 75: 376–389.

Segui, Juan & Jonathan Grainger (1990). Priming word recognition with orthographic neighbors: effects of relative prime-target frequency. *Journal of experimental psychology. Human perception and performance* 16/1: 65–76.

Silva-Pereyra, Juan; Mario Rodríguez-Camacho, Belén Prieto & Eduardo Aubert (2014). *LEXMEX: Diccionario de frecuencias del español de México*. México D.F.: Editorial FES Iztacala UNAM.

Stuart, Morag, Maureen Dixon, Jackie Masterson & Bob Gray (2003). Children's early reading vocabulary: Description and word frequency lists. *British Journal of Educational Psychology* 73/4: 585–598. DOI: 10.1348/000709903322591253

Taft, Marcus & Kenneth Forster (1975). Lexical storage and retrieval of prefixed words. *Journal of Verbal Learning & Verbal Behavior* 14/6: 638–647. DOI: 10.1016/S0022-5371(75)80051-X

Thorndike, Edward & Irving Lorge (1944). *The teacher's word book of 30,000 words.* New York: Teachers College, Columbia University.

Vergara-Martínez, Marta, Pablo Gómez & Manuel Perea (2020). Should I stay or should I go? An ERP analysis of two-choice versus go/no-go response procedures in lexical decision. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 46/11: 2034.

Yap, Melvin & David Balota (2015). Visual word recognition. En A. Pollatsek & R. Treiman (Eds.), *The Oxford handbook of reading.* Nueva York: Oxford University Press, 26–43.

Zeno, Suzanne, Steve Ivens, Robert Millard & Raj Duvvuri (1995). *The educator's word frequency guide*. Brewster, New York: Touchstone Applied Science Associates.