1º Congreso Internacional de Ciencias Humanas - Humanidades entre pasado y futuro. Escuela de Humanidades, Universidad Nacional de San Martín, Gral. San Martín, 2019.

Cuenta palabras: Diccionario de frecuencia léxica infantil para el español rioplatense.

Fumagalli, Julieta, Sanchez, María Elina, Picone, Muriel, Cancino, Matías, Oliva, María Paz, Melman, Martín, Checchi, Sofia y Giussani, Laura.

Cita:

Fumagalli, Julieta, Sanchez, María Elina, Picone, Muriel, Cancino, Matías, Oliva, María Paz, Melman, Martín, Checchi, Sofia y Giussani, Laura (2019). *Cuenta palabras: Diccionario de frecuencia léxica infantil para el español rioplatense. 1º Congreso Internacional de Ciencias Humanas - Humanidades entre pasado y futuro. Escuela de Humanidades, Universidad Nacional de San Martín, Gral. San Martín.*

Dirección estable:
https://www.aacademica.org/1.congreso.internacional.de.ciencias.humanas/1650

ARK: https://n2t.net/ark:/13683/eRUe/Uzh

Cuenta palabras: Diccionario de frecuencia léxica infantil para el español rioplatense

Fumagalli, J.[12]; Sanchez, M.E.[12]; Picone, M.[2], Cancino, M.[2] ; Oliva, M.P.[2] ; Melman, M.[2] ; Checchi, S.[2]; Giussani, L.[2]
[1] CONICET
[2] Universidad de Buenos Aires, Facultad de Filosofía y Letras, Instituto de Lingüística
julietafumagalli@filo.uba.ar

Resumen

Introducción: La frecuencia léxica es una de las variables psicolingüísticas más importantes en el estudio del procesamiento de la comprensión y la producción del lenguaje. Esta variable puede explicar por qué hay palabras que se procesan de manera más rápida y efectiva que otras. Toda investigación que se realice en el marco de disciplinas como la psicolingüística, neurolingüística, psicología cognitiva y neurociencias del lenguaje, en la que se evalúe la comprensión o la producción de ítems léxicos debe considerar la frecuencia de uso de las palabras en la selección de estímulos. En el caso del español rioplatense, no contamos con diccionarios de frecuencia para la región ni de la población adulta ni de la infantil. **Objetivo:** El propósito de esta investigación es mostrar los resultados preliminares de la elaboración de un diccionario de frecuencia léxica infantil (niños entre 5 y 12 años) para el español en su variedad rioplatense. **Materiales:** El corpus de palabras a presentar estará conformado a partir de 27 textos escolares de distintas asignaturas de 1º grado a 7º grado de nivel Primario. **Resultados:** Se calculará la frecuencia léxica total de cada ítem y la frecuencia según nivel escolar.

Palabras clave: Frecuencia léxica; Comprensión del lenguaje; Producción del Lenguaje; Niños; Rioplatense

Introducción

Para comprender y producir lenguaje tanto adultos como niños debemos 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 aprendemos a leer y a escribir) de todas las palabras que conocemos. Las investigaciones psicolingüísticas han utilizado diversas tareas para indagar acerca de la organización del léxico mental, las más difundidas son las tareas de decisión léxica, denominación y lectura.

A partir de estas tareas, que miden la precisión y el tiempo requerido para resolverlas, los investigadores pudieron detectar diversas variables que explican las respuestas de los participantes. Entre estas variables se identifican, por un lado, variables lingüísticas como la longitud de la palabra y la complejidad silábica u ortosilábica y, por otro, variables psicolingüísticas como la vecindad ortográfica y fonológica, la familiaridad, la edad de adquisición, los efectos de concretud e imaginabilidad y la frecuencia léxica.

Variables psicolingüísticas: La frecuencia léxica

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. Cualquier investigación que se realice en el marco de la psicolingüística, neurolingüística, psicología cognitiva y neurociencias, ya sea con niños o adultos sanos o con sujetos con patologías (dislexia, afasia, demencias, etc.) en la que se evalúe la percepción o la producción de ítems léxicos debe considerar la frecuencia léxica para la selección de estímulos con los que diseñará la tarea. Por lo tanto, los investigadores necesitan herramientas actuales que les permitan seleccionar los estímulos de acuerdo a esta variable. Sin embargo, en el caso del español rioplatense, no contamos con un diccionario de frecuencia para la región y, menos aún con un diccionario de frecuencia de la población infantil. En este contexto, las diversas investigaciones y herramientas de evaluación para nuestra población se realizan recurriendo a bases generadas en otros dialectos del español, generalmente el español ibérico.

Estudios lexicográficos: los diccionarios de frecuencia

Las investigaciones en el marco de la psicolingüística realizadas en diferentes lenguas utilizan bases de datos de palabras para obtener la frecuencia léxica. Estas bases se

obtienen a partir de la recolección de un número representativo de distintos tipos de textos escritos en soporte papel o web (diarios, revistas, textos literarios, ensayos, etc.) que son de lectura frecuente. Otra alternativa, de uso actual, es recurrir a sitios web especializados en subtítulos de películas o series para recopilar el corpus.

En español, la mayoría de los corpus disponibles son para la población adulta y se basan en datos recolectados para la variedad ibérica. Hasta mediados de los años 90 se utilizaba la base compilada por Juilland y Chang-Rodriguez (1964). Alameda y Cuetos (1995) publicaron el Diccionario de frecuencia de las unidades lingüísticas del español realizado en base a 2 millones de palabras provenientes de diversos tipos de textos escritos producidos entre 1978 y 1993. En el año 2000 se publicó otra herramienta muy difundida, el Lexesp (aplicación Corco), compilada por Sebastián Gallés, Martí, Carreiras y Cuetos (2000), que amplía la base de Alameda y Cuetos (1995) e incorpora otros índices relevantes para la selección de estímulos. En 2005 se creó una aplicación denominada Buscapalabras (B-pal) (Davis y Perea, 2005, versión en español de N-whatch (Davis 2005)) que retoma la base de datos del Lexesp e incorpora mayor información sobre variables lingüísticas y psicolingüísticas.Más recientemente, en el año 2013, el Basque Center onCognition, Brain and Language (BCBL) desarrolló el diccionario EsPAL (Duchon et al. 2013). Este se basa en un conjunto de datos de 300 millones de palabras escritas y de 460 millones de subtítulos e incluye entre otras propiedades la frecuencia de palabras, la estructura y vecindad ortográfica, la estructura fonológica, y la imaginabilidad.

En relación con los diccionarios de frecuencia infantil podemos citar muy pocos trabajos y solamente para la variedad ibérica. El primero es el diccionario de frecuencia de vocabulario productivo escrito desarrollado por Justica (1995) basado en un corpus de producciones verbales espontáneas escritas, y de un test asociación libre de palabras de niños de 6 a 10 años de edad de escuelas de zona rural y urbana de distinto nivel socioeconómico del Departamento de Granada. El segundo trabajo a destacar es el *Diccionario de Frecuencia del Castellano Escrito en niños de 6 a 12 años* de Martínez-Martín y García-Pérez (2004) que está construido sobre la base de libros de texto y de lectura utilizados por niños de primero a sexto grado de la escuela primaria española. Para la confección del corpus, se seleccionó un porcentaje de textos leídos por ocho niños de cada grado. Este diccionario permite consultar la frecuencia de cada palabra según curso y la frecuencia acumulada de cada ítem léxico

para todos los grados. Por último, haremos referencia a la herramienta *Lexin*(Corral, Ferrero y Goikoetxea, 2009), una base de datos léxicos de niños españoles de jardín de infantes y primer grado. El corpus fue compilado a partir de 134 libros de lectura y escritura españoles para niños que comienzan el proceso de alfabetización. Para el análisis, los textos se transcribieron para calcular la frecuencia y se les asignó la categoría gramatical en base al *Lexesp* (Sebastián-Gallés et al. 2000).

A pesar de contar con diccionarios para el español, al igual que ocurre con otras bases 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. Por otra parte, la mayoría de los diccionarios y corpus están pensados para la población adulta y son muy escasas las herramientas para controlar variables psicolingüísticas para la población infantil.

Contar con una herramienta que permita controlar diversas variables lingüísticas y psicolingüísticas, y fundamentalmente la frecuencia léxica a partir de una base confeccionada con textos producidos en la variedad rioplatense es de suma relevancia para la investigación, la práctica clínica y la educación en nuestro medio.

Objetivo y metodología

El objetivo de este trabajo es elaborar un diccionario de frecuencia infantil para el español rioplatense. Este comprenderá el vocabulario de niños entre 5 y 12 años, es decir de alumnos que concurren de 1ero a 7mo grado de nivel primario. Para esta instancia del diseño del instrumento, el corpus de palabras se conformará a partir de 27 textos escolares de 1º grado a 7º grado de nivel primario. En la tabla 1 se presenta la distribución de los textos.

Tabla 1. Distribución por áreas y grado de los textos escolares

Grado	Cs. Naturales	Cs. Sociales	Lengua	Biciencias
1º				2
2º				2
3º				2
4º	2	2	2	
5º	2	2	2	
6º	2	2	2	

7º	2	1

Resultados

Para confeccionar el corpus, el texto fue extraído de los manuales de modo directo (conversión de formato) o mediante el uso de OCR (reconocimiento óptico de caracteres) según el caso. Luego fue preprocesado mediante el uso de regex para filtrar los caracteres no alfabéticos y unir las palabras divididas por saltos de línea. Asimismo, se eliminaron indicaciones o aclaraciones para el docente o la familia que estaban intercaladas con el texto dirigido a los niños.

El procesamiento de los datos fue realizado con las herramientas tokenizer y FreqDist de la librería nltk (Python 3) y la versión para español del Stanford Log-linear PoSTagger. Para la lematización se utilizó SnowballStemmer y se eliminaron los nombres propios mediante NamedEntityRecognition.

Con los datos procesados hasta el momento, el corpus tiene un total de 810.799 palabras o tokens. El diccionario resultante (es decir, la cantidad de palabras únicas) contiene 34.480 palabras (incluyendo palabras funcionales (stopwords) y las formas flexivas de las palabras sin lematizar). De esas 34480, 13154 son palabras que aparecen una sola vez en el corpus (hapaxes). En la tabla que se presenta a continuación se brindan ejemplos de palabras de frecuencia alta, media y baja.

Tabla 2. Ejemplos de palabras de frecuencia alta, media y baja

Frecuencia					
Alta		Media		Baja	
de	42959	Estas	810	vendidos	6
la	30224	tiempo	810	acuosos	5
y	23851	animales	809	ortográfica	4
en	23580	Nos	800	paleontología	4
el	23512	Tierra	795	amoroso	3
que	22841	Todo	783	Bárbaro	3
los	18624	nuestro	780	fraternal	2
se	14379	donde	778	gallinero	2
las	13820	País	767	andanza	1
un	10016	información	765	hermandad	1

Conclusiones

Una vez finalizado el diseño, se contará con un instrumento que resultará de suma importancia tanto para la investigación como para la clínica y la educación. Asimismo,

es el puntapié inicial para confeccionar herramientas semejantes para otras variedades dialectales y permitirá también la creación de nuevas bases de datos importantes para la investigación en el área.

Bibliografía

Alameda, J.R. y Cuetos, F. (1995). *Diccionario de frecuencias de las unidades lingüísticas del castellano.* Oviedo: Servicio de Publicaciones de la Universidad de Oviedo.

Corral, S.; Ferrero, M. yGoikoetxea, E. (2009) LEXIN: A lexical database from Spanish kindergarten and first-grade readers, *Behavior Research Methods*, 41 (4), 1009-1017.

Davis, C. y Perea, M. (2005).BuscaPalabras: A program for deriving orthographic andphonological neighborhood statistics and other psycholinguistic indices in Spanish.*Behavior Research Methods*, 37, 665-671.

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

Duchon, A., Perea, M., Sebastián-Gallés, N., Martín, A. y Carreiras, M. (2013).EsPal: one-stop shopping for Spanish word properties.*Behav Res* 45: 1246.

Juilland, A. y Chang-Rodríguez, E. (1964). *Frequency dictionary of Spanish words*.La Haya: Mouton.

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

Martínez, J. A., y García, M. E. (2004). *Diccionario de frecuencias del castellano escrito en niños de 6 a 12 años*. Salamanca: Universidad Pontificia de Salamanca.

Sebastián Gallés, N., Martí, M.A., Carreiras, M. y Cuetos, F. (2000). *LEXESP. Léxico informatizado del español*. Barcelona: EdicionsUniversitat de Barcelona.