Derivación de una lista práctica de frecuencias basada en el Recuento de vocabulario español de Rodríguez Bou

Guadalupe Valdés
Rosalinda Barrera
Donald W. Dearholt
Manuel Cárdenas*

El uso práctico de listas de vocabulario de alta frecuencia ha quedado ya documentado desde hace tiempo. En inglés, por ejemplo, estas listas se han usado como una base sobre la cual se han elaborado textos infantiles; como un elemento importante de las medidas de lecturabilidad para analizar y evaluar la dificultad de materiales de lectura y como punto de partida para establecer normas mínimas de desarrollo de vocabulario.

Por lo general, los compiladores de listas de vocabulario de alta frecuencia, tanto en inglés como en español, han examinado el léxico utilizado en diferentes "universos" (la lengua utilizada en la prensa, la lengua de las novelas y el teatro, etc.) para reflejar así la frecuencia de uso de los distintos elementos en la realidad lingüística social. En inglés, entre las compilaciones de frecuencia de vocabulario se encuentran las listas de Thorndike y Lorge (1948), de Dale y Chall (1948), de Kucera y Francis (1967), de Carroll, Davies y Richman (1971) y de Harris y Jacobson (1972). En español, entre las compilaciones que se han producido en los Estados Unidos para guiar la enseñanza del castellano como lengua extranjera en este país se encuentran la lista de Keniston (1920) **A basic list of Spanish words and idioms**; la lista de Jamieson (1924) **A standardized vocabulary for Elementary Spanish**; la lista de Cartwright (1925) **A study of the vocabularies of eleven Spanish reading texts** y la lista de Buchanan (1929) **A graded Spanish word book**. Entre las listas de vocabulario de alta frecuencia que se han producido, con una sola excepción, en países hispanohablantes para reflejar el uso normal de la lengua en estos contextos se encuentran **Investigaciones acerca de las palabras usadas en castellano en Panamá** (Céspedes, 1929); **Recuento de vocabulario español** (Rodríguez Bou, 1952); **Vocabulario usual, común y fundamental** (García Hoz, 1953) y **Frequency dictionary of Spanish words** (Juilland y Chang Rodríguez, 1968).

En los últimos años, la implementación de programas bilingües en los Estados Unidos ha visto nacer en este país una necesidad de evaluar la dificultad de textos y materiales que se utilizan en dichos programas. En el español, se ha pensado que las listas de frecuencia de vocabulario pueden ser útiles tanto para el educador como para el investigador (Rodríguez Trujillo, 1981 y SCDC, 1974).

---

*Guadalupe Valdés (lingüística española-sociolingüística); Rosalinda Barrera (lectura-educación bilingüe); Donald W. Dearholt (computación) son miembros del personal docente de Nuevo México, Estados Unidos. Desde 1981 a 1984 trabajaron como equipo en un proyecto titulado: "Reading and bilungual education: analyzing texts and children's comprensión", que subvencionara el Instituto Nacional de Educación.

Sin embargo, el usar listas de frecuencia de vocabulario español para trabajar con materiales de instrucción, no es tan sencillo como pudiera creerse a primera vista. En primer lugar, de las listas ya mencionadas, solo el **Recuento de vocabulario español** de Rodríguez Bou refleja el vocabulario utilizado en textos infantiles. Las demás compilaciones se basan exclusivamente en el vocabulario del mundo adulto: prensa, literatura, revistas, teatro, etc.; y no reflejan la frecuencia de uso de elementos léxicos en el material escrito y dedicado a los niños. Claramente, el utilizar una lista de frecuencia, que sólo presenta un índice de vocabulario adulto, para evaluar la dificultad de un texto infantil, llevaría a un juicio decididamente equivocado.

Pero escoger de entre las listas mencionadas, la única compilación que refleja el mundo infantil, tampoco lleva a un juicio claro sobre el material que se desea investigar. Aunque es lógico pensar que podría usarse validamente la lista de Rodríguez Bou, ya que por lo menos incluye los elementos que aparecen con más frecuencia en el mundo hispanohablante en los materiales infantiles; aun el uso de esta compilación presenta varios problemas serios tanto al investigador como al practicante. El propósito de este trabajo es describir la compilación de Rodríguez Bou, dar ejemplos de los problemas que presenta, y finalmente ofrecer un segmento de la lista breve de vocabulario de alta frecuencia derivada del **Recuento** en su totalidad.

Descripción del Recuento de vocabulario español

El **Recuento de vocabulario español** es una obra enorme, de dos volúmenes, cada uno de los cuales incluye más de quinientas páginas. El primer volumen contiene las diez mil unidades léxicas más frecuentes en orden de rango; las diez mil unidades léxicas más frecuentes en orden alfabético, con sus frecuencias y rangos; las veinte mil formas de inflexión más frecuentes en orden alfabético, con sus frecuencias y rangos y las unidades léxicas con frecuencias menores de 16. El segundo volumen (de dos tomos) contiene las frecuencias ponderables de las 20.542 unidades léxicas y de las 62.888 formas de inflexión que se encontraron en las diez fuentes estudiadas.

Diferencia entre las formas léxicas y las formas de inflexión

Para emplear en forma válida el **Recuento de vocabulario español** con el fin de determinar frecuencias de palabras en otros textos, es necesario comprender claramente la diferencia entre las listas de unidades léxicas y las listas de formas de inflexión. Estas últimas contienen los elementos que se encontraron en las fuentes estudiadas exactamente en la forma en que allí aparecieron (verbos conjugados, sustantivos en singular y plural, adjetivos en masculino y femenino, etc.). En cambio, las listas de unidades léxicas contienen la forma base de cada palabra (infinitivos, sustantivos en el singular, adjetivos en el masculino singular y todas las formas que jamás se inflexionan como las conjunciones, adverbios, preposiciones, etc.). Esta diferenciación es importante, ya que el orden en que aparece, por ejemplo, un infinitivo, no indica que todas las formas conjugadas de ese verbo tengan la misma frecuencia.

La importancia de estas clases de listas puede apreciarse en el segundo volumen, en donde aparecen combinadas las 83.430 palabras que se encontraron en las diez fuentes estudiadas. En este volumen, las unidades léxicas se consignan con su letra inicial en mayúscula, mientras que las formas de inflexión que se encontraron aparecen inmediatamente debajo de cada unidad léxica de la cual se derivan. Como ejemplo, véase la **Tabla 1**, en la cual se da la información que aparece en el volumen II del **Recuento** sobre el verbo **abordar**.

Cada una de las columnas representa una fuente de estudio –un universo de usos léxicos (vocabulario oral, material religioso, programas de radio, libros de texto y demás)–. Al lado de las palabras más frecuentes se incluye entre paréntesis un número.

Ese número indica el rango de la palabra de acuerdo con su frecuencia empleando una clave numérica que se define al principio del volumen. En este caso, el verbo **abordar** (unidad léxica) se encontró en cada una de las fuentes estudiadas el número de veces que se indica. Debe notarse, sin embargo, que no todas las formas del verbo abordar se encontraron en el corpus de materiales. Las únicas formas de inflexión que aparecieron fueron las que se incluyen en la tabla: **aborda**, **abordaba**, **abordado**, **abordados**, **abordamos**, **abordan**, **abordar**, **abordaron**, **abordarse** y **abordó**.

Problemas que presenta el uso del Recuento

Si se desea usar el **Recuento** para evaluar la dificultad léxica de un pasaje cualquiera, surgen de inmediato una serie de dificultades. Para determinar la frecuencia de una palabra específica, por ejemplo, la palabra **estábamos**, es necesario consultar varias listas. El volumen I presenta las 10.000 unidades léxicas más frecuentes en 23 listas de grupos de unidades divididas en segmentos como sigue:

1. las 87 palabras más frecuentes entre las primeras 500;
2. las segundas 100 palabras más frecuentes entre las primeras 500.

Las primeras 500 palabras se presentan en grupos de 100 unidades. Las segundas 500 palabras se presentan también en grupos de 100. Los siguientes millares se presentan en grupos de 500 unidades, hasta llegar al sexto millar el cual se presenta como lista de 1000 unidades.

Las 20.000 formas de inflexión se presentan usando divisiones similares. En este caso, se utilizan 33 listas de grupos de formas empezando con las primeras 65 formas de inflexión de las primeras 500 y terminando con las 1000 formas del vigésimo millar.

Las listas de unidades léxicas –tal como se ha señalado– no contienen formas conjugadas. Si se busca el verbo **estar** en estas listas, la frecuencia que se da (como se vio en el caso del verbo abordar), no indica que sean de la misma frecuencia todas las formas del verbo. Es posible que algunas formas de un verbo sean frecuentes y otras no. La frecuencia del infinitivo (unidad léxica) poco indica. Por lo tanto, para investigar la verdadera frecuencia de la

forma **estábamos** es necesario examinar cada una de las 33 listas de inflexión para determinar cuál es su rango de frecuencia. En cambio, si se desea buscar la frecuencia de la forma **fácilmente**, esta tendrá que buscarse en las 23 listas de unidades léxicas, ya que como adverbio, no se encontrará entre las formas de inflexión.

Tabla 1

Palabra	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII	XIV
Abordar (8)					24	4		5	33			2	2	35
aborda						1				1				1
abordaba												2	2	2
abordado						2				2				2
abordados							2			2				2
abordamos						1				1				1
abordan						2				2				2
abordar						9	2			11				11
abordaron						4				4				4
abordarse						3				3				3
abordó						2				2				2

Rodríguez Bou, 1952, Vol. II. pp. 22-23

El primer volumen del Recuento consiste en dos grupos de listas complementarias que tienen que usarse juntas para poder determinar la frecuencia de las formas de inflexión y de las unidades que nunca se inflexionan. No es válido emplear una lista sin la otra. Sin embargo hay otra posible solución. Podrían utilizarse también los dos tomos del volumen II, en el cual se han combinado las listas de todas las palabras estudiadas (un total de 83.430 palabras). De esta lista podríamos determinar el índice de frecuencia para las dos formas ya citadas. Encontraríamos que **estábamos** es del rango 22 y **fácilmente** del 31. Es decir, **estábamos** se encuentra entre el segundo grupo de 500 palabras más frecuentes del segundo millar de las 20.000 formas de inflexión y **fácilmente** entre las 500 palabras más frecuentes del tercer millar de las 10.000 unidades léxicas. Una vez más tendríamos un rango de frecuencia sobre base distinta para las dos palabras de interés; resultado que no es de sencilla interpretación, y que además requiere que se utilicen dos tomos (cada uno de más de 500 páginas) para buscar cada palabra.

En resumen, el Recuento no contiene una lista práctica de palabras de alta frecuencia que combine unidades léxicas y formas de inflexión y que pueda usarse con facilidad en la docencia y en la investigación.

Proceso de derivar una lista combinada de unidades léxicas y formas de inflexión

Con el propósito de derivar una lista práctica de vocabulario, se llevó a cabo un proceso cuidadoso que dio como resultado una compilación combinada de 3619 elementos. A continuación se detallará el proceso que se usó al combinar las unidades léxicas y las formas de inflexión en el orden de frecuencia en que aparecen en el **Recuento**.

a. **Fusión de las listas de palabras más frecuentes que aparecen en el Recuento sin rango de frecuencia**

Al preparar la compilación combinada de elementos, se deseó retener las frecuencias exactas establecidas por Rodríguez Bou. El primer problema a solucionar se presentó cuando se intentó fundir la lista de 87 unidades léxicas más frecuentes con la lista de 65 formas de inflexión más frecuentes. Para estas dos listas, no se dan en el **Recuento** rangos de frecuencia, por considerarse las palabras de más alta frecuencia dentro del corpus estudiado. Para nuestra compilación se decidió combinar estas dos listas empleando las siguientes reglas:

1. Si un elemento aparece solamente en la lista de 65 formas de inflexión, se pone en la lista combinada.
2. Si un elemento aparece solamente en la lista de 87 unidades léxicas, pero también aparece en la lista de formas de inflexión, en cualquiera de las 33 listas, se acepta el rango de frecuencia de estas listas y no se pone el elemento en la lista combinada.
3. Si un elemento aparece en la lista de 65 formas de inflexión y en la lista de 87 unidades léxicas, se pone en la lista combinada.

La lista combinada que se obtuvo después de usar este proceso contiene un total de 105 elementos, los cuales se consideran los elementos de mayor frecuencia en la lista práctica.

b. **Fusión de las listas para las cuales se dan rangos de frecuencia**

Para derivar una lista breve que pudiera usarse fácilmente, se decidió incluir cuatro listas de frecuencia de unidades léxicas y cuatro listas de frecuencia de formas de inflexión. Se combinaron estas ocho listas empleando una computadora que se programó para que combinara las listas de acuerdo con ciertas reglas específicas.

El primer paso en el proceso fue introducir todos los elementos de cada una de las listas más la información sobre la frecuencia de cada elemento en los archivos de la computadora. Se introdujeron aproximadamente 2500 unidades léxicas y 2500 formas de inflexión. Las reglas que produjeron la lista combinada fueron las siguientes:

1. Si un elemento aparece en una sola lista, ya sea de unidades léxicas o de formas de inflexión, se pone en la lista combinada.
2. Si un elemento aparece en dos listas, se pone en la lista combinada. Debe utilizarse el rango de frecuencia que se consigna para ese elemento en la lista original de formas de inflexión.
3. Las divisiones entre grupos se identifican por el rango de frecuencia del último elemento del grupo de mayor frecuencia.

El resultado de este proceso fue una sola lista de 3514 elementos, divididos en cuatro grupos de frecuencia de acuerdo con los rangos que se establecen en el **Recuento** para las formas de inflexión. Combinada esta lista con la compilación de 105 elementos de mayor frecuencia, nuestra lista

práctica de frecuencias basada en el **Recuento** contiene un total de 3619 elementos divididos en cinco grupos. La lista de las 105 palabras de mayor frecuencia que forman el primer grupo de los cinco, se incluye en el **Apéndice**.

Una aplicación de la lista práctica

La lista práctica, derivada como se explicó anteriormente, se empleó en la derivación de una fórmula de lecturabilidad (Valdés, Barrera y Cárdenas, 1983). Esta fórmula se desarrolló para trabajar con materiales escritos en español que actualmente se utilizan en los programas bilingües de los Estados Unidos. La lista breve de vocabulario permite que el practicante y el investigador tengan acceso a un índice de dificultad de léxico de fácil uso.

Referencias bibliográficas

Buchanan, M.A. **A graded Spanish word book**. Toronto: American and Canadian Committees on Modern Languages, 1929.

Carroll, John B.: Davies, Peter; Richman, Barry. **American heritage word frequency book**. Boston: Houghton Mifflin, 1971.

Cartwright, C.W. A study, of the vocabularies of eleven Spanish grammars and fifteen Spanish reading texts. **Modern Language journal**. Vol. X, 1925.

Céspedes, A.T.R. **Investigaciones acerca de las palabras usadas en castellano**. Panama: Star and Herald, 1929.

Dale, Edgar y Chall, Jeanne S. A formula for predicting readability. **Educational Research Bulletin**, 27. 11-20, 28. 37-54.

García Hoz. E. **Vocabulario usual, común y fundamental**. Madrid: Instituto San José de Calasanz, 1953.

Harris, Albert J. y Jacobson, Milton D. **Basic elementary reading vocabularies**. New York: Macmillan. 1972.

Jamieson. E.I.A. Standardized vocabulary, for elementary modern Spanish. **Modern Language**, 1924.

Juilland, Alphonse y Chang Rodríguez, Eugenio. **Frequency dictionary, of Spanish words**. La Haya: Mouton, 1964.

Keniston, J. **A basic list of Spanish words and idioms**. 1920.

Kucera, Henry y Francis, Nelson. **Computation analysis of present day American English**. Providence. R.I.: Brown University Press, 1967.

Rodríguez Bou, Ismael. **Recuento de vocabulario español**. Río Piedras. PR.: Consejo Superior de Enseñanza de Puerto Rico, 1952.

Spanish Curricula Development Center. **Estimating difficulty of reading selections**. Miami Beach, Florida: Spanish Curricula Development Center, 1974.

Thorndike, Edward L. y Lorge, Irwing. **The teacher's word book of 30.000 words**. New York: Teacher's College Press. 1944.

Rodríguez Trujillo, Nelson. Listas de frecuencias de palabras: Una revisión de la literatura en español y de sus posibles usos en investigación. **Lectura y Vida**, 1980. Año 1, n°4 , 21-25.

Valdés, Guadalupe; Barrera, Rosalinda y Cárdenas, Manuel. Phase 1: Reading in bilingual education: Analyzing texts and children's comprehension. Final Report NIE-G-80-124, 1983.

Apéndice

Lista práctica de vocabulario
Las 105 palabras más frecuentes

a	están	papá
agua	este	para
ahora	flores	pero
al	fue	perro
allí	gato	por
aquí	grande	porque
así	gusta	que
bien	había	qué
bueno	hacer	se
casa	hay	si
como	he aquí	sí
con	iba	son
cuando	jugar	su
de	la	sus
del	las	tal vez
después	le	también
día	libro	tan
dice	lo	te
dijo	los	tengo
donde	maestra	tenía
dos	mamá	tiene
el	más	todos
él	me	tres
ella	mi	tu
en	mí	tú
entonces	mira	un
era	mucho	una
es	muy	uno
esa	nada	va
escuela	niño	vamos
ese	niños	ver
eso	no	vez
esta	nos	voy
está	otra	y
estaba	otro	yo