LÉXICO BÁSICO DEL ESPAÑOL DE CHILE: EL PROYECTO

M. NATALIA CASTILLO FADIĆ
Pontificia Universidad Católica de Chile

RESUMEN

Esta comunicación presenta la fundamentación teórica y metodológica y los resultados principales de nuestro proyecto Léxico Básico del Español de Chile, en el que, para determinar el léxico más usado en el país, se operó sobre un corpus de referencia propio de quinientas mil palabras en contexto, extraídas de cerca de mil libros y de más de tres mil ejemplares de diarios chilenos publicados por vez primera entre 1981 y 2006. Se trabajó con distintos géneros que constituyen "mundos" diferentes ─Narrativa, Ensayo, Drama, Técnico-Científico y Prensa─ y se establecieron comparaciones entre el léxico de cada uno de ellos, usando herramientas de estadística inferencial. El procesamiento del corpus ha utilizado herramientas de léxico-estadística y lingüística computacional y ha considerado aspectos semánticos y gramaticales. Los principales resultados de esta investigación son: a) La creación de un corpus de referencia etiquetado de 501211 palabras en contexto. b) La determinación del léxico básico del español de Chile. c) La creación de un diccionario no definitorio de 4832 entradas.

Palabras clave: léxico básico, español de Chile, estadística léxica, lingüística de corpus.

1. INTRODUCCIÓN

Esta macroinvestigación lexicoestadística tiene como antecedentes principales a Juilland & Chang-Rodríguez (1964) y a Morales (1986). Se centra en la determinación del léxico básico del español de Chile, conformado por los alrededor de 5.000 vocablos de mayor uso, a partir de un corpus de referencia etiquetado de más de 500.000 palabras, creado ex profeso para esta investigación (Castillo Fadić, 2012).

El estudio, sincrónico y sintópico, trabaja con distintos *mundos* o géneros –narrativa, ensayo, drama, técnico-científico y prensa– y establece comparaciones entre el léxico de cada uno. Se emplean herramientas de lingüística de corpus, lingüística computacional y estadística inferencial y se consideran aspectos lexicológicos, semánticos y gramaticales. Por tratarse de un estudio de réplica respecto de Morales (1986), tiene implicancias dialectales.

El objetivo general es obtener el léxico básico del español de Chile. Los específicos son: precisar los índices estadísticos de las unidades léxicas del corpus; determinar los vocablos de mayor frecuencia, dispersión y uso en el español de Chile; analizar los resultados, contrastando los índices de frecuencia de cada mundo entre sí y con los de dispersión y uso; se espera contribuir al estudio del español de Chile y sentar las bases para seleccionar las unidades léxicas pertinentes para elaborar diccionarios del español de Chile.

Los resultados de la investigación pueden aplicarse a la enseñanza-aprendizaje del español y pueden contribuir también al desarrollo de aplicaciones de lingüística computacional en esta lengua.

2. LÉXICOS BÁSICOS

2.1. Lexicoestadística

Los léxicos básicos corresponden al núcleo estadístico léxico de mayor uso en una comunidad y corresponden a los aproximadamente cinco mil vocablos de mayor uso; un rasgo característico de estos vocablos es su carácter atemático: mientras más arriba se sitúen en los rangos de uso, menor será su dependencia de variables externas. Para obtenerlos, se requieren investigaciones lexicoestadísticas.

La estadística léxica exige operar en el nivel del vocablo –unidad de léxico– y no meramente en el de la palabra –unidad de texto– (Müller, 1973, págs. 225-226); el vocablo, invariante, unidad de cita, "representa todo el paradigma de flexiones, derivaciones o conjugaciones que se forman a partir de una raíz o de un núcleo morfemático y que se reconocen como palabras" (Lara, 2006, pág. 138). El tipo, por su parte, se ubica en un nivel intermedio entre vocablo y palabra, en la medida en que agrupa las distintas ocurrencias de cada palabra bajo una sola etiqueta, permitiendo el conteo estadístico de las variantes.

Para determinar el léxico básico del español de Chile, operamos con tres índices estadísticos esenciales: *frecuencia absouta (f_i), dispersión (D)* y *uso (U)*. La f_i da cuenta del número de veces que aparece un elemento en un corpus o subcorpus (un mundo, por ejemplo); no usamos *frecuencias relativas*, puesto que un índice que se observa en relación con el tamaño de la muestra (fi/n) no nos resulta necesario para cotejar subcorpus de tamaño similar; por otra parte, el trabajar con frecuencias relativas supone la manipulación de cifras con un elevado número de decimales, lo que dificulta el posterior traslado de los índices estadísticos a publicaciones tales como diccionarios no definitorios organizados por rangos.

Puesto que las f_i solas no brindan toda la información necesaria para obtener un léxico básico, las ponderamos por la dispersión (D), que mide cuán equilibrada es la distribución de las frecuencias en los distintos subcorpus o mundos; la medida de dispersión compleja utilizada, se calcula mediante la fórmula mejorada de Juilland, Traversa, Beltramo, & Di Blasi (1973), donde n= número de mundos= 5, x_i= suma de las frecuencias reales de cada palabra o vocablo en cada mundo y T= la suma de las frecuencias totales de cada palabra o vocablo:

$$D = 1 - \frac{\sqrt{nx_i^2 - T^2}}{2T}$$

Esta fórmula arroja resultados que oscilan entre 0 (dispersión mínima) y 1 (dispersión máxima). Los vocablos con dispersión mínima son exclusivos de un mundo, independientemente de su f_i. Los que presentan dispersión máxima tienen idéntica f_i en todos los mundos.

El uso (U), en tanto, es el producto de la f_i por la D.

2.2. Desambiguación y segmentación

Para trascender el nivel de la palabra y pasar al del tipo y luego al del vocablo, en la *lematización*, es preciso desambiguar los homónimos –homófonos y homógrafos– (cf. Figura 1). Los lematizadores automáticos presentan la mayor efectividad en la desambiguación de la homonimia sintáctica y la menor en la desambiguación de la homonimia léxica, que suele requerir mayor intervención manual.

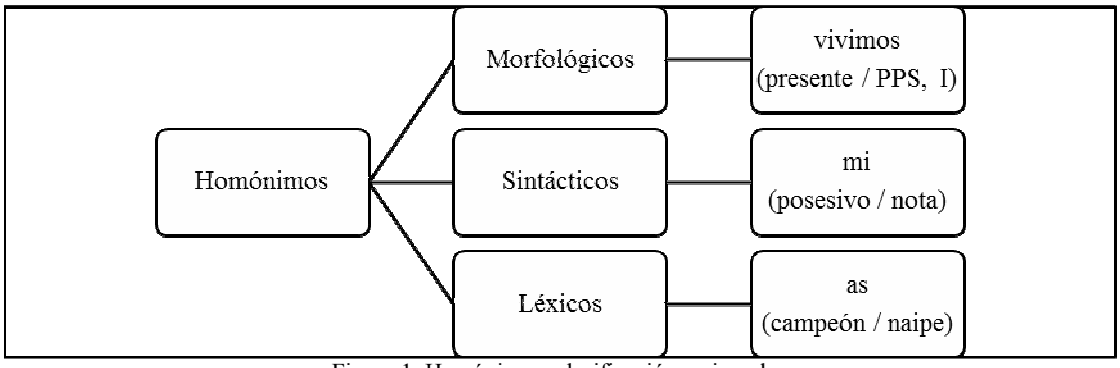

Figura 1. Homónimos: clasificación y ejemplos

La segmentación (cf. Figura 2) supone la demarcación de los límites entre las unidades léxicas susceptibles de conteo. La dificultad surge cuando estos límites no están suficientemente claros, ya sea porque no han sido descritos aún por la gramática o la lexicología, ya sea porque no hay acuerdo en su disposición, ya sea porque las reglas que subyacen a su delimitación no han sido adecuadamente formuladas.

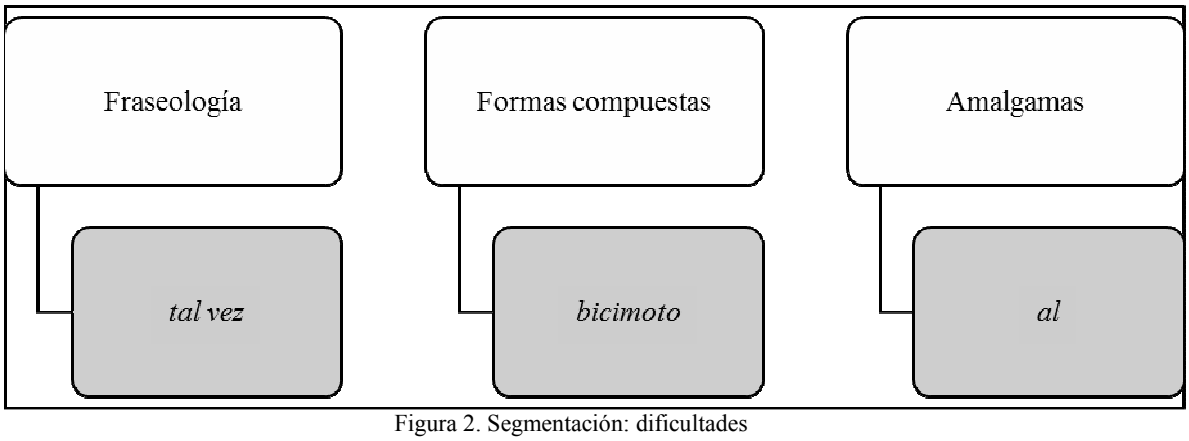

Figura 2. Segmentación: dificultades

Esta investigación procesa las unidades fraseológicas como tales, en la medida en que forman parte del diccionario interno de Freeling 2.2., base de LexBas 1.0, nuestro lematizador Respecto de las formas compuestas, se consideran como todos independientes y no se segmentan en sus formantes. Sobre las amalgamas, sólo <al> y <del> se segmentan en <a>+<el> y <de>+<el>, con lo que cada ocurrencia de una de estas amalgamas se cuenta como una ocurrencia de la preposición correspondiente y otra del artículo <el>.

3. METODOLOGÍA

3.1 Obtención y constitución del corpus

El corpus, de 501.211 palabras en contexto, se estructura en cinco mundos de más de 100.000 palabras (cf. Figura 3). Fue extraído de fuentes primarias escritas originalmente en español por autores chilenos, publicadas en español y por primera vez entre 1981 y 2006: libros clasificables como Drama, Narrativa, Ensayo y Técnico-Científico; diarios de circulación nacional, estratificados en las cuatro secciones generales a los medios estudiados, a saber, Opinión, Nacional, Deportes y Espectáculos. El tamaño del corpus, de más de 500.000 palabras, ha sido tradicional en las investigaciones de este tipo (cf. Juilland & Chang-Rodríguez, 1964; Juilland, Traversa,

Beltramo, & Di Blasi, 1973; Morales, 1986; Ávila, 1998; y Ávila, 1999), por cuanto permite obtener resultados representativos estadísticamente con mayor eficiencia y menor costo que los de tamaño superior.

Mundos	**Número de palabras**
Narrativa	100323
Drama	100220
Ensayo	100033
Técnico-Científico	100578
Prensa	100057
Total	**501211**

Figura 3. N° de palabras por mundo

Para aumentar la representatividad y la pureza del corpus, se procuró una estratificación a ultranza, que incluyera el máximo de fuentes por mundo, el mínimo de páginas por obra y el mínimo de oraciones por página. Esto significó la revisión manual del corpus, lo que permitió descartar obras repetidas bajo títulos diferentes, reediciones, traducciones, adaptaciones, citas a otros autores, etc.

El corpus se obtuvo tras la realización de cuatro muestreos estratificados por mundo (cf. Figura 4); el primero, aleatorio sin reemplazo, permitió determinar el número de oraciones requeridas para reunir 100.000 palabras por mundo, a través del cálculo de la media de palabras por oración; el segundo, también aleatorio sin reemplazo, sorteó el 15% del universo relativo de publicaciones chilenas, estratificadas por mundo, aumentado para corregir el error muestral de acuerdo con la desviación estándar observada tras el análisis de los datos obtenidos en el primer muestreo; el tercer y cuarto muestreo, por azar sistemático, permitieron identificar las páginas de las que se extraerían oraciones, así como las líneas de esas páginas. El universo sobre el cual se realizaron los sorteos se delimitó previamente según bases de datos bibliográficas previamente elaboradas a partir de datos proporcionados por la Biblioteca Nacional de Chile.

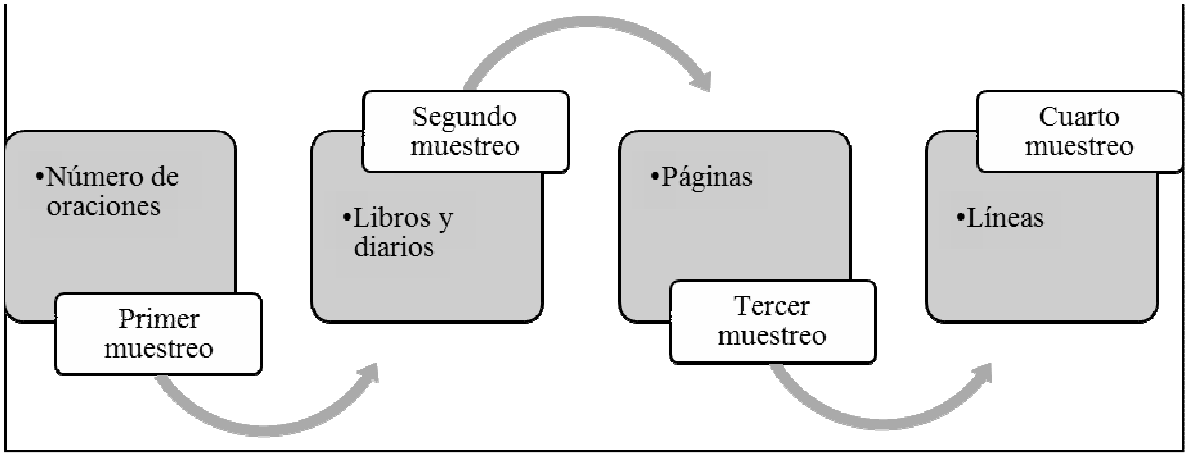

Figura 4. Muestreos

3.2 Procesamiento y análisis del corpus

Para la lematización y determinación de los índices de frecuencia, dispersión y uso, se usó LexBas 1.0, basado en FreeLing 2.2, programa creado especialmente para esta investigación, gracias al financiamiento del Dr. Humberto López Morales.

Las particularidades dialectales del corpus dificultaron el procesamiento automático y requirieron distintos grados de procesamiento manual. Para la desambiguación sintáctica y morfológica se usaron etiquetas EAGLES en formato *.XML. Para la desambiguación semántica, se crearon etiquetas propias, operacionales, también en *.XML. El voseo chileno, con formas conjugadas diferentes de las registradas por el diccionario interno de Freeling 2.2, exigió la creación de un diccionario de conjugaciones voseantes etiquetadas en EAGLES para alimentar el diccionario interno del programa y mejorar el reconocimiento automático de los verbos y, con ellos, el de las unidades léxicas factibles de ser reconocidas de acuerdo con la aplicación de reglas combinatorias.

4. RESULTADOS

4.1 Corpus Básico del Español de Chile

Se obtuvo un corpus etiquetado de 501.211 palabras, que puede servir de base para múltiples investigaciones sobre el español de Chile.

4.2 Léxico básico del español de Chile

El rango de corte se determinó con base en criterios estadísticos y es similar al de autores previos; Juilland & Chang-Rodríguez (1964) lo fijan en el índice de uso >=3,08, mientras que Morales (1986, pág. 29), que desglosa su diccionario en dos partes, lo establece en>=3,55 para la primera y entre 3,55 y 3,08 para la segunda.

Como Juilland & Chang-Rodríguez (1964), Juilland, Traversa, Beltramo, & Di Blasi (1973), Morales (1986) y Ávila (1999), estimamos que el número de unidades consideradas básicas debe estar alrededor de 5000, ya que la representatividad que alcanza la muestra en este rango, en relación con el total del corpus, tiende a ser >=90%. Como Morales (1986), consideramos que las unidades deben presentarse en a lo menos tres mundos. Por último, como Juilland & Chang-Rodríguez (1964) y Morales (1986), consideramos que la f_i mínima de los vocablos debe ser >=4, para evitar que una unidad que alcance el índice de U requerido por su alta dispersión pueda ser considerada básica pese a una f_i demasiado baja; no obstante, este último criterio resultó redundante, puesto que ninguna de las unidades de alto uso presentó una f_i inferior a cuatro.

Aplicando los parámetros mencionados, ordenamos los vocablos por U total; observamos que el U en el rango 5000 era =2,97; seleccionamos, los vocablos con uso >= 2,97, más allá del rango 5000 y revisamos que estuvieran presentes en al menos tres mundos. Al aplicar este filtro, los vocablos con uso>= 2,97 salieron del léxico básico, pues sólo tenían presencia en dos mundos. Con ello, el corte quedó establecido en el uso 3,0.

Así, el léxico básico del español de Chile quedó constituido por 4832 vocablos de U >= 3,0, presencia en al menos tres mundos y representatividad acumulada (R)= 91,6708% (cf. Figura 5).

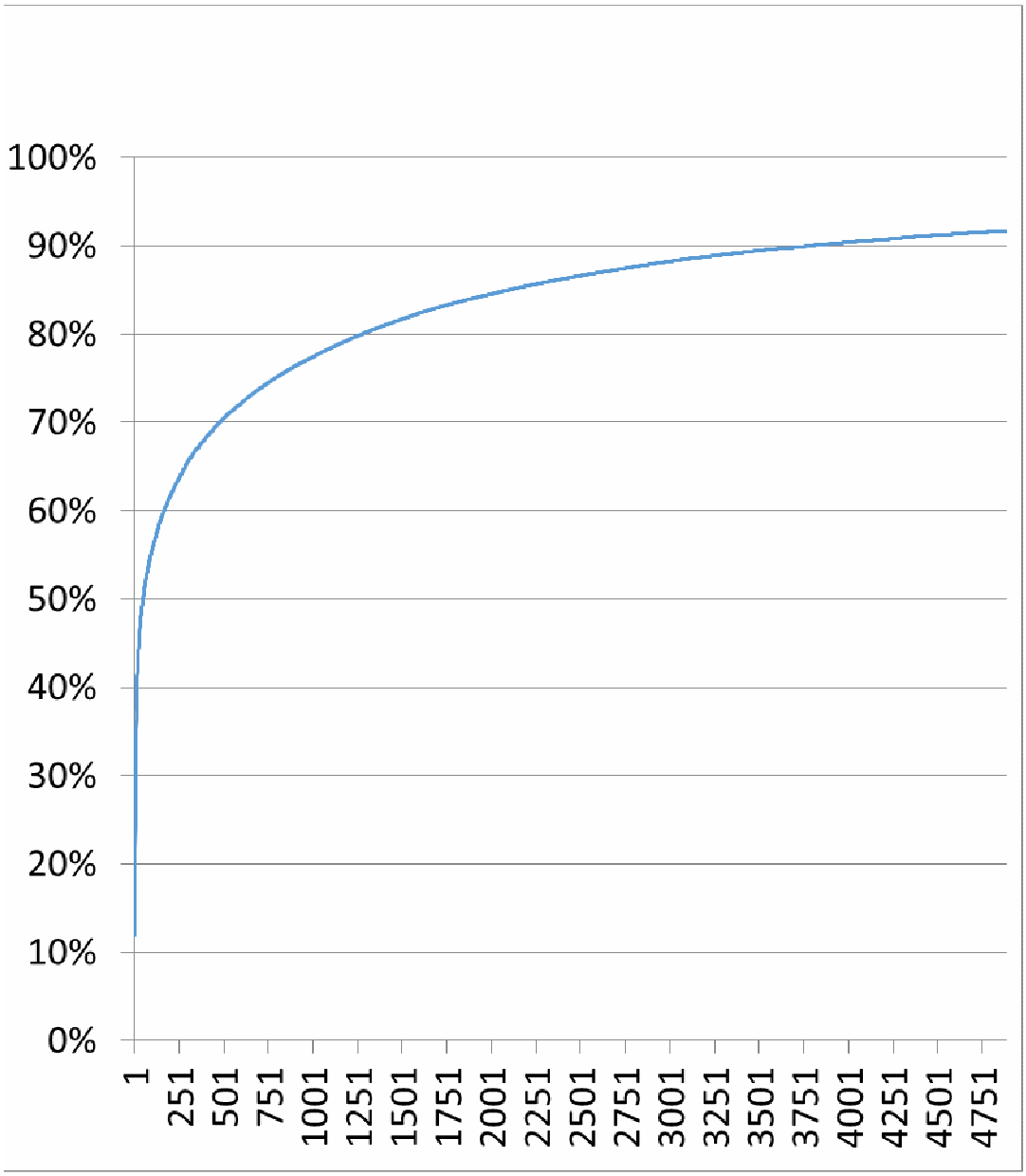

Figura 5. Vocablos de mayor uso: representatividad acumulada

4.2 Léxico Básico del Español de Chile: el diccionario

Estos vocablos con sus respectivas variantes se organizaron en un diccionario no definitorio, ordenado alfabéticamente, cuya microestructura se aprecia en la Figura 6.

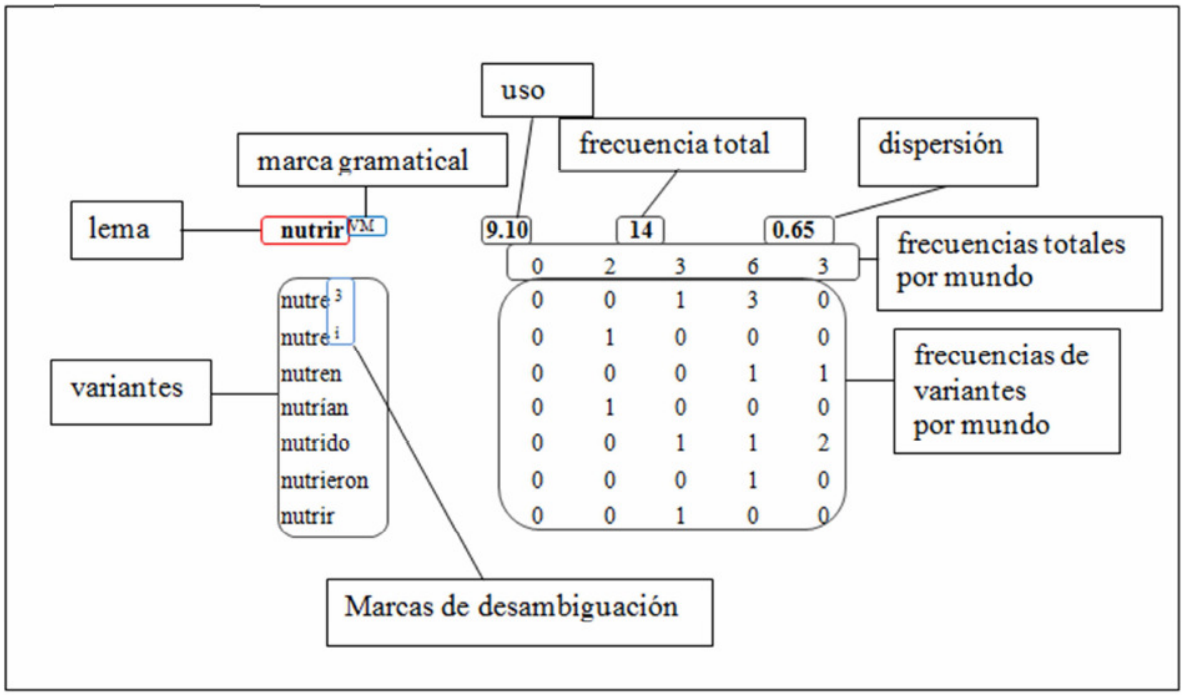

Figura 6. Léxico Básico del Español de Chile: microestructura del diccionario

5. CONCLUSIONES

Resulta fundamental mejorar las herramientas de análisis lingüístico en español, especialmente en lo relativo a particularidades dialectales, por medio de la clara descripción de las reglas subyacentes, requisito necesario para aplicaciones de lingüística computacional. Esto permitirá procesar con mayor rapidez y eficiencia corpus léxicos diversos, especialmente los más alejados de la norma estándar, que hoy presentan grandes dificultades en el análisis automático.

Una vez creado el *Corpus Básico del Español de Chile*, definido el léxico básico del español de Chile y generado un diccionario no definitorio homónimo, invitamos a nuestros lingüistas a emprender investigaciones similares en otros países del mundo hispánico, para generar léxicos básicos susceptibles de comparación y corpus básicos que faciliten la realización de diversas investigaciones lexicoestadísticas más allá de las fronteras nacionales.

REFERENCIAS BIBLIOGRÁFICAS

Ávila, A. M. (1998). *Elaboración, anotación y análisis del corpus oral del Proyecto V.U.M.* Málaga: Universidad de Málaga, Departamento de Filología Griega, Estudios Árabes y Traducción e Interpretación, Área de Lingüística General.

Ávila Muñoz, A. M. 1999. *Léxico de frecuencia del español hablado en la ciudad de Málaga*. Málaga: Universidad de Málaga.

Castillo Fadić, M. N. 2012. *Corpus Básico del Español de Chile*.

Juilland, A., & Chang-Rodríguez, E. 1964. *Frequency Dictionary of Spanish Words, The Romance Languages and their Structures*, First Series SI. La Haya: Mouton.

Juilland, A., Traversa, V., Beltramo, A., & Di Blasi, S. 1973. *Frequency Dictionary of Italian Words*. The Hague- Paris: Mouton.

Morales, A. 1986. *Léxico básico del Español de Puerto Rico*. San José de Puerto Rico: Academia Puertorriqueña de la Lengua, Editorial La Muralla, S.A.