10.18800/lexis.202002.004
Artículo

Corpus Básico del Español de Chile ©: metodología de procesamiento y análisis
Corpus Básico del Español de Chile ©: Processing and Analysis Methodology

María Natalia Castillo Fadić[1]

[1]Pontificia Universidad Católica de Chile - Chile

Resumen

Este artículo revisa la metodología empleada para procesar y analizar el *Corpus Básico del Español de Chile* ©. Se centra en los criterios para ordenar los materiales, segmentar y lematizar, mediante un programa computacional elaborado *ad hoc* para esta investigación y especialmente preparado para procesar y analizar corpus de español de Chile.

Palabras claves*:* lingüística de corpus, lingüística computacional, lexicología, estadística léxica, español de Chile

Abstract

This article reviews the methodology to process and analyze the *Basic Corpus of Chilean Spanish* © (in Spanish *Corpus Básico del Español de Chile* ©). It focuses on the criteria for ordering linguistic materials, segmentation and lemmatization, using a computer program developed ad hoc for this research and specially prepared to process and analyze Chilean Spanish.

Keywords: corpus linguistics, computational linguistics, lexicology, lexical statistics, Chilean Spanish

I. Introducción

El presente artículo se centra en los fundamentos metodológicos del procesamiento y análisis del *Corpus Básico del Español de Chile*) ©Castillo Fadić 2012a). Este corpus, de algo más de 500.000 palabras en contexto, recibe este nombre porque a partir de él se elaboró el *Léxico Básico del Español de Chile* (Castillo Fadić 2020a). Fue obtenido mediante una serie de muestreos estratificados sobre publicaciones de autores chilenos de las categorías Drama, Narrativa, Ensayo, Técnico-Científico y Prensa, de un periodo de 26 años que abarca los siglos XX y XXI (véase Castillo Fadić 2020b). Los fundamentos teóricos de la investigación que dio origen a este corpus se encuentran en Castillo Fadić (2020a). Damos cuenta, a continuación, de los criterios para ordenar los materiales, segmentar y lematizar, empleando un programa elaborado *ad hoc* para esta investigación y especialmente preparado para procesar y analizar corpus de español de Chile.

II. Procesamiento del corpus: pasos y métodos

El procesamiento del corpus revistió muchísima complejidad y exigió el empleo de herramientas informáticas de alta especificidad. Para ello, utilizamos un programa computacional elaborado

ad hoc para esta investigación[1], que permite procesar de manera adecuada los materiales obtenidos tras los diversos muestreos, tanto en lo correspondiente a la lematización automática, como en lo relativo a la obtención de la frecuencia de cada unidad y a su dispersión, producto de las cuales se determina el uso. El programa computacional, que recibe el nombre de LexBas 1.0 y que se apoya en FreeLing 2.2[2], inició su marcha blanca a mediados de enero de 2011. Durante todo el primer semestre de ese año, se realizaron múltiples pruebas para perfeccionarlo en virtud de los objetivos de la investigación y se establecieron codificaciones para resolver casos que resultaban ambiguos para el procesamiento automático. "Los datos fueron preparados para el tratamiento informático y las fórmulas utilizadas fueron convertidas, actualizando los postulados iniciales y las fórmulas de Juilland y Chang-Rodríguez (1964), Juilland, Traversa, Beltramo, y Di Blasi (1973) y también los más actualizados de Morales (1986)" (Humberto López Morales 2020: en prensa).

Dado que para un procesamiento automático efectivo del lenguaje natural es fundamental la labor interdisciplinaria (véase Alvar Ezquerra, Blanco Rodríguez y Pérez Lagos 1994, Alvar Ezquerra y Corpas Pastor 1994, y Lavid 2005), trabajamos en estrecha colaboración con el informático encargado de LexBas 1.0, para corregir detalles y lograr un eficaz procesamiento automático que permitiera obtener el listado de frecuencia, dispersión y uso, que era lo central de nuestra investigación (véase Castillo

[1] Agradecemos muy especialmente al Dr. Humberto López Morales por haber encargado y financiado este programa específicamente para esta investigación. El programa fue diseñado por un matemático e informático de la Universidad de Salamanca, quien aportó las herramientas informáticas necesarias para el procesamiento lingüístico automático y el análisis estadístico.

[2] FreeLing 2.2 era, al momento de la creación de LexBas 1.0, la última versión probada de un programa informático libre de análisis lingüístico, desarrollado por el Centre de Tecnologies i Aplicacions del Llenguatge i la Parla (TALP) de la Universitat Politècnica de Catalunya (UPC). Si bien opera básicamente a partir de reglas combinatorias, contiene diccionarios de distintas lenguas, incluido uno de español. Este diccionario presta especial utilidad en el procesamiento de formas conjugadas de verbos, agrupaciones y nombres propios.

Fadić 2015a). En conjunto, revisamos el funcionamiento de distintos modos de consulta del programa[3] y detectamos falencias en el análisis automático que estimamos necesario corregir, en virtud de los objetivos de la investigación; algunas de estas inexactitudes derivaban de la falta de acuerdo que existe aún en el ámbito teórico respecto de la clasificación de ciertas unidades, especialmente en las llamadas *palabras gramaticales*; otras tenían relación con limitaciones propias del procesamiento automático, que pierde precisión ante construcciones inhabituales (Almela, Cantos, Sánchez, Sarmiento y Almela 2005); otras muchas se desprendían de las necesidades particulares de esta investigación, centrada en un corpus extraído de fuentes chilenas, donde la variedad idiomática representada exhibe múltiples peculiaridades que la alejan del estándar panhispánico (veáse, por ejemplo, Rona 1962; Zamora Munné y Guitart 1982; y Moreno Fernández 2016) y que, por lo mismo, no están contempladas por el diccionario interno de FreeLing 2.2. Respecto de las primeras imperfecciones, que consideramos inevitables en un trabajo de esta naturaleza, intentamos establecer ciertas precisiones teóricas; sobre las segundas, vimos la necesidad de incorporar al programa mecanismos de edición manual, para modificar el análisis realizado por LexBas 1.0, cuando resultara necesario; como es lógico, la revisión manual acabada de todas y cada una de las oraciones suponía un esfuerzo de largo aliento, por lo que decidimos comenzar por la revisión de las oraciones correspondientes a vocablos contenidos en el recuento inicial de los 5.000 más usados, especialmente cuando el índice de certeza de análisis del software era distinto de 100%. Tomamos esta decisión basados en que los errores de análisis que incrementen o disminuyan el conteo de vocablos de muy bajo uso no

[3] Para ello, nos trasladamos a España los meses de julio y agosto de 2011, gracias al patrocinio de la Facultad de Letras de la Pontificia Universidad Católica de Chile.

tendrían prácticamente relevancia desde el punto de vista estadístico. Por el contrario, las imprecisiones en el análisis de vocablos de alto uso podrían alterar su rango o, incluso, podrían incidir en que quedaran dentro o no de las 100 palabras de mayor uso, o incluso de las 5000, sobre todo si estaban en posiciones de corte. Por ejemplo, si observamos el caso de los nombres propios, que solo se detectaron para ser eliminados, puesto que no tienen interés en un léxico básico, la aplicación encontró 51 645 formas distintas; si de estas hubiera 50 ocurrencias de una palabra analizadas incorrectamente, es decir, clasificadas como nombres propios siendo nombres comunes, o a la inversa, estaríamos hablando de una desviación de un 1 ‰ (0, 1%), lo que, desde el punto de vista estadístico, sería un margen irrelevante. Por último, respecto de las últimas imperfecciones, su solución pasa no solo por la incorporación de nuevas etiquetas y el incremento del procesamiento manual, sino que exige el mejoramiento del diccionario interno de FreeLing 2.2; vistas nuestras particularidades regionales (algunas de las cuales pueden revisarse en el apartado 2.4.3), consideramos que de todas ellas la más -susceptible de análisis automático por medio de un diccionario[4] y la de mayor impacto estadístico era el voseo verbal; por ello, creamos un diccionario de formas conjugadas voseantes, etiquetadas en EAGLES (Expert Advisory Group on Language Engineering Standards), a fin de alimentar el diccionario interno de FreeLing 2.2.

Los pasos que seguimos para procesar de manera automática el corpus fueron los siguientes:

[4] El procesamiento automático puede basarse también en reglas combinatorias.

2.1. Orden de los materiales

Los materiales recogidos se ordenaron inicialmente en una hoja de cálculo por cada mundo, como se observa en la Figura 1, donde registramos el número de sistema (columna A) con que cada obra se identificaba en la bibliografía proporcionada por la Biblioteca Nacional de Chile (véase Castillo Fadić 2020b), el nombre del autor (B), el título de la obra (C), los detalles de impresión (lugar, editorial y año) (D), la descripción (formato y número de páginas) (E) y la clasificación Dewey (F). A estas informaciones, agregamos la interpretación de la clasificación Dewey[5], con indicación de género o materia (G-H), y el año de publicación (I), en columna separada.

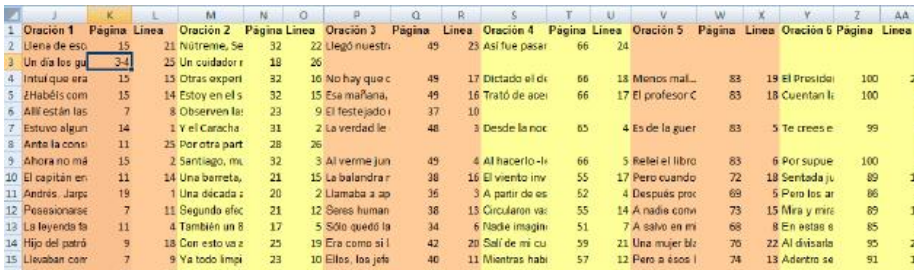
Acto seguido (véase Figura 2), listamos alternadamente las oraciones que extraeríamos de cada una de las obras (J, M, P, S, V, Y, etc.), indicando la página de la que se sacaba la cita (K, N, Q, T, W, Z, etc.) y la línea sorteada (L, O, R, U, X, AA, etc.). Puesto que los sorteos fueron por azar sistemático, los números de página se ingresaron en las planillas de antemano; respecto de los números de línea, se ingresó el primero sorteado y luego se completó con números consecutivos; las columnas correspondientes a las oraciones fueron las últimas en llenarse.

De este modo, pudimos mantener un registro de las oraciones extraídas de cada fuente, con sus referencias completas, lo que permite citarlas si es preciso.

[5] Obtenida caso a caso, mediante consulta al tomo correspondiente de Dewey (1989) en la Biblioteca de Humanidades de la Pontificia Universidad Católica de Chile.

	A	B	C	D	E	F	G	H	I
1	Nº sistema	Autor	Título	Impresi	Descripc	Clasificación	DEWEY		Año
2	000248264	Cury Urzúa, Enrique, 1933-	Cuestiones penales / Enrique Cury Urzúa.	[Santiag	29 p. ;	24 345 C982	Leyes		1988
3	000290546	Contreras B., Marta	Griselda Gambaro : teatro de la descomposición / Marta Contreras.	[Concep	188 p. ;	2 A862 G138	Literatura hispán	Drama	1994
4	000250456	Ghisolfo Araya, Francisco, 1929-	El poder naval frente al derecho del mar / Francisco Ghisolfo Araya.	[Chile] :	16 p. ;	22 359 G427	Administración pública		1982
5	000269163	Gómez, Galo, 1926-	Origen y desarrollo de la Universidad en Chile / Galo Gómez Oyarzún.	México	109 p. ;	2 378.83 G633	Educación		1992
6	000276730	Riego, Cristian	La eficacia del proceso penal frente a los delitos de robo / Cristian Riego.	Santiag	39 h. ;	28 345.8305 R554	Leyes		1993
7	000323660	Fanna Vicuña, María del Carmen, 1945-	Marco constitucional de la representación democrática / María del Carmen Fan	[Santiag	37 p. ;	27 323.042 F225	Ciencia política		1984
8	000272610	Pujadas H., Gabriel de	Educación : desafíos de hoy y mañana : institucionalidad escolar, curriculum,	Santiag	120 p. ;	2 370.983 P979	Educación		1993
9	000254787	Carbone Campo, Leonardo	Fumigantes.	Santiag	27 p. ;	27 668.651 C264	Ingeniería química		1988
10	000253041	Carrasco Cortés, Sergio	Metodología y técnica del handbol / Sergio Carrasco Cortés.	[Santiag	81 p. ;	31 796.31 C313	Recreación y artes escénicas		1982
11	000018561	Gubbins B., Víctor	Formación y ejercicio profesional del arquitecto en Chile / Víctor Gubbins B.	Santiag	37 p. ;	27 720.2383 G921	Arquitectura		1987
12	000256843	Israel, Alberto	Publicidad técnica & práctica / Alberto Israel.	Santiag	554 p. ;	2 659.1 I85	Administración y servicios auxi		1989
13	000326481	Leporati, Ariel, 1925-	Psicología del deporte / Ariel Leporati P.	[Santiag	184 p. ;	2 796.01 L598	Recreación y artes escénicas		1996
14	000286262	Lynch Gaete, Patricio	Liderazgo : cuatro perspectivas para una dirección eficaz / Patricio Lynch Gaet	Concep	175 p. ;	2 658.4092 L987	Administración y servicios auxi		1993
15	000597858	Mellado, Justo Pastor, 1949-	Estética de la dificultad.	Buenos	41 p. ;	21 759.983 M524	Pintura y pinturas		1993
16	000547057	Núñez Atencio, Lautaro	Ocupación paleoindio en Quereo : reconstrucción multidisciplinaria en el territo	Antofag	131 p.,	17 930.1	Historia del mundo antiguo		1983
17	000278394	Solar Madariaga, Iván, 1943-	Control automático de procesos químicos / Iván Solar Madariaga, Ricardo Pére	[Santiag	225 p. ;	2 660.281 S684	Ingeniería química		1993
18	000020306	Vergos Ramírez, Jaime	Perfiles diferenciales de los programas infantiles chilenos / Jaime Vergos R.	Santiag	27 p. ;	27 791.4550983 V496	Recreación y artes escénicas		1986
19	000303081	Lagos, Luis Felipe	Mecanismos de transmisión bajo tipos de cambio fijo y flexible / Luis Felipe La	Santiag	21 p. ;	26 332.45 L177	Economía		1981
20	000251208	Acevedo F., Cecilia, 1943-	Planificación curricular en el nivel de transición de la educación parvularia / M.	Santiag	64 p. ;	dia 372.190983 A173	Educación		1982
21	000033814	Aguilera, Pablo, 1954-	En la frontera vida/muerte : problemas bioéticos / Pablo Aguilera.	Santiag	208 p. ;	1 174.9574 A283	Ética (moral filos	Economía y é	1991
22	000239318	Ahumada Acevedo, Pedro	Principios y procedimientos de evaluación educacional / Pedro Ahumada Acev	Valparai	107 p. ;	d 371.26 A287	Educación		1983
23	000595072	Ahumada Acevedo, Pedro	La evaluación en una concepción de aprendizaje significativo / Pedro Ahumada	Valparai	126 p. ;	2 379.15 A287	Educación		2002
24	000270213	Alegría Ch., Arturo	Huelga : enfoques teóricos y efectos económicos de distintas regulaciones / A	Santiag	67 p. ;	dia 331.8920983 A366	Economía		1992
25	000233950	Aliaga, Hamilton	Evaluación y aprendizaje de una experiencia de asesoría y capacitación a coo	[Santiag	204 p. ;	m 334.6830983 E92	Economía		1986
26	000027951	Allendes, Cecilia	Tormenta en familia / Cecilia Allendes, Ana María Foxley, Luisa Ulibarri.	[Santiag	93 p. ;	18 306.810983 A425	Ciencias sociale	Cultura e insti	1989
27	000228864	Alvarez Martín, Francisco	Realidad rural y educación / Fco. Alvarez, Cecilia Cardemil, Luis Bustos.	Santiag	161 p. ;	2 371.1460983 A473	Educación		1985
28	000252637	Andreani, Ricardo	Capacitación para jóvenes desocupados : formación personal y social / Ricard	Santiag	63 p. ;	24 365.2350983 A556	Ciencias sociale	Grupos social	1984
29	000414414	Aninat, Andrés, 1916-1992	La fuerza del amor : líneas de fuerza de la moral cristiana / Andrés Aninat.	[Santiac	109 p. ;	2 230 A587	Teología cristiana		1997

Figura 1. Listado de obras y sus referencias. Ejemplo de base Técnico-científica



	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA
1	Oración 1	Página	Línea	Oración 2	Página	Línea	Oración 3	Página	Línea	Oración 4	Página	Línea	Oración 5	Página	Línea	Oración 6	Página	Línea
2	Llena de eso	15	21	Nútreme, Se	32	22	Llegó nuestra	49	23	Así fue pasar	66	24						
3	Un día los gu	34	25	Un cuidador r	18	26												
4	Intuí que era	15	15	Otras exporta	32	16	No hay que c	49	17	Dictado el de	66	18	Menos mal...	83	19	El Presidei	100	20
5	¿Habéis com	15	14	Estoy en el s	32	15	Esa mañana,	49	16	Trató de acer	66	17	El profesor C	83	18	Cuentan la	100	1
6	Allí estaban las	7	8	Observen las	23	9	El festejado r	37	10									
7	Estuvo algun	14	1	Y el Carocha	31	2	La verdad le	48	3	Desde la noc	65	4	Es de la guer	83	5	Te crees e	99	6
8	Ante la consi	11	25	Por otra part	28	26												
9	Ahora no má	15	2	Santiago, mi	32	3	Al verme jun	49	4	Al hacerlo -le	66	5	Relei el libro	83	6	Por supue	100	7
10	El capitán en	11	14	Una barreta,	21	15	La balandra r	38	16	El viento inv	55	17	Pero cuando	72	18	Sentada ju	89	19
11	Andrés, Jarpa	19	1	Una década a	20	2	Llamaba a ap	36	3	A partir de es	52	4	Después proc	69	5	Pero los ar	86	5
12	Posesionarse	7	11	Segundo efec	21	12	Seres human	38	13	Circularon vas	55	14	A nadie conv	73	15	Mira y mira	89	16
13	La leyenda ta	11	4	También un 8	17	5	Sólo quedó la	34	6	Nadie imagin	51	7	A salvo en mi	68	8	En estas e	85	9
14	Hijo del patró	9	18	Con esto va a	25	19	Era como si l	42	20	Salí de mi cu	59	21	Una mujer blo	76	22	Al divisarla	95	23
15	Llevaban coir	7	9	Ya todo limpi	23	10	Ellos, los jefe	40	11	Mientras habi	57	12	Pero a ésos l	74	13	Adentro se	91	14
16	Pero a la tarc	15	11	Rómulo habi	32	12	Al recibir aqu	49	13									
17	No se enojat	19	5	Y allí leí lo siy	24	6	Los curas del	41	8	Temiendo es	58	9	Esté enferm	75	10	Por otra pa	92	11
18	En una estac	9	1	En la tierra e	25	2												
19	Volvió a senti	13	9	Aunque creo	30	10	Es más, incl	47	11	Liam Elgee/M	64	12	Se habrá enti	81	13	No me gus	98	14
20	Entonces, él	13	13	Respiró el pa	21	14	Era verdader	38	15	Pero lo más	55	16						
21	Solitaria, cor	9	4	Experimentó	22	5	Se desplaza l	39	6	Guardaron la	56	7	Recibe hono	73	8	El nombra	90	9
22	Se apartó de	13	17	Ella dice que	25	18	Nunca había	42	1	Hijo de un ex	59	2	El predio dab	76	3	En el local	93	4
23	Sus padres h	19	17	Al pasar al in	25	18	David quería	42	19	Después de	59	20	La fecha del	76	21	Bueno esc	93	22
24	Su torre y to	7	10	En sus alrede	19	11	Estaba junto	36	12	Nada de mal	53	13	Las dos estat	71	14	La voluntá	87	15
25	Se hundió en	13	10	Mañana iré a	23	11	No obstante,	40	1	La casona qu	57	2	Antonio Figu	74	3	Las señorit	91	4
26	En efecto, en	13	4	Como es de s	30	5												
27	Y lo hace sin	15	20	Crucé con de	23	21	Retrocedía, p	40	22	Oí un grito en	57	23	El mayor, sir	74	24	Una atmós	91	25

Figura 2. Orden de las oraciones en las planillas. Ejemplo de base Narrativa

Una vez finalizado el proceso de recolección de corpus, el material de cada uno de los mundos se reunió en un archivo Excel único, con una hoja para cada mundo, respetando el orden y estructura de las bases iniciales y procurando ingresar solo un valor por campo, de modo de permitir el procesamiento automático. Este archivo fue ingresado sin cambios a LexBas 1.0, que lo procesó de manera automática distinguiendo mundos, referencias bibliográficas y oraciones.

2.2. Segmentación

2.2.1. Palabras y vocablos

Esta investigación considera como unidad mínima la palabra. Una de las ventajas de esta opción es que el hecho de que gráficamente esté separada de otras unidades similares por espacios en blanco permite una identificación automática eficaz.

No obstante, recurrimos también a otras unidades que han exigido procesamientos de mayor complejidad: los *vocablos*, que permiten agrupar distintas realizaciones bajo un único lema (véase Lyons 1997, Lara 2006), y las *agrupaciones de palabras*, que engloban todas las unidades fraseológicas, vale decir, todas aquellas formas que, estando compuestas por más de una palabra, funcionan como un todo (Corpas Pastor 1997).

2.2.2. Agrupaciones frecuentes

Se confeccionó un listado de agrupaciones frecuentes de palabras, para evitar que el programa lematizara separadamente pa-

labras que, funcionalmente, constituyen una sola unidad; consideramos aquí tanto formas compuestas de los verbos como locuciones de distintos tipos (véase Figura 3). No obstante, dejamos fuera todo aquello que no fuera susceptible de sistematización, por cuanto no existía la posibilidad de un procesamiento automático. Este es el caso de perífrasis verbales y de formas cuyo análisis está sujeto a discusión.

#	Text	Lemma	Pos	Eagle	Drama	Narrativa	Ensayo	Técnico	Prensa	Total	Dispersion	Use total
1	no_obstante	no_obstante	CC	CC	1	3	12	17	11	44	0.66	29.04

1 result

Figura 3. Ejemplo de agrupación: <no_obstante>

Respecto de los verbos, se tratan como una unidad los tiempos verbales compuestos y las construcciones en forma pasiva. Esto es: <he comido>, que responde a la estructura <*haber* conjugado+ participio>, se analiza como una aparición del verbo en participio <comer> y no como una ocurrencia del verbo <haber>; la pasiva <ha sido olvidado>, que obedece a la estructura <*ser* conjugado+ participio>, se considera como ocurrencia del verbo en participio <olvidar>. En este último caso, el programa LexBas 1.0 analiza la pasiva como unidad, aun cuando haya palabras intercaladas, aunque ciertas ocurrencias deben tratarse manualmente; así, <es frecuentemente observado> se analiza como <es observado> + <frecuentemente> y se lematiza bajo <observar> y <frecuentemente> respectivamente.

No se unifican formas verbales de mayor complejidad, como perífrasis del tipo <ir buscando>, pues la falta de consenso en el análisis y la alta variabilidad impiden el tratamiento automático, como bien lo indica Morales (1986: 21):

> Si bien en una secuencia como *ha comido* o *había ido* nadie duda de su funcionamiento como una sola unidad, hay un proceso gradual de independencia en *está cosiendo*, *es castigado*, *ha de comenzar*, *tiene señalado*, etc., que ya no la ofrecen con tanta claridad. Para mayor complejidad, todas estas formas, incluso las formas compuestas tradicionales, permiten la intercalación de un elemento adverbial, con lo cual se viola uno de los requisitos que exigen las formas para su consideración de independientes.

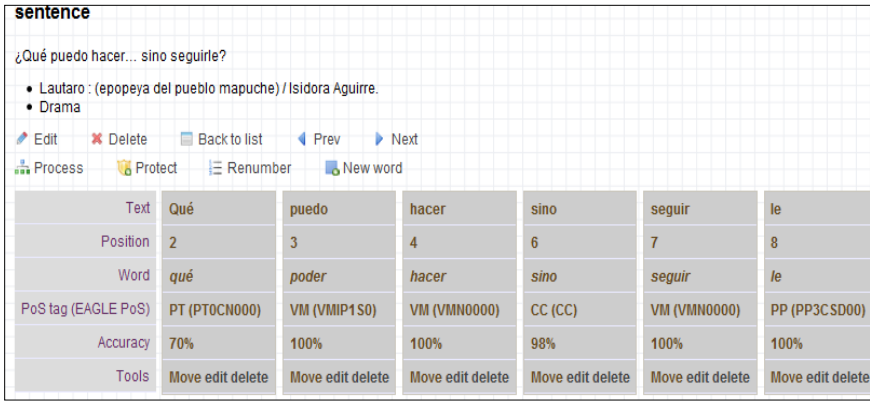
Respecto de las locuciones y demás unidades fraseológicas, se distinguen todas aquellas que forman parte del diccionario interno de FreeLing 2.2, lematizador de base del programa LexBas 1.0. A estas se han agregado manualmente otras combinaciones frecuentes que admiten sistematización, básicamente porque no permiten elementos intercalados o porque presentan una forma estable. En este sentido, adoptamos una solución distinta de la de Morales (1986: 22), quien no tiene en cuenta lo que llama "«unidad de función», por lo menos en un primer nivel de análisis", y más parecida a la de Juilland y Chang-Rodríguez (1964), aun cuando intentamos aquí superar algunas de las falencias que presenta la mencionada obra en el procesamiento de unidades complejas, por medio de la combinación entre el tratamiento automático y el manual.

2.2.3. Amalgamas

En el caso de las amalgamas, procuramos seguir el tratamiento de Morales (1986: 20), a saber: "El español presenta dos casos típicos de amalgama que son los artículos contractos, *al* y *del*, y

en cuanto a su tratamiento todos los estudiosos están de acuerdo en el conteo separado de las unidades que representan (*a el* y *de el*)". De este modo, el programa LexBas 1.0 separa *al* y *del* en sus formantes y lematiza las ocurrencias de dichos formantes bajo los vocablos *a, el* y *de* respectivamente.

Respecto de los verbos con pronombres enclíticos, como Juilland y Chang-Rodríguez (1964) y Morales (1986: 20-21), tratamos separadamente los verbos y los pronombres (véase-Figura 4).



Text	Qué	puedo	hacer	sino	seguir	le
Position	2	3	4	6	7	8
Word	*qué*	*poder*	*hacer*	*sino*	*seguir*	*le*
PoS tag (EAGLE PoS)	PT (PT0CN000)	VM (VMIP1S0)	VM (VMN0000)	CC (CC)	VM (VMN0000)	PP (PP3CSD00)
Accuracy	70%	100%	100%	98%	100%	100%
Tools	Move edit delete	Move edit delete	Move edit delete	Move edit delete	Move edit delete	Move edit delete

Figura 4. Ejemplo de análisis de verbo más pronombre enclítico

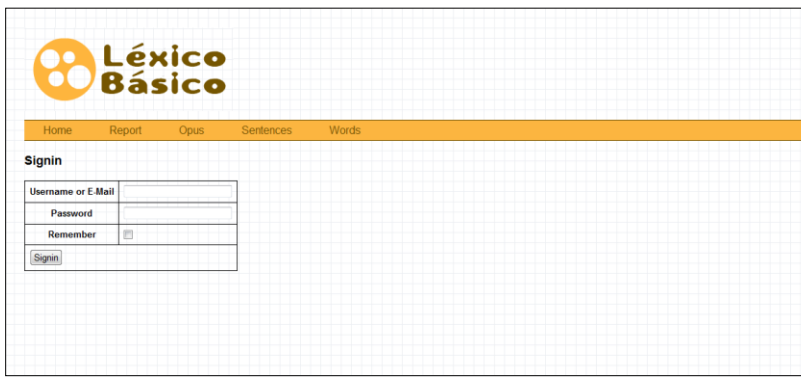
2.3. Unidades excluidas

Nuestra investigación excluye números en dígitos, siglas y nombres propios. Puesto que el lematizador considera la presencia de mayúsculas en una posición distinta de "después de punto" como indicador de que se encuentra ante un nombre propio, se

elaboró un listado etiquetado de las palabras que no son nombres propios pese a escribirse con mayúscula inicial, como <Gobierno> o <Padre> “sacerdote”, y otro de las que pueden no ser nombres propios pese a escribirse en mayúscula, como <Estado> o <Benigno>, para que el programa indicara su poca certeza en el tratamiento de esos casos y, de ese modo, los casos pudieran marcarse para desambiguación manual.

2.4. Lematización

Para la lematización, se utilizó el programa computacional LexBas 1.0. Este programa fue instalado para nuestro uso exclusivo en <http://vls43.dinaserver.com/> [6] (véase Figura 5) y estuvo operativo en ese servidor hasta el 15 de febrero de 2012. Actualmente, se encuentra operativo por medio de una máquina virtual instalada en nuestro ordenador que, para estos efectos, funciona como servidor.



[6] La gestión de una base de datos compleja por medio de herramientas diversas requiere de un ordenador con una enorme capacidad de procesamiento y memoria. Por lo mismo, es usual que aplicaciones de este tipo funcionen exclusivamente a través de una red.

Figura 5. Portal de LexBas 1.0

LexBas 1.0 es un sistema de análisis morfológico automático que permite, basándose en FreeLing 2.2, asociar palabras a un vocablo, indicando a qué lema pertenecen, y categorizarlas, distinguiendo su clase gramatical por probabilidades, según su combinatoria. No permite, sin embargo, discriminar automáticamente entre acepciones, por lo que la homonimia léxica deberá ser resuelta en una etapa posterior[7]. De hecho,

> […] tanto si utilizamos un sistema estadístico como otro basado en el conocimiento lingüístico, el problema de la desambiguación morfológica es un problema, con todas las salvedades, resuelto, ya que son varios y de diverso tipo los sistemas que ofrecen márgenes de acierto superiores al 97%. Por desgracia, […] no sucede lo mismo con la desambiguación semántica. (Marín 2009: 475)

En efecto, "los índices de acierto que hallamos en la actualidad (rara vez superiores al 70%) distan de ser mínimamente aceptables" (Marín 2009: 480), lo que no tiene que ver solo con las dificultades para traspasar al ordenador reglas o pautas de discernimiento, sino también con la baja precisión humana al realizar manualmente la misma tarea, que llega apenas al 80%, mientras que en las demás tareas de procesamiento de lenguajes naturales ronda el 95% (Marín 2009: 481). Esto se debe, por una parte, a la enorme desproporción cuantitativa entre la ambigüedad semántica y la morfológica: mientras la última presenta un inventario acotado de homónimos, la primera debe lidiar con un elevado número de acepciones; por otra parte, si las clases

[7] Para resolver la homonimia léxica, hemos elaborado un diccionario de homónimos, que combina la etiquetación *XML para la desambiguación semántica con la etiquetación EAGLE. No obstante, el adecuado procesamiento de estas etiquetas implica la realización de una serie de ajustes a LexBas.

o categorías que permiten distinguir homónimos morfológicos o sintácticos pueden ubicarse en una lista cerrada y claramente delimitada, los distintos significados de las unidades léxicas homónimas no siempre se encuentran delimitadas del mismo modo por los diversos lexicógrafos (Marín 2009: 480), lo que afecta tanto al número de acepciones contempladas para cada palabra, como al límite entre las acepciones reconocidas.

2.4.1. Desambiguación y etiquetación

Puesto que el lematizador funciona sobre la base de combinaciones sintácticas, fue necesario desambiguar las palabras que presentaban homonimia, mediante el uso de etiquetas en formato *XML en el archivo original del corpus. Este trabajo, arduo y manual, resulta básico para distinguir homónimos morfológicos o sintácticos, que en ciertos contextos podrían ser confundidos por el lematizador, especialmente si no hay marcas expresas en el texto, como un pronombre antes de un verbo, un sustantivo antes de un adjetivo o un artículo antes de un sustantivo. En estos casos, usamos etiquetas EAGLES (Expert Advisory Group on Language Engineering Standards, s/f)[8], para que el lematizador pudiera reconocerlas. Estas etiquetas permiten asignar palabras a clases y se basan en las anotaciones y comentarios apuntados en el momento de la recolección del corpus, toda vez que se detectaba riesgo de ambigüedad, especialmente cuando se observaba que la oración, desprovista de un contexto mayor, podría presentar interpretaciones diversas. Esto fue más frecuente en Drama, donde las oraciones eran más breves que en los demás

[8] Las etiquetas EAGLES 2.0 aparecen presentadas y ejemplificadas en <http://nlp.lsi.upc.edu/freeling/doc/userman/parole-es.html>

mundos (las de una palabra sola no eran inhabituales) y formaban, por lo general, parte de un diálogo; ello, sumado al recurso a lenguaje paraverbal, propio del género, significaba la común ausencia de sujetos expresos, lo que redundaba en la necesidad de indicar con gran frecuencia si el verbo se encontraba en segunda o tercera persona, ya se tratara del singular o del plural, por el sincretismo existente entre las conjugaciones correspondientes a *usted* y él/ella, y a *ustedes* y *ellos/ellas*, respectivamente, y entre *yo* y él/ella en algunos tiempos verbales.

Sin embargo, la relevancia mayor de las etiquetas aparece en el caso de la homonimia léxica, puesto que el programa no tiene la capacidad de distinguir por sí solo entre dos unidades de igual categoría gramatical que presenten diferencias semánticas de extensión o comprensión, como sucede con <as> "carta de naipe" y <as> "campeón". Las etiquetas utilizadas procuran ser lo más simples posibles y no necesitan definir con exactitud las unidades léxicas, sino únicamente desambiguar, como en los casos de <carta>as</carta>, v/s <campeón>as</campeón>. En los casos de homonimia léxica, usamos definiciones operacionales propias y las insertamos en formato *XML. Las etiquetas EAGLES, en tanto, se concentran en aspectos gramaticales.

Ahora bien, más allá de la acuciosidad que se desee en la etiquetación de las unidades, es preciso enfatizar con Morales (1986: 24) que "cualquier decisión que se tome ofrecerá reparos debido a la asistematicidad que aún presenta la teoría en muchos de estos casos, lo cual impide llegar a acuerdos generales en cada una de las lenguas". Esta asistematicidad, evidente al pretender resolver, por ejemplo, la complejidad funcional de <se> (véase Morales 1986: 23-25) o al intentar definir cuán fino se debe hilar en la distinción de acepciones de una palabra, no se queda en lo teórico, sino que alcanza el plano metodológico: así, el procesamiento automático será más o menos eficiente en

la medida en que se le puedan proporcionar reglas claras de funcionamiento que operen en todos los casos. En otras palabras, la complejidad de la tarea no radica tan solo en entender o explicar teóricamente la diferencia entre homónimos, sino en lograr distinguir las marcas formales que pueden dar cuenta de la diferencia entre uno y otro, para luego traspasar ese conocimiento a un lenguaje inteligible para el ordenador. Con todo ello, en una lengua como el español, donde las posibilidades de ordenación de elementos al interior de la oración presentan una flexibilidad mayor que en otras lenguas, es imposible predecir todas las asistematicidades que pudieran producirse en un texto, por lo que cada una de ellas significará una pérdida de fiabilidad en el procesamiento automático de ese caso en particular. Por lo mismo, nuestro programa computacional explicita en términos de porcentajes el grado de certeza del análisis de cada una de las unidades (véase Figura 6, fila “Accuracy”). Como se puede observar en “viejos sistemas políticos” (véase Figura 6, fila “Text”, posiciones

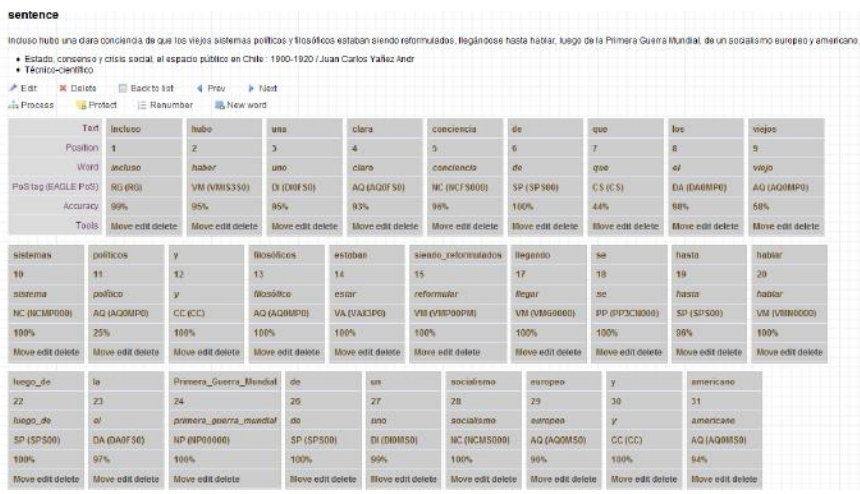


Figura 6. Ejemplo de análisis automático de una oración

9, 10 y 11), la certeza del *software* tiende a disminuir cuando un sustantivo se ve modificado por dos o más adjetivos, especialmente cuando estos adjetivos no están coordinados expresamente, sino que uno se ubica antes y el otro después del mismo sustantivo. Esto puede explicarse por el hecho de que: 1) las formas nominales *pueden* cumplir tanto la función sustantiva como la adjetiva y 2) los adjetivos *suelen* ubicarse después del sustantivo, aunque *pueden* situarse también antes de este, según lo expliquen o especifiquen; como es lógico, el computador no tiene la capacidad humana de discernir entre estas opciones, de modo que requiere reglas que no tengan excepciones; todas las condiciones incrementan la pérdida de certeza del análisis automático.

2.4.2. Precisiones respecto del análisis automático

2.4.2.1 Las formas nominales

Las formas singulares y plurales, masculinas y femeninas de los adjetivos se lematizan bajo el masculino singular. En el caso se los sustantivos, cuando la flexión de género redunda en un cambio de significado, como es el caso de el <editorial> o la <editorial>, ambas formas se lematizan por separado; si, en cambio, la flexión de género no obedece a un cambio semántico, se procede del mismo modo que con los adjetivos, como ocurre con el <gato> y la <gata>; del mismo modo se procede con las unidades que poseen diferente número. Con esta decisión, “no hace sino seguirse una tradición ampliamente justificada en teoría

gramatical” (Morales 1986: 25).

2.4.2.2. Los pronombres personales

Los pronombres personales se agrupan, de acuerdo con los procesamientos de FreeLing 2.2, por persona, número y caso. En el caso de la tercera persona, que tiene flexión de género, los femeninos se lematizan bajo la forma masculina que corresponde.

La opción de listar separadamente los pronombres personales plurales de los singulares (véase Figuras 7 y 8) permite, a nuestro juicio, una mejor discriminación de las unidades léxicas. Cabe precisar que esta alternativa es distinta de la de Morales (1986), quien los contabiliza en conjunto.

#	Text	Lemma	Pos	Eagle	Drama	Narrativa	Ensayo	Técnico	Prensa	Total	Dispersion	Use total
1	ella	él	PP	PP3FS000	139	179	61	59	38	476	0.72	342.72
2	ello	él	PP	PP3NS000	2	23	49	48	46	168	0.72	120.96
3	él	él	PP	PP3MS000	158	180	65	34	38	475	0.68	323.00
4		él	PP		299	382	175	141	122	1119	0.78	872.82

4 results

Figura 7. Pronombres personales de 3ª persona singular: caso nominativo

#	Text	Lemma	Pos	Eagle	Drama	Narrativa	Ensayo	Técnico	Prensa	Total	Dispersion	Use total
1	ellas	ellos	PP	PP3FP000	19	13	29	36	16	113	0.81	91.53
2	ellos	ellos	PP	PP3MP000	41	84	59	74	56	314	0.88	276.32
3		ellos	PP		60	97	88	110	72	427	0.90	384.30

3 results

Figura 8. Pronombres personales de 3ª persona plural: caso nominativo

A su vez, los pronombres personales que se encuentran en casos distintos se contabilizan por separado (véase Figuras 9 y 10). Inicialmente, este era uno de los aspectos que nos interesaba enmendar, subordinado al desarrollo de una solución por parte de LexBas. No obstante, pese a que lematizar todos los casos de un pronombre bajo un mismo vocablo facilitaría la comparación con otros estudios, como el de Morales (1986), nos parece también de interés facilitar la revisión de las frecuencias de los pronombres correspondientes a los distintos casos por separado. Más aún, desde un punto de vista lexicográfico, estimamos que el abordaje de un diccionario estadístico como el de Castillo Fadić (2020a), de por sí complejo, puede dificultarse en demasía si se exige al usuario discriminación de casos gramaticales para las búsquedas y no solo conocimiento del orden alfabético.

#	Text	Lemma	Pos	Eagle	Drama	Narrativa	Ensayo	Técnico	Prensa	Total	Dispersion	Use total
1	les	les	PP	PP3CPD00	132	100	52	48	58	390	0.79	308.10
2		les	PP		132	100	52	48	58	390	0.79	308.10

2 results

Figura 9. Pronombres personales de tercera persona plural: caso dativo

#	Text	Lemma	Pos	Eagle	Drama	Narrativa	Ensayo	Técnico	Prensa	Total	Dispersion	Use total
1	te	te	PP	PP2CS000	792	162	2	1	7	964	0.21	202.44
2		te	PP		792	162	2	1	7	964	0.21	202.44

Figura 10. Pronombres personales de segunda persona singular: caso dativo

2.4.2.2.1. El caso de tú /vos / usted

Puesto que en español de Chile estos tres pronombres personales coexisten, los hemos registrado separadamente, con indicación de sus respectivos índices de frecuencia, dispersión y uso. Todos ellos se agrupan bajo el lema <tú>. Como se aprecia en la Figura 11, en la columna correspondiente a la etiqueta EAGLES, tanto <vos> como <usted> incluyen al final de la etiqueta la letra <P>, correspondiente al valor “Polite”, que marca la existencia de una deferencialidad marcada, ya sea en sentido positivo <usted> como negativo <vos>.

#	Text	Lemma	Pos	Eagle	Drama	Narrativa	Ensayo	Técnico	Prensa	Total	Dispersion	Use total
1	tú	tú	PP	PP2CSN00	213	43	0	0	2	258	0.20	51.60
2	usted	tú	PP	PP2CS00P	238	33	5	0	12	288	0.21	60.48
3	vos	tú	PP	PP2CS00P	27	2	0	0	0	29	0.08	2.32
4		tú	PP		478	78	5	0	14	575	0.20	115.00
4 results												

Figura 11. Pronombres personales de segunda persona singular: caso nominativo

2.4.2.2.2. El caso de ustedes /vosotros

Como hemos indicado, a diferencia de Morales (1986), los pronombres personales de personas plurales se listan separadamente de los singulares. Por lo mismo, <vosotros> y <ustedes>

no se lematizan bajo <tú>, como en Morales (1986), sino bajo <vosotros>[9], como se observa en la Figura 12.

#	Text	Lemma	Pos	Eagle	Drama	Narrativa	Ensayo	Técnico	Prensa	Total	Dispersion	Use total
1	ustedes	vosotros	PP	PP2CP00P	59	14	1	0	0	74	0.23	17.02
2	vosotros	vosotros	PP	PP2MP000	5	2	0	0	0	7	0.30	2.10
3		vosotros	PP		64	16	1	0	0	81	0.24	19.44

3 results

Figura 12. Pronombres personales de segunda persona plural: caso nominativo

En atención a la pérdida en Latinoamérica de la oposición +/- deferencial de <vosotros> y <ustedes>, y al uso generalizado de <ustedes>, nos parece preferible lematizar bajo esta última forma y no bajo <vosotros>, como hace automáticamente LexBas 1.0, basándose en FreeLing 2.2. No obstante, hemos preferido, en esta etapa de análisis, no modificar manualmente esta lematización, en espera de que la siguiente versión de nuestro programa de análisis acoja este rasgo dialectal. Cabe precisar que en Castillo Fadić (2020a) esto ya se encuentra corregido.

2.4.3. Dificultades adicionales de análisis automático de corpus chilenos

Hemos detectado ciertas dificultades especiales en el procesamiento automático de nuestro corpus, no contempladas por los

[9] FreeLing 2.2., creado en España, lematiza automáticamente como <vosotros>, de acuerdo con la variante más extendida en la Península.

lematizadores centrados en el español estándar, preferentemente peninsular. Al respecto, destacamos: tratamiento de formas verbales, aspiración de /-s/ e inestabilidad vocálica y consonántica. A continuación, damos cuenta de cada una de estas problemáticas y del modo en que las hemos abordado.

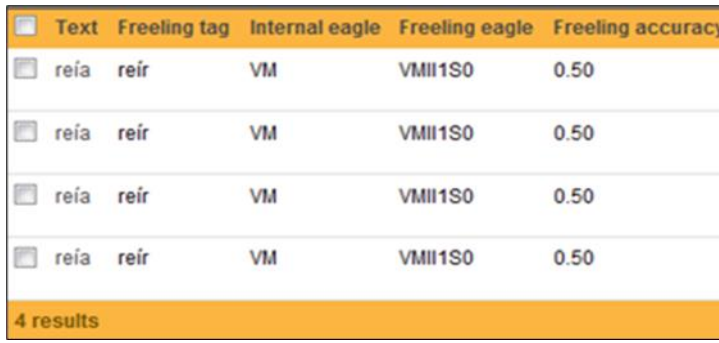
2.4.3.1. Formas verbales

2.4.3.1.1. Conjugaciones verbales de segunda y tercera persona

Como bien indica (Morales 1986: 24), Juilland y Chang-Rodríguez (1964) “no distinguen entre las formas verbales de tercera persona y segunda con *usted*, y teniendo en cuenta la importancia de esta distinción en la norma hispanoamericana, en el recuento de Puerto Rico se hizo el desglose de estas formas” Puesto que esta solución nos parece adecuada para reflejar más fielmente la realidad lingüística, hemos decidido, como Morales, distinguir entre estas formas verbales.

No obstante, el procesamiento de las formas conjugadas de los verbos presentó ciertas dificultades, originadas por el hecho de que FreeLing 2.2 está enfocado a la norma peninsular. Por lo mismo, no siempre discrimina adecuadamente las conjugaciones verbales de tercera y segunda persona, cuando la segunda persona está representada por <usted> o <ustedes>. En los casos en que el sincretismo alcanza también a la primera persona, la fiabilidad del análisis baja drásticamente y ni siquiera la presencia expresa del sujeto permite una desambiguación automática totalmente fiable. Por ello, LexBas 1.0 incorpora un índice de fiabilidad del análisis (*Freeling accuracy*); mientras más alejado esté del 100% (1), más evidente es la necesidad de revisión

manual. En el caso de <reía> (véase Figura 13), por ejemplo, la etiqueta indica que se trata de un verbo principal, modo indicativo, imperfecto, primera persona de singular; la razón de que el índice de fiabilidad esté solo en el 50% (0.50) radica en que los verbos conjugados en este modo, tiempo y persona presentan sincretismo con segunda y tercera persona de singular.



Text	Freeling tag	Internal eagle	Freeling eagle	Freeling accuracy
reía	reír	VM	VMII1S0	0.50
reía	reír	VM	VMII1S0	0.50
reía	reír	VM	VMII1S0	0.50
reía	reír	VM	VMII1S0	0.50

4 results

Figura 13. Ejemplo de análisis verbo <reír>

Este índice de fiabilidad no solo aparece junto a cada palabra analizada, sino que constituye también un parámetro de búsqueda de unidades dentro de todo el corpus. Así, por ejemplo, es posible buscar todas las palabras cuyo análisis tenga una fiabilidad igual o menor a un número determinado; en el ejemplo de la Figura 14, hemos ingresado 0.5 en la casilla *Freeling accuracy*, lo que indica que se mostrarán los casos cuya desambiguación tenga una fiabilidad igual o menor al 50%.

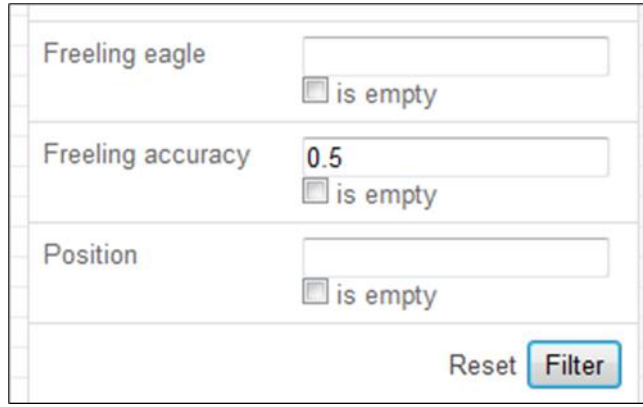


Figura 14. Filtro de búsqueda: índice de fiabilidad del análisis

Este filtro puede combinarse con otros, como el lema, la categoría gramatical, la etiqueta EAGLES, etc.

Además, para fortalecer la desambiguación de casos como estos, procuramos identificar los modos, tiempos y personas que pueden requerir revisión manual, por tender a presentar sincretismos. Esto nos permitió concentrar la revisión manual en los verbos conjugados más susceptibles a la homonimia. En la Tabla 1, las etiquetas EAGLES de la última columna permiten discernir de manera inequívoca si un verbo conjugado del corpus corresponde a uno u otro tiempo, a una u otra persona, o incluso, a uno u otro modo.

Tabla 1. Etiquetación de homónimos morfológicos: verbos

Modo	Tiempo	Persona y número			Etiqueta EAGLES[10]
		Primera conjugación	Segunda conjugación	Tercera conjugación	
Indicativo	Presente	2ª sing. 3ª sing (canta)	2ª sing. 3ª sing. (come)	2ª sing. 3ª sing. (vive)	VMIP2S0 VMIP3S0
		2ª pl. 3ª pl. (cantan)	2ª pl. 3ª pl. (comen)	2ª pl. 3ª pl. (viven)	VMIP2P0 VMIP3P0
	Presente PPS[11]	1ª pl. (cantamos)		1ª pl. (vivimos)	VMIP1P0 VMIS1P0
	Futuro	2ª sing. 3ª sing. (cantará)	2ª sing. 3ª sing. (comerá)	2ª sing. 3ª sing. (vivirá)	VMIF2S0 VMIF3S0
		2ª-pl- 3ª pl. (cantarán)	2ª pl. 3ª pl. (comerán)	2ª pl. 3ª pl. (vivirán)	VMIF2P0 VMIF3P0
	P. Imperf.[12]	1ª sing. 2ª sing. 3ª sing. (cantaba)	1ª sing. 2ª sing. 3ª sing. (comía)	1ª sing. 2ª sing. 3ª sing. (vivía)	VMII1S0 VMII2S0 VMII3S0
		2ª pl. 3ª pl. (cantaban)	2ª pl. 3ª pl. (comían)	2ª pl. 3ª pl. (vivían)	VMII2P0 VMII3P0
	PPS	2ª sing. 3ª sing. (cantó)	2ª sing. 3ª sing. (comió)	2ª sing. 3ª sing. (vivió)	VMIS2S0 VMIS3S0
		2ª pl. 3ª pl. (cantaron)	2ª pl. 3ª pl. (comieron)	2ª pl. 3ª pl. (vivieron)	VMIS2P0 VMIS3P0
	Condicional	1ª sing. 2ª sing. 3ª sing. (cantaría)	1ª sing. 2ª sing. 3ª sing. (comería)	1ª sing. 2ª sing. 3ª sing. (viviría)	VMIC1S0 VMIC2S0 VMIC3S0
		2ª pl. 3ª pl. (cantarían)	2ª pl. 3ª pl. (comerían)	2ª pl. 3ª pl. (vivirían)	VMIC2P0 VMIC3P0

Modo	Tiempo	Persona y número			Etiqueta EAGLES[10]
		Primera conjugación	Segunda conjugación	Tercera conjugación	
Subjuntivo	Presente	1ª sing. 2ª sing. 3ª sing. (cante)	1ª sing. 2ª sing. 3ª sing (coma)	1ª sing. 2ª sing. 3ª sing (viva)	VMSP1S0 VMSP2S0 VMSP3S0
		2ª pl. 3ª pl. (canten)	2ª pl. 3ª pl. (coman)	2ª pl. 3ª pl. (vivan)	VMSP2P0 VMSP3P0
	P. Imperf.	1ª sing. 2ª sing. 3ª sing. (cantara o cantase)	1ª sing. 2ª sing. 3ª sing. comiera o comiese	1ª sing. 2ª sing. 3ª sing. (viviera o viviese)	VMSI1S0 VMSI2S0 VMSI3S0
		2ª pl. 3ª pl. (cantaran o cantasen)	2ª pl. 3ª pl. (comieran o comiesen)	2ª pl. 3ª pl. (vivieran o viviesen	VMSI2P0 VMSI3P0
	Futuro	1ª sing. 2a sing. 3ª sing. (cantare)	1ª sing. 2a sing. 3ª sing. (comiere)	1ª sing. 2a sing. 3ª sing. (viviere)	VMSF1S0 VMSF2S0 VMSF3S0
		2ª pl. 3ª pl. (cantaren)	2ª pl. 3ª pl. (comieren)	2ª pl. 3ª pl. (vivieren)	VMSF2P0 VMSF3P0
Subjuntivo Imperativo	Presente	2ª sing. (cante)	2a sing. (coma)	2a sing. (viva)	VMSP2S0 VMM02S0
Subjuntivo Imperativo	Presente	2ª pl. (canten)	2ª pl. (coman)	2ª pl. (vivan)	VMSP2P0 VMM02P0

2.4.3.1.2. Formas voseantes

En julio de 2011, apreciamos que el voseo morfológico chileno, el cual se expresa en las desinencias verbales de un modo único en Hispanoamérica, especialmente en la primera conjugación (véase Rona 1962), no era reconocido por el lematizador. Planteado el problema a nuestro informático, quien lo trasladó a los

creadores de FreeLing, advertimos que esto no tenía una solución contemplada en la versión 3.0 del mencionado programa —en preparación entonces—, por lo que actualizar LexBas a partir de la versión por salir no era una alternativa; la única solución era, entonces, que nosotros proporcionáramos un listado de verbos frecuentes en conjugación voseante, etiquetados en EAGLES según un patrón establecido. Para conocer cómo se realiza esta etiquetación, puede observarse el ejemplo del verbo conjugado *hablaste*, que se etiqueta como sigue: <hablaste hablar VMIS2S0>, donde <hablaste> es la forma conjugada que se analiza, <hablar> es el infinitivo que debe funcionar como lema y <VMIS2S0> señala que se trata de un verbo principal <VM>, en modo indicativo <I>, tiempo pasado <S>, segunda persona <2>, singular <S>[10].

Por esta razón, creamos un diccionario de conjugaciones voseantes, etiquetado, que permitió procesar automáticamente y de manera exitosa el voseo verbal chileno. Cabe precisar que el voseo chileno puede aparecer en todos los tiempos verbales y tanto en modo indicativo como subjuntivo, excepto en futuro simple y pretérito perfecto simple del modo indicativo y en modo imperativo[11], como ejemplificamos en la Tabla 2. Se usa en todos los niveles sociales, en estilo coloquial, y puede combinarse con el pronombre <tú> (no marcado), o con <vos> (marcado pragmáticamente)[12].

[10] <0> es un casillero vacío, que solo se llena en los verbos si están en participio, para señalar su género femenino o masculino.

[11] A diferencia de variedades de voseo de otras zonas dialectales que contemplan una forma para el imperativo.

[12] A diferencia de lo que sucede con el voseo morfológico, el voseo pronominal presenta además covariación respecto de la variable social.

Tabla 2. Modelos de voseo verbal por conjugación

	Pronombre	Primera conjugación (verbos en –AR). Ej. cantar:	Segunda conjugación (verbos en –ER). Ej. Comer	Tercera conjugación (verbos en –IR). Ej. Vivir
Presente. Modo indicativo	Vos / tú	caminái	comís	vivís
Pretérito imperfecto. Modo indicativo	Vos / tú	caminabai	comíai	vivíai
Presente. Modo subjuntivo	Vos / tú	caminís	comái	viváis
Pretérito imperfecto. Modo subjuntivo	Vos/ tú	caminarai caminasei	comierai comiesei	vivierai viviesei
Futuro. Modo subjuntivo	Vos/ tú	caminarei	comierei	vivierei
Condicional	Vos /tú	caminaríai	comeríai	viviríai

Como es lógico, y puesto que el voseo morfológico es exclusivo de situaciones informales, si bien existe la posibilidad de que aparezca en la opción b) del imperfecto subjuntivo (véase Castillo Fadić y Sologuren Insúa 2018) y en futuro subjuntivo —en cuyo caso se conjugarían como se indica—, su frecuencia ha de ser bajísima, justamente porque estas dos conjugaciones se estiman en Chile como propias de intercambios formales e incluso, en el caso del futuro, de comunicaciones legales. Por lo mismo, en los listados que preparamos para alimentar el diccionario del lematizador, solo incluimos los siguientes tiempos y etiquetas consignados en la Tabla 3.

Tabla 3. Modelo de ingreso de formas verbales voseantes al

diccionario interno del lematizador

Conjugaciones	Etiquetas	Ejemplo de ingreso completo
Presente. Modo indicativo	VMIP2S0	hablái hablar VMIP2S0
Pretérito imperfecto. Modo indicativo	VMII2S0	hablabai hablar VMII2S0
Presente. Modo subjuntivo	VMSP2S0	hablís hablar VMSP2S0
Pretérito imperfecto (tipo A). Modo subjuntivo	VMSI2S0	hablarai hablar VMSI2S0
Condicional	VMIC2S0	hablaríai hablar VMIC2S0

2.4.3.2. Aspiración de /–s/

La aspiración de /-s/, rasgo característico del español de Chile, conlleva un serio problema para el análisis automático. Aunque solo está presente gráficamente en Drama y Narrativa, toda vez que los autores intentan reproducir en la escritura el lenguaje oral, su presencia constituye una severa traba para el lematizador, dada la enorme asistematicidad de su presentación gráfica y la ausencia de una regla fija que pueda orientar al programa.

Así, en algunos casos la aspiración de /-s/ es representada por el autor mediante una <-h> (véase Figura 15).

#	Text	Lemma	Pos	Eagle	Drama	Narrativa	Ensayo	Técnico	Prensa	Total	Dispersion	Use total
1	voh	voh	I	I	14	0	0	0	0	14	0.00	0.00
2	voh	voh	NP	NP00000	3	0	0	0	0	3	0.00	0.00
3		voh	NP		3	0	0	0	0	3	0.00	0.00
4		voh	I		14	0	0	0	0	14	0.00	0.00
4 results												

Figura 15. Ejemplo de /-s/ realizada como <-h>. Análisis julio 2011, versión inicial de LexBas 1.0.

En otros, mediante <□> (véase Figura 16).

#	Text	Lemma	Pos	Eagle	Drama	Narrativa	Ensayo	Técnico	Prensa	Total	Dispersion	Use total
1	sabí	sabí	VM	VMIS1S0	4	0	0	0	0	4	0.00	0.00
2		sabí	VM		4	0	0	0	0	4	0.00	0.00
2 results												

Figura 16. Ejemplo de /-s/ realizada como <□>. Análisis julio 2011, versión inicial de LexBas 1.0.

En ambos casos, como se observa, la versión inicial de LexBas 1.0 realiza análisis incorrectos. Al buscar alguna regla que pudiera guiar al lematizador, se revisó la posibilidad de programar al software para que reconociera como <s> las <h> que no antecedían a una vocal, por el hecho de que la distribución de <h> la sitúa, en español, en posición prenuclear y no postnuclear; no obstante, el hecho de que interjecciones como <oh>, <ah> o <eh>, de frecuencia no despreciable en el corpus, contuvieran este grafema en posición final, nos llevó a descartar esta posibilidad. La revisión de estos casos debió hacerse manualmente.

En el caso de las conjugaciones voseantes con aspiración, el ingreso de las variantes aspiradas al diccionario de conjugaciones verbales permitió solucionar errores, como se aprecia en el análisis mejorado de la palabra <sabí>, reconocida junto a <sabís> como variantes del indicativo, presente, segunda persona singular del verbo <saber>, por lo que comparten la misma etiqueta EAGLES (véase Figura 17).

#	Text	Lemma	Pos	Eagle	Drama	Narrativa	Ensayo	Técnico	Prensa	Total	Dispersion	Use total
1	sabí	saber	VM	VMIP2S0	4	0	0	0	0	4	0.00	0.00
2	sabís	saber	VM	VMIP2S0	10	0	0	0	0	10	0.00	0.00
2 results												

Figura 17. Palabra <sabí>. Lematización mejorada

2.4.3.3. Inestabilidad vocálica y consonántica

La inestabilidad, tanto vocálica como consonántica, que caracteriza al español de Chile, especialmente popular, y que se expresa preferentemente en situaciones coloquiales, es también una traba para el análisis automático, cuando los autores intentan reflejar la realidad hablada. En el caso de <dispertaste>, por ejemplo (véase Figura 18), aunque el programa logra determinar que se trata de un verbo conjugado, basado en reglas de combinatoria, no posee las herramientas para vincularlo con el verbo <despertar>, lo que redunda en una lematización inadecuada. La bajísima frecuencia de esta unidad, amén de su baja dispersión, la dejan en un margen irrelevante, desde el punto de vista estadístico, como suele ocurrir con estas formas populares que, por no estar estandarizadas, presentan una enorme asistematicidad.

#	Text	Lemma	Pos	Eagle	Drama	Narrativa	Ensayo	Técnico	Prensa	Total	Dispersion	Use total
1	dispertaste	dispertaste	VM	VMIS2S0	1	0	0	0	0	1	0.00	0.00
1 result												

Figura 18. Ejemplo de análisis de forma que reproduce la oralidad

Puesto que el lematizador opera a partir de combinaciones sintácticas, la dificultad para procesar una de las unidades de una oración puede repercutir en el análisis incorrecto de toda la estructura, como se observa en la Figura 19.

Figura 19. Ejemplo de análisis automático de discurso que reproduce la oralidad (julio 2011)

Para resolver estos casos, trabajamos con herramientas de edición al interior del mismo programa de análisis (véase Figura 20); al seleccionar la opción *Edit,* se abre una ventana que permite realizar cambios manuales; para ello, es preciso revisar todos los campos asociados a cada unidad léxica, sin limitarse exclusivamente a indicar el lema correcto en la casilla *Freeling tag*; en *Internal eagle* se despliegan las distintas categorías gramaticales que considera el programa, pero esta sola categoría es insuficiente para el análisis, pues solo *Freeling Eagle* permite precisar género y número —en el caso de las formas nominales— y modo, tiempo, persona y número —en el caso de las formas verbales—.

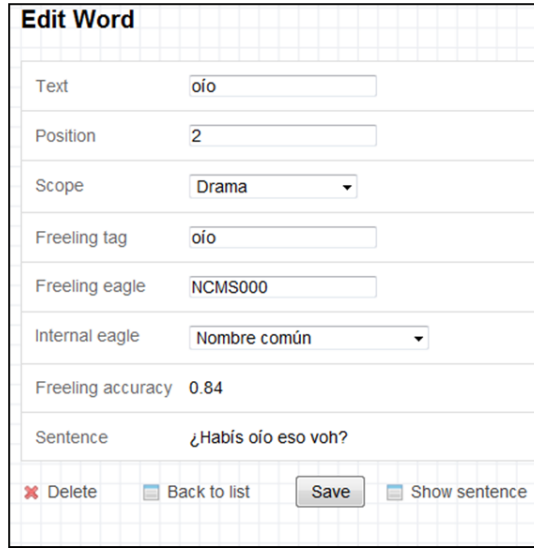


Figura 20. Editor de palabras

III. Análisis del corpus

Una vez concluido el proceso de orden, segmentación, depuración y lematización del corpus, se realizó un análisis estadístico para obtener la frecuencia de las palabras y vocablos, su dispersión y su uso. Para ello, se contó con la ayuda de LexBas 1.0 y se usaron las fórmulas mejoradas de Morales (1986). Así, se contabilizaron las frecuencias absolutas de las unidades y se calculó su dispersión mediante la fórmula

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

Para el cálculo del uso, se multiplicó la frecuencia por la dispersión: U = F x D.

Puesto que LexBas 1.0 solo permite hacer consultas predeterminadas y no soporta ciertos análisis, de manera complementaria recurrimos a Excel. Esto nos permitió no solo acceder al léxico básico del español de Chile —presentado en la forma de un diccionario no definitorio (Castillo Fadić 2020a)—, sino también obtener listados comparados de los vocablos de mayor uso y frecuencia, estableciendo cortes en los rangos 100, 500 y 1505, a fin de cotejar, además, con obras previas que realizan cortes en los mismos rangos.

Por otra parte, la posibilidad que ofrece LexBas 1.0 de aislar las unidades por categoría gramatical (véase Figura 21), nos brindó la opción de revisar separadamente unidades pertenecientes a distintas categorías, lo que permitió un análisis fino de casos particulares (véase Figura 22).

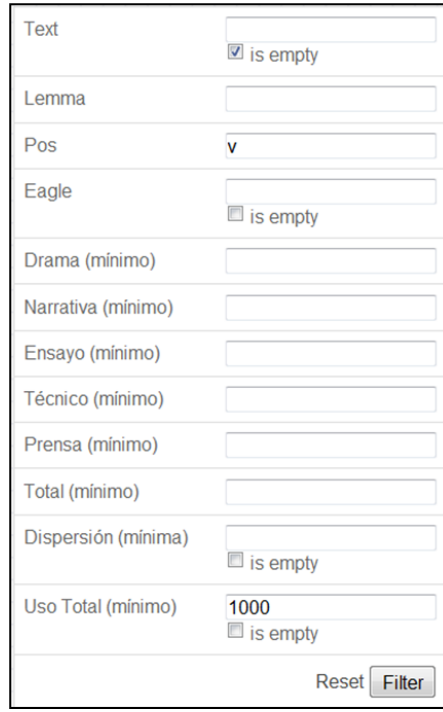


Figura 21. Ejemplo de búsqueda: vocablos clasificados como verbos (V), con uso total mínimo de 1000

#	Text	Lemma	Pos	Eagle	Drama	Narrativa	Ensayo	Técnico	Prensa	Total	Dispersion	Use total
1		ser	VS		2066	1338	1442	1389	890	7125	0.87	6198.75
2		estar	VA		903	492	277	240	363	2275	0.74	1683.50
3		tener	VM		830	426	271	342	307	2176	0.77	1675.52
4		poder	VM		482	312	342	449	187	1772	0.85	1506.20
5		hacer	VM		598	414	249	240	276	1777	0.81	1439.37

5 results

Figura 22. Verbos de uso superior a 1000

En otros casos, la posibilidad de buscar palabras —*Text*— a

partir de un segmento permitía acceder a unidades que compartieran una misma base léxica o, al menos, una misma configuración grafemática, lo que facilitaba la revisión del procesamiento automático. En el caso de la Figura 23, la búsqueda de <pudr>, sin precisión del lema ni de la categoría gramatical, nos remite a un listado de tipos; para acceder a las oraciones en las que aparecen las palabras correspondientes, es preciso clickear en *Show*.

Mediante las mencionadas herramientas informáticas, dimos cuenta de las características del corpus, tanto en lo relativo al número de vocablos y palabras registrados en total y por mundo, como en lo referente a la proporción en la que se presentan las distintas categorías gramaticales, según se refleja en las etiquetas EAGLES correspondientes.

Aplicando índices complementarios de medición propuestos por López Morales (1984) y retomados por Haché de Yunén 1991, determinamos la riqueza léxica de nuestro corpus como indicador de su suficiencia[13] y observamos la covariación de la riqueza respecto de la variable mundo.

#	Text	Lemma	Pos	Eagle	Drama	Narrativa	Ensayo	Técnico	Prensa	Total	Dispersion	Use total	Actions
1	pudran	podrir	VM	VMSP3P0	4	0	0	0	0	4	0.00	0.00	Edit Show
2	pudre	podrir	VM	VMIP3S0	2	0	0	0	0	2	0.00	0.00	Edit Show
3	pudrición	pudrición	NC	NCFS000	1	0	0	0	0	1	0.00	0.00	Edit Show
4	pudriría	podrir	VM	VMIC1S0	1	0	0	0	0	1	0.00	0.00	Edit Show
5	pudriéndote	podrir+te	VM	VMG0000+PP2CS000	0	1	0	0	0	1	0.00	0.00	Edit Show

5 results

Figura 23. Consulta a partir de un segmento de una palabra

[13] "La única manera de lograr que una muestra léxica sea relativamente representativa [...] es [...] cuidando que [...] esté compuesta de una rica variedad, [...] en una cantidad que resulte *suficiente*. [...] La suficiencia de un corpus depende, primero, de que hayamos tratado de eliminar posibles sesgos en la muestra, asegurándonos de que la selección de los datos haya sido *aleatoria*; luego, de la *variedad* que le hayamos dado a nuestra recolección; después, de su *cantidad*; por último y de manera más importante, de la *riqueza léxica* que el propio corpus nos va mostrando durante su análisis [...]" (Lara 2006: 155).

Establecimos la representatividad acumulada (R) de nuestro corpus en general y de nuestro léxico básico en particular, mediante la fórmula $R = \frac{\sum_{i=1}^{n} F_i}{N}$, que observa el cociente entre la sumatoria de las frecuencias totales de los vocablos contemplados dentro de un rango y el número total de vocablos contenidos en ese mismo rango (véase Castillo Fadić 2012b y Castillo Fadić y Sologuren Insúa 2020). En la fórmula, n representa el número de rango del vocablo de mínimo uso o frecuencia dentro del rango (la unidad ubicada en el rango de corte), mientras que N corresponde al número total de palabras consideradas en la muestra, vale decir, desde el rango uno (1) al de corte. La representatividad del corpus se calculó a partir de las frecuencias totales de los vocablos ordenados por uso de mayor a menor.

Se consideró de interés, además, determinar la curva de cobertura por mundo, para lo cual la fórmula se aplicó considerando cada mundo como una base independiente, donde, por no ser pertinente el índice de dispersión, las unidades se ordenaron por frecuencia total de modo decreciente.

Precisamos también qué unidades de alta frecuencia tienen baja dispersión y a la inversa, y cotejamos los resultados obtenidos en los distintos mundos desde diferentes enfoques, observando la covariación de diferentes variables en un estudio de implicancias sociolingüísticas. En esta línea, revisamos las unidades con dispersión máxima que no presentan necesariamente una alta frecuencia, sino que se caracterizan por el equilibrio de frecuencia entre los mundos; dimos cuenta también de las unidades con dispersión mínima y generamos listados de vocablos de alta frecuencia, pero con dispersión cero, organizados por mundos. Al respecto, levantamos gráficos para comparar la representación de las distintas clases gramaticales en cada mundo, observando diferencias de frecuencia y de inventario.

Nos pareció también de interés evaluar posibles afinidades entre mundos, para lo cual aislamos los vocablos que aparecen en solo dos de ellos, con independencia de sus índices de frecuencia, dispersión y uso, lo que nos permitió apreciar tendencias marcadas de afinidad entre unos y otros mundos y graficar dicha afinidad en general y por mundo.

Por último, aplicamos criterios de selección complementarios al índice de uso para determinar, dentro de las unidades de mayor uso, las que forman parte del núcleo estadístico de mayor estabilidad dentro del español de Chile, denominado léxico básico, y organizamos los resultados en un diccionario de frecuencia no definitorio (véase Castillo Fadić 2020a).

IV. Conclusiones

Los métodos empleados para procesar y analizar el corpus han resultado provechosos. Nos han permitido configurar un corpus de referencia etiquetado y estratificado del español de Chile, a partir del cual hemos podido no solo obtener el léxico básico del español de Chile (Castillo Fadić 2020a), sino también realizar una serie de investigaciones de implicancias sociolingüísticas (véase, por ejemplo, Castillo Fadić 2015b y 2019; y Castillo Fadić y Sologuren Insúa 2017 y 2018).

Esperamos que los lineamientos para ordenar los materiales y los criterios propuestos para segmentar, excluir y lematizar unidades léxicas puedan ser de utilidad para quienes requieran procesar corpus lingüísticos hispánicos. Muy especialmente, esperamos que nuestras propuestas ofrezcan alguna orientación a quienes decidan emprender el procesamiento automático de corpus de español no castellano □hispanoamericano, andaluz,

canario, africano, etc.□, que probablemente se encuentren, como nosotros, ante soluciones pensadas para otras lenguas o para otras variedades de español.

En lo relativo al análisis estadístico, los cálculos de frecuencia, dispersión y uso permitieron obtener el *Léxico Básico del Español de Chile* (Castillo Fadić 2020a). Los cálculos adicionales de riqueza léxica, representatividad y curvas de cobertura por mundo, entre otros, pueden prestar utilidad en el ámbito educativo (véase, por ejemplo, Castillo Fadić y Sologuren Insúa 2020) y, particularmente, en la planificación de la enseñanza del español como lengua materna y como segunda lengua, donde pueden complementarse con otros métodos centrados en la selección (Santos Díaz 2017) y enseñanza del léxico (Santos Díaz, Trigo Ibáñez y Romero Oliva 2020a, 2020b).

Particularmente, en lo que atañe a los repertorios de léxico básico, esta línea de investigación es tan relevante como poco abordada, por lo que invitamos a los lingüistas interesados en lingüística de corpus a emprender trabajos en esta línea, dentro de sus respectivas comunidades.

Referencias bibliográficas

1. Almela, Ramón, Cantos, Pascual, Sánchez, Aquilino, Sarmiento, Ramón y Almela, Moisés (2005) *Frecuencias del español. Diccionario y estudios léxicos y morfológicos.* Madrid: Universitas S.A.
2. Alvar Ezquerra, Manuel y Villena Ponsoda, Juan Antonio (eds.)(1994) *Estudios para un corpus del español*. Málaga: Universidad de Málaga.
3. Alvar Ezquerra, Manuel y Corpas Pastor, Gloria (1994)

"Criterios de diseño para la creación de córpora". En *Estudios para un corpus del español*. Coords., Manuel Alvar Ezquerra y Juan Antonio Villena Ponsoda. Málaga: Universidad de Málaga: 31-40.

4. Alvar Ezquerra, Manuel, Blanco Rodríguez, María José y Pérez Lagos, Fernando (1994) "Diseño de un corpus español en el marco de un corpus europeo". En *Estudios para un corpus del español*. Coords., Manuel Alvar Ezquerra y Juan Antonio Villena Ponsoda. Málaga: Universidad de Málaga: 9-30.
5. Ávila, Manuel Antonio (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.
6. Castillo Fadić, María Natalia (2012a) *Corpus Básico del Español de Chile* ©.
7. Castillo Fadić, María Natalia (2012b) "Léxico Básico del Español de Chile". Tesis de doctorado. Universidad de Valladolid, España.
8. Castillo Fadić, María Natalia (2015a) "Léxico Básico del Español de Chile: el proyecto". *E-Aesla. Revista digital*. Consultado: 25 de diciembre de 2017. <https://cvc.cervantes.es/lengua/eaesla/pdf/01/51.pdf>
9. Castillo Fadić, María Natalia (2015b) "El verbo <hacer> en el español de Chile: tipos y combinaciones frecuentes en el género ensayo". E-Aesla. Revista digital. 1. 1-9.

Consultado: 25 de diciembre de 2017. <https://cvc.cervantes.es/lengua/eaesla/pdf/01/50.pdf>

10. Castillo Fadić, María Natalia (2019) "¿Qué se dice de la mujer y el hombre en el español de Chile?: estudio exploratorio de las combinaciones frecuentes de los vocablos mujer y hombre en un corpus de referencia estratificado". *Boletín de Filología*. 54. 1, 95-117. <https://doi.org/10.4067/S0718-93032019000100095>
11. Castillo Fadić, María Natalia (2020a) *Léxico Básico del Español de Chile.* Santiago de Chile: Liberalia Ediciones, Fondo del Libro y la Lectura (en prensa).
12. Castillo Fadić, María Natalia (2020b) "*Corpus Básico del Español de Chile ©:* metodología de obtención, revisión y constitución definitiva". En *Boletín de Filología, Estudios en homenaje a Alfredo Matus Olivier*. Eds., Abelardo San Martín, Darío Rojas y Soledad Chávez.
13. Castillo Fadić, María Natalia y Sologuren Insúa, Enrique (2017) "El reformulador <es decir> en el español de Chile: una propuesta de clasificación funcional". *Lenguas modernas*. 49, 77-92.
14. Castillo Fadić, María Natalia y Sologuren Insúa, Enrique (2018) "Pretérito imperfecto de subjuntivo en el español de Chile: ¿existe alternancia libre entre las desinencias –ra y –se?". *Onomázein*. 42, 153-171.
15. Castillo Fadić, María Natalia y Sologuren Insúa, Enrique 2020) "Léxico frecuente, riqueza léxica y estereotipos sobre la lectura de profesores en formación". *Logos Revista de Lingüística, Filosofía y Literatura*. Universidad de La

Serena.

16. Corpas Pastor, Gloria (1994) “Anotación semántica y ambigüedad”. En *Estudios para un corpus del español. Anejo n.º 7 de Analecta Malacitana*. Coords., Manuel Alvar Ezquerra y Juan Antonio Villena Ponsoda. Málaga: Universidad: 103-112.
17. Corpas Pastor, Gloria (1997) *Manual de fraseología española.* Madrid: Gredos.
18. Dewey, Melvil (1989) *Dewey decimal classification and relative index*. Vols. I, II, III y IV. Vigésima edición. Ed., John Comaromi. Albany: Forest Press.
19. Expert Advisory Group on Language Engineering Standards(s/f*) Welcome to EAGLES on line*. Consultado: 16 de febrero de 2012. <http://www.ilc.cnr.it/EAGLES96/home.html>
20. Haché de Yunén, Ana Margarita (1991) “Aportes de las pruebas de riqueza léxica a la enseñanza de la lengua materna”. En *La enseñanza del español como lengua materna.* Ed., Humberto López Morales. Río Piedras: Universidad de Puerto Rico: 49-60
21. Juilland, Alphonse y Chang-Rodríguez, Eugenio (1964) *Frequency Dictionary of Spanish Words, The Romance Languages and their Structures, First Series SI.* La Haya: Mouton.
22. Juilland, Alphonse, Traversa Vicenzo, Beltramo Antonio y Di Blasi, Sebastiano (1973) *Frequency Dictionary of Italian Words.* The Hague-Paris: Mouton.
23. Lara, Luis Fernando (2006) *Curso de lexicología.* México:

El Colegio de México.

24. Lavid, Julia (2005) *Lenguaje y nuevas tecnologías. Nuevas perspectivas, métodos y herramientas para el lingüista del siglo XXI.* Madrid: Cátedra.

25. López Morales, Humberto (1984) *La enseñanza de la lengua materna.* Madrid: Playor.

26. López Morales, Humberto (2020) "Prólogo". En *Léxico Básico del Español de Chile*, de María Natalia Castillo Fadić. Santiago de Chile: Liberalia Ediciones, Fondo del Libro y la Lectura: en prensa.

27. Lyons, John (1997) *Semántica lingüística. Una introducción.* Barcelona: Paidós.

28. Maríns, Rafael (2009) "El tratamiento computacional del léxico y sus aplicaciones". En *Panorama de la lexicología*. Ed., Elena de Miguel. Barcelona: Ariel: 465-486.

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

30. Moreno Fernández, Francisco (2016) *La lengua española en su geografía: manual de dialectología hispánica*. 3ª edición. Madrid: Arco.

31. Rona, José Pedro (1962) "El problema de la división del español americano en zonas dialectales". En *PFLE* I. Madrid: Ediciones de Cultura hispánica, 215-226.

32. Santos Díaz, Inmaculada Clotilde (2017) "Selección del léxico disponible: propuesta metodológica con fines didácticos". *Porta Linguarum*. 27. 122-139.

33. Santos Díaz, Inmaculada Clotilde, Trigo Ibáñez, Ester y

Romero Oliva, Manuel Francisco (2020a) “La activación del léxico disponible y su aplicación a la enseñanza de una lengua”. *Porta Linguarum.* 33. 75-93.

34. Santos Díaz, Inmaculada Clotilde, Trigo Ibáñez, Ester y Romero Oliva, Manuel Francisco (2020b) Propuesta de una taxonomía de los centros de interés en los estudios de disponibilidad léxica”. *Delta, Documentação e Estudos em Linguística Teórica e Aplicada.* 36. 4, 1-28.
35. Zamora Munné, Juan y Guitart, Jorge (1982) *Dialectología hispanoamericana. Teoría – Descripción – Historia.* Salamanca: Ediciones Almar.

Recibido: 01/12/2018
Aceptado: 22/05/2020