Investigaciones fraseológicas y corpus textuales

Guillermo Rojo[1]
Universidade de Santiago de Compostela, España
Real Academia Española

1. INTRODUCCIÓN

En los últimos años, incluso desde antes de que el Pleno de ASALE aprobara en el congreso de Ciudad de México (2015) la propuesta de emprender la elaboración del *Diccionario panhispánico de fraseología* (*DPF*), he tenido ocasión de hablar con Alfredo Matus y con algunos otros colegas (entre los que ocupa un lugar destacado Francisco Javier Pérez) de la forma en que cabe enfocar la explotación de la información contenida en los corpus de español que podemos manejar en la actualidad y la elaboración del *DFP*. No siempre hemos estado de acuerdo, como sucede con frecuencia, pero mi visión se ha enriquecido con esas discusiones y me ha parecido oportuno, en consecuencia, exponer las líneas generales de este campo de trabajo en un volumen dedicado a homenajear al amigo y colega que durante tantos años ha dirigido la Academia Chilena de la Lengua.

La lingüística de corpus (LC) tiene una historia muy corta todavía, pero las posibilidades que brinda y la orientación metodológica que proporciona hacen que haya pasado a ocupar un lugar central en la investigación lingüística, con una perspectiva transversal, que la hace resultar útil en las más diversas disciplinas

[1] Para correspondencia, dirigirse a: Guillermo Rojo (guillermo.rojo@usc.es), Facultad de Filología. Universidade de Santiago de Compostela, 15782 Santiago de Compostela, España.

y subdisciplinas lingüísticas e incluso, aunque solo hasta cierto punto, con independencia de la aproximación teórica adoptada. La investigación de las unidades fraseológicas (UF) no es, por supuesto, una excepción, pero creo que las causas de esta proximidad pueden encontrarse en los mismos orígenes de la LC.

Tal como es presentada habitualmente, la historia de la LC tiene dos aspectos necesitados de una reformulación radical. El primero de ellos consiste en el hecho de que los tratados generales y también los estudios específicos acerca de esta aproximación se mueven exclusivamente en el universo de la lingüística inglesa, con lo que los antecedentes y las contribuciones procedentes de otras tradiciones son totalmente ignorados. El segundo es menos importante en apariencia, pero tiene efectos visibles sobre el tema que nos ocupa. Según la visión generalizada, la LC surge en los Estados Unidos a mediados de los años 60 del siglo pasado, en un momento en que la lingüística de orientación chomskyana está despegando y convirtiéndose en la vanguardia de la investigación en nuestra disciplina. En ese contexto, una aproximación basada en el procesamiento automático de datos no procedentes de la introspección, el papel de la frecuencia, análisis estadísticos, etc. era mal considerada, de modo que la LC pasó unos cuantos años de vida mortecina, poco menos que clandestina, de la que tardó bastante tiempo en salir. Todo esto es cierto, pero solo en lo referente a Estados Unidos. En la preparación del Brown Corpus participó ya Randolph Quirk, que tenía la experiencia de haber puesto en marcha los trabajos de construcción del Survey of English Usage, iniciado en 1959. Esta línea de trabajo se vio reforzada con la publicación del Brown Corpus, que tuvo muy pronto su contrapartida británica, el Lancaster - Oslo / Bergen Corpus (LOB). La LC tuvo un desarrollo rápido e intenso en el Reino Unido y algunos países nórdicos, casi siempre con corpus de lengua inglesa. Lo que me interesa destacar aquí es el hecho de que la línea metodológica correspondiente a la LC encaja de modo muy natural con el estructuralismo inglés, especialmente en la corriente que tiene su origen en Firth y continúa con Quirk, Halliday y Sinclair, entre muchos otros. En esa orientación, la caracterización del significado de una unidad en función de las que se dan a su alrededor y la existencia de unidades léxicas "extendidas", superiores a la palabra resultan perfectamente naturales y el título de la obra seminal de Sinclair 1991 (*Corpus*, *Concordance*, *Collocation*) resulta suficientemente ilustrativo de esta orientación[2]. La posibilidad de usar corpus textuales para el análisis e incluso la detección de las UF está presente, pues, desde los mismos orígenes de la LC, al menos de la LC practicada en Europa.

[2] Rundell (2018) analiza con amplitud la importancia que Sinclair dio a las "extended units of meaning" desde sus primeras incursiones en el trabajo con corpus.

La primera gran ventaja de los corpus textuales radica en su tamaño: los que consideramos corpus de referencia constan de cientos de millones de formas, como el Corpus de Referencia del Español Actual (CREA) o el Corpus del Español del Siglo xxi (CORPES XXI), o de miles de millones como el Corpus del Español Web / Dialectos (CdEweb) o el Es-Ten-Ten (Spanish Corpus from the Web). Dado que están en formato electrónico, es sencillo programar rutinas que procesen su contenido y proporcionen información elaborada automáticamente, como, por ejemplo, listas de frecuencias de las formas o de las combinaciones de formas (los llamados *n-gramas*), que nos pueden facilitar la identificación de las unidades multipalabra en general y las unidades fraseológicas (UF) en particular. Por otro lado, los textos que componen los corpus están codificados, lo cual significa que cada uno de ellos lleva la indicación correspondiente acerca del valor que presenta en cada uno de los parámetros pertinentes: país, medio, soporte, tipo de texto, etc. Gracias a esta característica, es posible obtener la frecuencia con que se registra un cierto elemento o fenómeno en los diferentes subcorpus que se pueden construir de forma dinámica en función de los intereses de la investigación planificada. Este es un punto decisivo en general y también para lo que aquí nos ocupa, puesto que lo realmente importante no es la frecuencia de una expresión como *me importa un rábano* en todo el CORPES, sino la comparación entre las que presenta en cada uno de los países del ámbito hispánico. Por último, los corpus son procesados lingüísticamente, de modo que reciben la información léxica y gramatical asociada a cada uno de los elementos que hay en su interior. Gracias a este proceso, que, como es evidente, tiene que realizarse de forma automática, podemos obtener datos que corresponden no a patrones ortográficos, sino a características léxicas (como todas las formas del verbo *importar*), gramaticales (todos los casos de copretérito de indicativo de todos los verbos), combinatorias (las palabras que aparecen con mayor frecuencia en torno al verbo *importar*) o diferentes combinaciones de todas ellas. De todo esto deriva que la utilización de corpus electrónicos en las investigaciones fraseológicas es poco menos que forzosa[3]. No, por supuesto, para resolver problemas de índole teórica, como, por ejemplo, si *importar un rábano* e *importar un pimiento* son dos UF distintas o dos variantes de la misma UF o bien si la UF es *importar un pimiento* o solo *un pimiento*. Los datos contenidos en los corpus contribuyen decisivamente, en cambio, a mejorar nuestro conocimiento acerca del uso de estas expresiones en los ejes de variación, como trataré de mostrar en el apartado 2, y a la detección de expresiones que responden a un mismo esquema constructivo, como veremos en el apartado 3.

[3] Como parte de una colaboración mucho más amplia entre la lexicografía y el procesamiento del lenguaje natural (PLN). *Cf*. Gantar 2019.

2. DISTRIBUCIÓN

Localizar e identificar todas las variantes de una (candidata a) UF en un diccionario usual presenta dificultades de varios tipos. La más superficial, pero también la más importante para quien no conoce el significado de una expresión como *me importa un comino* y necesita por ello consultar un diccionario, consiste en saber en qué entrada tiene que buscarla. Es, pues, una dificultad de tipo práctico. En la tradición lexicográfica hispánica, las expresiones complejas se sitúan habitualmente en la entrada correspondiente al primer sustantivo que figura en ellas, de modo que no sirve de nada ir a la entrada *importar*, sino que hay que ir a *comino* y allí se encuentra, en el *DLE*, la locución verbal *importar a alguien un comino*, que se define como 'ser insignificante, o de poca o ninguna importancia para esa persona'. Es una convención fácilmente manejable en la consulta de expresiones concretas, pero es evidente que esa organización de los datos hace imposible en la práctica recuperar el inventario de variantes, puesto que habría que revisar, una a una, al menos todas las entradas de los sustantivos candidatos a entrar en una locución con este significado. Los diccionarios en formato impreso recurren a las remisiones para paliar este problema: en el *DLE*, *cielo borreguero* está en la entrada correspondiente a *cielo*, pero en *borreguero* se remite también a la expresión compleja. Sin embargo, esto no sucede con las locuciones verbales construidas con *importar*, por lo que en la entrada correspondiente al verbo no figuran esas expresiones ni se remite a ellas. Naturalmente, lo mismo sucede en los formatos electrónicos que se limitan a reproducir la estructura y organización de los diccionarios impresos. Es importante tener en cuenta en este punto que la inclusión en la entrada correspondiente al sustantivo se produce en el *DLE* tanto en los casos en los que la construcción es considerada como *importar* + locución adverbial (*un pito*) como en aquellos en los que se habla de una locución verbal (*importar un comino*)[4].

La conversión de los diccionarios en bases de datos lexicográficos (BDL) da muchas más posibilidades, puesto que permite entrar en el interior de cada uno de los campos en que se estructura la entrada y recuperar la información que interesa en cada caso. Esa es la posibilidad que ofrece la opción denominada "Diccionario avanzado" en el Enclave RAE. Las locuciones encabezadas por *importar* pueden ser recuperadas directamente incluyendo este verbo en el campo de búsqueda por lema. Escribiendo en él *importar* o *importarle* aparecen las que se construyen con los sustantivos *caracol*(*es*), *chita*, *clavo*, *comino*, *cuerno*, *paja*, *pimiento*, *rábano*, así como *importar a alguien madre algo*, que no responde al mismo esquema constructivo, pero posee el mismo significado

[4] El diccionario *LEMA* remite, en la entrada *importar*, a *comino*, *cuerno*, *pepino* y *pito*. El *DEA* incluye una subacepción para este tipo de estructura y remite a *bledo*, *carajo*, *comino*, etc.

y, por tanto, se define de modo similar. Todas ellas son caracterizadas en el *DLE* como locuciones verbales con *importar* y, en muchos casos, también con *valer*. Podemos incluso aumentar el ámbito de búsqueda y escribir *importar* en el campo de la definición, lo cual recupera *valer algo charra*. No aparecen, sin embargo, las expresiones formadas con *ardite*, *bledo* o *pepino* y tampoco las que usan *carajo*, *pijo* o *pito*. La explicación es clara y se relaciona con el segundo problema vinculado a las UF. Según el *DLE*, *importar un ardite* / *bledo* / *pepino* no son locuciones verbales sencillamente porque esos tres sustantivos significan, también 'cosa de poco valor', lo cual hace que se considere que su construcción *importar* (o *valer*) tiene significado recto. Distinto es el caso de las otras tres posibilidades. El *DLE* considera que *un carajo*, *un pijo* y *un pito* son locuciones adverbiales que significan 'muy poco o nada', por lo que tampoco en este caso figuran explícitamente las construcciones con *importar*. Todas ellas aparecen si, aprovechando las ventajas del acceso a la BDL incluido en Enclave RAE, se escribe *importar* en el campo de la definición o bien de los ejemplos, pero está claro que con eso nos salimos de lo que se podrían considerar expresiones del tipo *importar un rábano*. En el cuadro 1 aparece la relación de expresiones que figuran en el DLE con la caracterización que se les da y la definición utilizada.

Así pues, la recuperación de expresiones del tipo (*no*) *importar* + *cuantificador* + *sustantivo* devuelve, en el *DLE*, construcciones de tres tipos distintos, lo cual pone de relieve el segundo problema mencionado: decidir si estamos ante una UF y, en caso positivo, cuál es y a qué subtipo pertenece:

locución verbal:	*importar un rábano*
verbo + locución adverbial:	*importar un pito*
verbo + frase nominal:	*importar un pepino.*

Casi todas las expresiones mencionadas hasta ahora llevan marca de uso coloquial y solo (*no*) *importar madre*(*s*) lleva marca geográfica (México). Hay que entender, por tanto, que las demás son de uso generalizado en todo el ámbito hispánico. Para poder valorar lo que se obtiene del *DLE* vamos a realizar dos operaciones distintas. Por una parte, la comparación con lo que se encuentra en el *Diccionario fraseológico documentado del español actual* (*DFDEA*). También en esta obra se sigue la convención de situar las locuciones en la entrada correspondiente al sustantivo que contienen, pero incluye una relación inicial que sigue el orden alfabético estricto, de modo que todas las que contienen *importar* aparecen seguidas, lo cual hace sencillo localizarlas de una sola vez. En el cuadro 1 aparecen todas las expresiones de este tipo que figuran en uno de estos diccionarios o en ambos, con la caracterización utilizada en cada caso. Al lado de las coincidencias esperables, se observan algunas discrepancias de interés. La más llamativa, en mi opinión, radica en el número de UF que son caracterizadas como de carácter adverbial (o pronominal), grupo en el que se incluyen algunas que el *DLE* da como verbales (*importar un comino*, por ejemplo). El sistema utilizado habitualmente consiste en presentar la frase nominal (*un*

comino) como una UF de carácter adverbial (o pronominal), dar 'nada' como equivalente e indicar luego que se construye con los verbos *importar* y *valer*. *Importar un comino* y *valer un comino* figuran en la relación de UF[5]. La mayor parte de ellas son consideradas coloquiales, pero también se dan algunas como vulgares. Como muestra el cuadro, varias de estas expresiones no figuran en el *DLE* y, a cambio, en el *DFDEA* no figuran *importar un caracol* / *una chita* / *un clavo* ni *un pijo*.

DLE					**DFDEA**			
ardite				Cosa insignificante o de muy poco valor	*[importar] un ardite*	loc. pron. y tb. adv.	lit.	Nada
bledo				Cosa insignificante o de poco valor. +Ejemplo con *importar*	*[importar] un bledo*	loc. adv.	col.	Nada
no importar algo dos caracoles, o un caracol; o no dársele algo dos caracoles, o un caracol; o no valer algo dos caracoles o un caracol	locs. verbs.			U. para demostrar el desprecio o la poca estimación de algo.				
un carajo	loc. adv. y pron.	coloq.		Muy poco o nada + Ejemplo con *importar*	*[importar] un carajo*	loc. pron. y tb. adv.	vulg.	Nada
					[importar] un carallo	loc. pron. y tb. adv.	euf.	→ un carajo
no importar, o no valer, una chita	locs. verbs.	coloqs.		No importar un bledo				
no importar un clavo algo	loc. verb.	coloq.		Merecer poco aprecio.				
					importar tres cojones	loc. verb.	vulg.	No importar en absoluto

[5] Esa doble posibilidad es la que hace que no se considere que la UF es *importar un comino*.

importar a alguien un comino	loc. verb.			Ser insignificante, o de poca o ninguna importancia para esa persona.	*[importar] un/tres comino(s)*	loc. adv. y tb. pron.		Nada
importarle a alguien un cuerno algo o alguien	loc. verb.	coloq.		Traerle sin cuidado.	*[importar] un cuerno*	pron. y tb. adv.		Nada
					importar una higa	loc. verb.	lit.	No importar en absoluto
					[importar] un higo	loc. pron. y tb. adv.	col.	Nada
					[importar] un huevo	loc. adv. y tb. pron.	vulg.	Nada
					[importar] una mierda	loc. pron. y tb. adv.	vulg.	Nada
importar a alguien madre algo	loc. verb.	malson. coloq.	Méx.					
					importar tres (pares de) narices	loc. verb.	col.	No importar en absoluto
no importar, o no montar, una paja	locs. verbs.			Valer muy poco algo, por inútil o de poca entidad.				
					importar tres pelotas	loc. verb.	vulg.	No importar en absoluto
					[importar] un pepino	loc. pron. y tb. adv.	col.	Nada
pepino				Cosa insignificante o de poco o ningún valor + ejemplo con *importar*	*[importar] un / tres pepino(s)*	loc. adv. y tb. pron.	col.	Nada
un pijo	loc. pron. y tb. adv.	coloq.	Ej. con *importar*					
importar, o no importar, algo un pimiento	locs. verbs.	coloqs.	Importar poco o nada		*[importar] un pimiento*	loc. pron. y tb. adv.	col.	Nada
					[importar] un pitoche	loc. adv.	col. rara	Un pito / nada
un/tres pito(s)	loc. adv.	coloq.	Ej. con *importar*		*[importar] un / tres pitos*	loc. adv. y tb. pron.	col.	Nada

					[importar] tres puñetas	loc. adv. y tb. pron.	vulg.	Nada
importar, o no importar, algo un rábano	locs. verbs.	coloqs.	Importar poco o nada		*[importar] un rábano*	loc. adv. y tb. pron.	col.	Nada

Cuadro 1: Relación de expresiones del tipo *importar* + cuantificador + sustantivo registradas en el *DLE* y el *DFDEA*

El *DFDEA* se centra en el español de España, por lo que la comparación con el *DLE* no puede ser completa, aunque, como ya se ha señalado, la única expresión que lleva marca geográfica en el *DLE* es *importar madre*(*s*). Una comparación más adecuada en lo referido al inventario en todo el dominio hispánico podría hacerse con las variantes recogidas en el proyecto VARILEX. El resultado, que aparece en el cuadro 2, no es, sin embargo, tan amplio como se podría esperar. Las personas encuestadas por VARILEX mencionan únicamente las combinaciones formadas con *madre*(*s*), *ajo*, *bledo*, *pepino*, *pito* y *rábano*[6]. Sorprende lo escaso de las variantes registradas en una expresión tan coloquial como la analizada y especialmente la ausencia de *carajo*, *cojones*, *cuerno*, *huevo*, *mierda*, *narices* o *pimiento*, todas ellas bastante extendidas en muchos países.

En los datos de VARILEX se observa que casi todas las expresiones registradas se dan en todos o la mayor parte de los países hispánicos. En los datos originales, únicamente *no importar(le) madre(s)* se restringe a México, Cuba y España[7]. Mucho más variado, como es natural, resulta el panorama que se obtiene del *Diccionario de americanismos* (*DAm*). En esta obra, que pretende recoger palabras y significados que se dan en los países hispánicos y, según el *DLE*, no tienen uso general, las expresiones que nos interesan están todas ellas agrupadas en la entrada *importar*. Se define *importar un cacahuate* como 'no dar valor o importancia a alguien o algo' y todas las demás remiten a ella. El cuadro 2 contiene la relación de las expresiones recogidas e indica los países en los que los redactores han podido registrarlas. Se observa con claridad que las incluidas corresponden a un número muy limitado de países (uno en muchos casos), que es lo esperable, puesto que los usos más generalizados deberían figurar ya en

[6] Es la pregunta F127 (*not to give a damn*). Las respuestas mencionadas son las correspondientes a los datos recogidos inicialmente. Los revisados por los expertos añaden *comino* (en casi todos los países) y *verga* (en Chile). Hay también algunos cambios en la adscripción a países (*cf. infra*, nota 7). El *Diccionario panhispánico Varilex* (*DPV*) reproduce los datos iniciales, aunque también con cambios en algunos países.

[7] Por supuesto, no todas las variantes se registran en todos los países, pero sí en la mayoría. De todas formas, es necesario tener en cuenta las modificaciones que se han dado en las diferentes fases del proyecto. *No importar(le) madre(s)* se registra en España, Cuba y México en los datos originales; en los revisados por los expertos desaparecen España y Cuba, pero se añade Nicaragua. En el *DPV*, a los países que figuran en los datos originales se añade Estados Unidos, donde no se pasaron cuestionarios inicialmente.

el *DLE*. Destaca la ausencia de (*no*) *importar madre(s)*, recogido en el *DLE* y VARILEX y que, por supuesto, figura en el *Diccionario del español de México* (*DEM*)[8]. Sí aparece, en cambio, *valer madre*, que el *DAm* define como 'no importar nada algo que concierne a alguien'[9].

El cuadro 2 incluye también la indicación de presencia (o ausencia) de cada una de estas expresiones en la versión 0.92 del Corpus del Español del Siglo xxi (CORPES XXI), que tiene algo más de 312 millones de formas ortográficas, y el Corpus del Español Web / Dialectos (CdEweb), con 2000 millones de formas. En ambos corpus se indica el número de casos registrados para esas expresiones y los países en los que se documentan cuando son solo uno o dos y su número cuando son más. Trabajar con corpus supone manejar una documentación real muy superior a la utilizada tradicionalmente en la confección de diccionarios como el *DLE* o el *DFDEA* y un procedimiento distinto de la respuesta individual a un cuestionario, como el empleado en VARILEX. No sorprende, por tanto, que los resultados coincidan en buena parte con los examinados anteriormente, pero presenten también algunas diferencias de interés, como vamos a ver a continuación.

Destaca, en primer lugar, la falta de documentación de algunas variantes registradas en el *DLE (caracol, charra, clavo, paja)*, el *DFDEA (pitoche)* o el *DAm (pitajaya)*. Naturalmente, eso no significa que no existan: los corpus son una muestra que aspira a ser representativa, pero no pueden contener todo lo que es posible en una lengua. De todas formas, la ausencia de una determinada combinación en conjuntos que constan de cientos o miles de millones de formas debería ser tomada en cuenta a la hora de decidir si se incluye o no una cierta expresión en un diccionario. La indicación de la frecuencia general y también de la registrada en diferentes tipos de texto (países, ámbito de uso, etc.) es la gran ventaja que proporcionan los corpus y la que permite complementar o matizar la información que figura en los diccionarios. Conjuntando los datos del CORPES y del CdEweb sabemos que las construcciones con *bledo, carajo, comino, mierda, pepino y pito* son las más frecuentes (más de 500 casos en el CdEweb y presencia en todos o casi todos los países), seguidas por *pimiento* y *rábano* (más de 300 casos en el CdEweb). En el caso de *importar un culo*, el *DAm* lo sitúa en la República Dominicana, Colombia y Bolivia, pero el CORPES localiza en Colombia los 18 casos documentados de esta construcción, mientras que el CdEweb lo registra en 13 países un total de 65 ocasiones, el 55 % de las

[8] *Valer* o *importar algo madre(s)*, *valerle* o *importarle madre(s) algo a uno* Tener muy poco valor o ninguno; no importarle a uno en absoluto o no ser de su incumbencia: "*Me importa madre* que tú ya no me quieras", "Tú, tus millones y tus tías *me valen madres*", "Y dile que ni se meta. ¡A él *le vale madres*!" (*DEM, s.v. madre*).

[9] La lista de expresiones con *valer* es mucho más amplia que la registrada para *importar*. La mayor parte de ellas son definidas mediante *importar*: *no importar nada*, *no importar algo a alguien*, *no importar nada algo que afecta a alguien*, etc.

cuales corresponden a Colombia[10]. Otro dato de interés se da en *importar un pimiento* (no recogido en VARILEX ni, por su caracterización general, en el DAm). El CORPES registra 63 casos, pero 58 de ellos (el 92 %) proceden de España y los 5 restantes corresponden a 5 países distintos. Bastante más extensa es la distribución que muestra en el CdEweb: 431 casos en 16 países distintos, de los cuales 343 (el 79,6 %) corresponden a España.

	DLE	**DFDEA**	**DAm**	**Varilex**	**CORPES**		**CdEweb**	
ajo				Es.			1	EU
ardite	(+)	+			3	Es, Bo.	15	[6]
bledo	(+)	+		+	195	[19]	1349	[21]
cacahuate	?		Mx, Ho, Ni.		2	Mx.	33	[7]
cachinflón			Ho, Ni.					
callampa			Ch.					
caracol	+							
carajo	(+)	+			184	[19]	1204	[21]
carallo		+					4	Ar, Es.
charra	+							
chita	+							
chorizo			Co.				4	Co.
cinco			Co.					
clavo	+							
cojón		+			14	Es + 2	15	Es + 3
comino	+	+		(+)	131	[15]	829	[21]
corno			Bo, Py, Ar, Uy.				10	Ar + 2
cuerno	+	+			15	[7]	57	[13]
cuesco			Ch.		3	Ch.		
cuete			Ch.				1	Bo.
culo			RD, Co, Bo.		18	Co.	65	[13]
higa		+					51	Es + 5
higo (seco)		+			2	Co, Uy.	6	[3]
huevo		+			27	[7]	138	[16]

[10] Destaca también Argentina, con 8 casos según el cuadro inicial, pero las concordancias proporcionan únicamente 6 y tres de ellos son el mismo texto en diferentes publicaciones periódicas. Por tanto, son solo 4 casos, aunque es importante tener en cuenta que este país no figura entre los registrados en el *DAm* ni en VARILEX y tampoco hay casos en el CORPES.

madre(s)	Mx.			Es, Cu, Mx.	6	Mx, Es.	12	Mx, EU.
maní			Ch.				2	Ch.
mierda		+			127	[13]	678	[20]
narices		+					10	[4]
paja	+							
papa			PR				35	[12]
pelotas		+					8	[5]
pepino	(+)	+		+	55	[14]	684	[21]
pijo	(+)						23	Es.
pimiento	+	+			63	Es + 5	464	Es + 15
pinga			Cu.				3	[3]
pitajaya			Bo.					
pito	(+)	+		+	105	[18]	622	[21]
pitoche		+						
puñetas		+					1	Es.
rábano	+	+		+	69	[11]	327	[18]
verga				Mx, Pn.			4	[3]

Cuadro 2: Expresiones del tipo *importar un* + sustantivo registradas en algunos diccionarios y corpus.

3. IDENTIFICACIÓN Y RECUPERACIÓN DE VARIANTES

Hay un segundo aspecto en el cual la utilización de corpus de referencia puede ser de gran utilidad en las investigaciones fraseológicas. Dado que los corpus que empleamos están lematizados y anotados morfosintácticamente, es sencillo recuperar todos aquellos ejemplos que responden a un cierto esquema constructivo. En el caso que nos ocupa, el esquema es del tipo IMPORTAR + ARTÍCULO INDETERMINADO / CUANTIFICADOR + SUSTANTIVO, esto es, cualquier forma del paradigma del verbo *importar* seguida del artículo indeterminado o de un elemento cuantificador y luego un sustantivo cualquiera. Aparecen así todos los casos del tipo *importar un pepino*, *importar tres pepinos*, etc. En búsquedas tan amplias como esta, en las que se trata, precisamente, de localizar las variantes, se requiere un cierto trabajo de manipulación de los ejemplos recuperados. En el CORPES, por ejemplo, las concordancias aparecen ordenadas inicialmente por el año del texto, pero es sencillo reordenarlas en función del segundo lema por la derecha, con lo que aparecerán seguidos todos los casos con *bledo*, *pimiento*, etc. y el recuento y el análisis de la distribución se hace con facilidad. En el CdEweb se obtienen de una vez las expresiones que responden a cada variante y su frecuencia, pero diferenciadas según la forma del verbo

y también del sustantivo, de modo que *importa un pimiento*, *importaba un pimiento*, *importaba dos pimientos*, etc. están en líneas distintas y tienen que ser agrupadas para hallar los totales.

Es necesario tener en cuenta que este procedimiento recupera los casos de una determinada construcción, pero eso no garantiza que todas ellas tengan el significado que nos interesa. Con un ejemplo claro, *estar hasta las narices* y *estar hasta la bandera* responden al mismo esquema constructivo, pero tienen significados diferentes y no pueden ser considerados variantes[11]. En el caso de *importar*, además, tenemos el problema añadido de que el verbo tiene otra acepción que es compatible también con sustantivos como *pimiento*, *higo*, *comino*, etc. y, por supuesto, muchos otros sustantivos que pueden aparecer en ese esquema. El análisis del contenido de un corpus no proporciona automáticamente, por tanto, la lista de variantes de una determinada UF. Permite, eso sí, identificar candidatas, analizar sus frecuencias, estudiar su distribución, controlar su significado… Con todo ello es evidente que podemos disponer de una relación más amplia de variantes y de mayor información acerca de sus condiciones de uso.

En el caso que nos ocupa, la versión 0.92 del CORPES contiene, además de los ya incluidos en el cuadro 1, los siguientes: *batata*, *belín*, *bleda*, *cañoto*, *cero a la izquierda*, *chingada*, *coño*, *corajo*, *diablo*, *guañamo*, *güevá*, *huea*, *joraca*, *jota*, *leche*, *nabo*, *pico*, *pincho*, *rabino*, *rabito*, *raja*, *repepino*, *sorete*, *soto*. Es cierto que la mayoría de estos términos aparecen en la expresión solo una o dos veces, así que no se puede excluir siquiera la opción de que no sean expresiones de frecuencia muy baja, sino creaciones ocasionales. Los únicos destacables son *coño* (8 casos, en 6 países distintos) y *raja* (7 casos, localizados en Chile todos ellos)[12]. El cuadro proporciona la relación de sustantivos que el CORPES documenta en esta expresión, con indicación de su frecuencia general y los países en los que se localiza mayoritariamente cada una de ellas.

Sustantivo	**Frec.**	**Distribución geográfica**
ardite	3	2 Esp. 1 Bolivia
batata	1	Esp.
belín	1	Arg.
bleda	1	Esp.
bledo	195	19 países. 50 % de los casos en Esp.
cacahuate	2	Méx.
cañota	1	Esp.

[11] Está claro también que el mismo problema aparece en las listas ordenadas alfabéticamente que se incluyen en, por ejemplo, el *DFDEA*.

[12] Esta expresión no figura en la selección incluida en las *640 frases que caracterizan a los chilenos*, editada por la Academia Chilena de la Lengua.

carajo	184	Más frec. en Perú, Argentina, Panamá, Puerto Rico y Venezuela
cero a la izquierda	1	Esp.
chingada	4	Méx.
chorizo	1	Colombia
comino	131	Colombia, Perú, Paraguay, Cuba, El Salvador
coño	8	Ch, Esp, Pn, PR, RD, Ven.
corajo	1	Chile
corno	1	Urug.
cuerno	15	Río de la Plata, Esp.
cuesco	2	Chile
culo	18	Colombia
diablo	1	Méx.
guañano	1	Chile
güevá	1	Chile
higo	2	Col, Uy
huea	1	Chile
huevo	27	Perú, España, Paraguay y 4 más
joraca	1	Urug.
jota	1	Esp.
leche	1	Esp.
madre	6	Méx. (5) El Salvador (1)
mierda	127	Esp (80 % de los casos) y 12 países más
nabo	1	Esp.
pepino	55	Perú, EE.UU, El Salvador y otros más. España, el más bajo
pico	2	Chile
pijo	2	Esp.
pimiento	63	Esp. 56. 1 caso en otros cinco países
pincho	1	Perú
pito	105	Más frec. en Guat. Perú, Ecuador, Argentina
rábano	69	Más en Bolivia, Chile, Perú, España, Paraguay
rabanito	1	Arg.
rabino	1	Arg.
rabito	1	Arg.
raja	8	Chile
repepino	1	Perú
sorete	1	Arg.
soto	1	Arg.

Cuadro 3: Frecuencia y distribución de expresiones del tipo *importar un* + sustantivo en el CORPES.

4. CONCLUSIÓN

Las páginas anteriores constituyen un intento, sin duda muy superficial, de mostrar el modo en que el análisis de los corpus textuales puede contribuir a mejorar nuestro conocimiento acerca de las unidades fraseológicas, un caso concreto en el terreno mucho más amplio de la colaboración entre las investigaciones lexicográficas y el procesamiento del lenguaje natural. La utilización de los datos incluidos en los corpus que tenemos a nuestra disposición no solucionan los diversos problemas teóricos que plantea el reconocimiento de las UF y su clasificación, pero puede, sin duda, realizar una aportación crucial a la determinación de sus condiciones de uso y la identificación de variantes.

CORPUS Y OTROS RECURSOS ELECTRÓNICOS MENCIONADOS EN EL TEXTO

Brown Corpus. The Standard Corpus of Present-Day Edited American English. Dirs. W. Nelson Francis y Henry Kučera. http://www.helsinki.fi/varieng/CoRD/corpora/BROWN/.

CdEweb: Corpus del Español (Web / Dialectos). Dir. Mark Davies. https://www.corpusdelespanol.org/web-dial/.

CORPES. Real Academia Española: *Corpus del español del siglo xxi*. http://rae.es/recursos/banco-de-datos/corpes-xxi.

CREA. Real Academia Española: *Corpus de referencia del español actual*. http://rae.es/recursos/banco-de-datos/crea.

DAm: Asociación de Academias de la lengua española (ASALE): *Diccionario de americanismos*. http://lema.rae.es/damer/.

DLE: Real Academia Española y Asociación de Academias de la lengua española. *Diccionario de la lengua española* (https://dle.rae.es/).

DPV: *Diccionario Pahnispánico Varilex*. https://www.scribd.com/document/294524780/Diccionario-Panhispanico-VARILEX.

Es-Ten-Ten: Spanish Web Corpus. https://www.sketchengine.eu/estenten-spanish-corpus/.

LOB. Lancaster-Oslo/Bergen Corpus. http://www.helsinki.fi/varieng/CoRD/corpora/LOB/.

SEU. Survey of English Usage. Dir. Randolph Quirk. http://www.ucl.ac.uk/english-usage/index.htm.

Varilex: Variación Léxica en Español del Mundo. Dir. Hiroto Ueda. https://lecture.ecc.u-tokyo.ac.jp/~cueda/varilex-r/.

REFERENCIAS BIBLIOGRÁFICAS

Academia Chilena de la Lengua. 2015. *640 frases que caracterizan a los chilenos*. Santiago de Chile: Uqbar.

DEA. Seco, Manuel, Olimpia Andrés y Gabino Ramos. 1999. *Diccionario del español actual*. Madrid: Aguilar.

Gantar, Polona, Lut Colman, Carla Parra Escartín, Héctor Martínez Alonso. 2019. Multiword expressions: Between Lexicography and NLP. *International Journal of Lexicography* 32/2: 138-162.

LEMA. 2001. *Diccionario de la lengua española*. Dir. Paz Battaner. Barcelona: Vox.

Rundell, Michael. 2018. Searching for extended units of meaning –and what to do when you find them. *Lexicography. Journal of ASIALEX*, marzo 2018, pp. 1-17.

Sinclair, John. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.