

Un estudio sobre la frecuencia de uso de locuciones

TSUTAHARA Ryo

1. Introducción

La pregunta de investigación que orienta este estudio es: ¿Cuáles son las locuciones más frecuentes en español? Con esto pretendemos presentar una lista de las locuciones por orden de frecuencia.

El número de palabras que existen en una lengua natural es enorme y, además, se siguen creando nuevas palabras de forma constante. Sin embargo, la mayor parte de la actividad lingüística está cubierta por las palabras más frecuentes. En el caso del español, las 5,000 palabras más frecuentes constituyen aproximadamente el 90% del corpus *Spanish Web 2018* (esTenTen). A continuación, se muestra una tabla que indica el porcentaje del corpus cubierto por las 5,000 palabras más frecuentes.¹

Tabla 1. Número de las palabras más frecuentes y su cobertura del corpus

Las X palabras más frecuentes	Cobertura del corpus	La Xª palabra
Las 100 palabras más frecuentes	53.32 %	<i>servicio</i>
Las 200 palabras más frecuentes	58.64 %	<i>relación</i>
Las 500 palabras más frecuentes	67.14 %	<i>sacar</i>
Las 1000 palabras más frecuentes	74.49 %	<i>facilitar</i>
Las 1500 palabras más frecuentes	78.86 %	<i>suyo</i>
Las 2000 palabras más frecuentes	81.76 %	<i>impresión</i>
Las 2500 palabras más frecuentes	83.84 %	<i>tratado</i>
Las 5000 palabras más frecuentes	89.22 %	<i>preocupante</i>

Así, el español depende fuertemente de las palabras básicas y existe un amplio debate sobre sus usos y los métodos de enseñanza en la educación de lenguas extranjeras, así como en la investigación básica en lexicología. (Nation 2001, Davies & Face 2006, Schmitt & Schmitt 2020, etc.)

Por esta razón, este estudio se centra en las locuciones del español, que son un subtipo de unidades pluriverbales (UPV, en adelante). Estas UPV son secuencias de palabras lexicalizadas que funcionan como una unidad y como un hiperónimo para locuciones, colocaciones, perífrasis y modismos, entre otros.

¹ Desarrollamos la lista de las palabras singulares por orden de frecuencia utilizando la función analítica Wordlist del corpus. Esta función dispone una lista de lemas en el corpus por orden de frecuencia. Por eso, la lista obtenida directamente del corpus contiene signos que no son unidades léxicas como números arábigos, etc. Descartamos manualmente tales elementos y elaboramos una lista constituida solo por entidades léxicas.

Las UPV han recibido menos atención que las palabras singulares en el campo de la enseñanza de ELE. No obstante, en las últimas décadas han aumentado los estudios que coinciden en defender su importancia didáctica (Meunier & Granger (eds) 2008 y Leal Riol 2011). Por ejemplo, Christiansen & Arnon (2017: 542) explican lo siguiente:

The new millennium, however, has seen a shift toward construing multiword linguistic units not as linguistic rarities, but as important building blocks for language acquisition and processing.

Esos estudios defienden la importancia de las UPV en la enseñanza del español como lengua extranjera, pues su uso adecuado aumenta significativamente la fluidez de los estudiantes. Altenberg (1998) y Erman & Warren (2000) demuestran estadísticamente que las UPV forman parte de la gran mayoría de la actividad lingüística; mientras que Martínez & Murphy (2011), Moreno (2012) Siyanova-Chanturia & Spina (2020) enfatizan la dificultad en su aprendizaje. Según las investigaciones anteriores, lo que dificulta la comprensión y el uso de las UPV es que, sin una instrucción adecuada, los alumnos tienden a no darse cuenta de que una secuencia de palabras forma una unidad léxica única. Por ejemplo, la siguiente frase es difícil de interpretar para los estudiantes de español como lengua extranjera incluso con la ayuda de un diccionario:

- (1) Aprende a gestionar estas situaciones de forma que lleves a tu equipo y a tu empresa al éxito.²

La frase no puede interpretarse correctamente aunque se busque cada palabra en el diccionario. Esto se debe a que *de forma que* es una locución, cuyo significado o función no es equivalente a la suma de los significados de las palabras individuales *de*, *forma* y *que*. Asimismo, como las UPV suelen estar compuestas por palabras muy frecuentes en el español, los estudiantes pueden considerar erradamente que comprenden secuencias como *de forma que* sin darse cuenta de que se trata de una locución.³ Por esta razón, al igual que la gramática y las palabras complicadas, también es necesaria una instrucción explícita sobre las UPV.

² Los ejemplos en este estudio son citados del corpus esTenTen.

³ Martínez & Murphy (2011) detallan este aspecto. Este estudio tiene como participantes a 101 estudiantes brasileños de inglés. Los autores les muestran oraciones que incluyen UPV y oraciones que no las contienen para examinar la dificultad de entender las UPV y si eran capaces de detectar la presencia de UPV. El estudio confirmó que las oraciones con UPV se entendieron con menos precisión en comparación. Además, se observó la tendencia a que los informantes consideraran que podían entender correctamente las oraciones mostradas a pesar de que se equivocaban en realidad. Según el análisis esto se debe a que los informantes no se daban cuenta de las UPV. Los autores analizan que a los estudiantes no se les ocurre la posibilidad de que la secuencia ante ellos fuera UPV puesto que las UPV más frecuentes suelen ser formadas solo por las palabras más básicas.

1.2. Estudios previos y estado actual

La importancia y la dificultad de la enseñanza de las UPV está ampliamente reconocida en el campo de la enseñanza de idiomas extranjeros. Sin embargo, la enseñanza de las UPV en español todavía está en sus primeras etapas aunque ya existen algunos estudios teóricos sobre ellas (Mendivil Giró 1990, Koike 2001, Alonso Ramos 2004, García Fernández 2006, etc.). En japonés, solo hay un libro de texto que trata las UPV (Takahashi & Ito 2013); mientras que en español (De Prada et al. 2012) o en inglés (Weiber 2004, Muñoz Basols et al. 2014 y Bregstein 2016) los materiales no son muy numerosos. Además, todos estos libros están diseñados para estudiantes de niveles intermedios o avanzados y no existen equivalentes para los niveles principiantes. Por ejemplo, en De Prada et al. (2012) *Uso interactivo del vocabulario* (para nivel B2-C2) se tratan colocaciones aparte de las palabras singulares pero en Encinar (2012) *Uso interactivo del vocabulario* (para nivel A1-B1) solo se tratan palabras singulares. Aparte, en la introducción del libro de Muñoz-Basols et al. (2014) *Spanish Idioms in Practice* se aclara que este material es para estudiantes de nivel intermedio y avanzado.

Además, a veces se encuentran discrepancias entre el uso real de la lengua y la selección de UPV en los libros de texto. Algunos estudios previos señalan problemas con el tratamiento de las UPV en libros de texto; como la selección de las UPV basada en juicios subjetivos de los autores, la carencia de criterios de selección verdaderamente útiles y la inserción de elementos de utilidad limitada (cf. Koprowski (2005: 330) y Martínez (2013: 185)). Por ejemplo, en el capítulo de preposiciones de Bregstein (2016) se incluye una lista de locuciones preposicionales pero no se encuentra la locución preposicional más frecuente en español, *a través de*, como mostraremos más adelante. Por otro lado, Weiber (2004) y Muñoz-Basols et al. (2014), prestan atención a las locuciones, pero solo tratan aquellas de carácter léxico y no incluyen las gramaticales, como las locuciones preposicionales y conjuntivas. *A través de*, tampoco aparece en estos materiales.⁴

Estos problemas pueden deberse a que se haya investigado poco sobre la frecuencia de uso de las UPV. En otros términos, todavía no sabemos qué UPV son de mayor importancia didáctica. Existen estudios que tratan la frecuencia de uso de las palabras singulares no solo en el dominio de la enseñanza de inglés (*General Service List*) sino también en la de español (Davies & Hayward Davies 2018) y estos estudios permiten una enseñanza eficaz y sistemática del vocabulario. No obstante, hasta donde sabemos no hay estudios de ELE al respecto. Para una enseñanza práctica de locuciones, debería elaborarse una lista de locuciones por orden de frecuencia.

Una de las razones de la ausencia de una lista de las UPV más frecuentes, a pesar de su

⁴ En general, las locuciones gramaticales reciben menos atención. Por ejemplo, Pérez Cinco (2000) presta atención al uso de las preposiciones, pero las locuciones preposicionales solo se presentan en una lista que ocupa dos páginas.

reconocida importancia, puede deberse a que es más difícil identificar la frecuencia de uso de una UPV que la de una palabra. La razón principal de la dificultad para medir con precisión la frecuencia de uso de las UPV es que existen secuencias de palabras que pueden ser una UPV o una mera secuencia de las palabras dependiendo del contexto y esto complica la medición de su frecuencia de uso a través de los corpus. Por ejemplo, si se quiere saber la frecuencia de la locución *de nuevo*, lo primero que debe hacerse es analizarla con Concordance. Según Concordance de esTenTen, en el corpus hay 3,849,309 ocurrencias de *de nuevo*. No obstante, este número no equivale a la frecuencia de uso de la locución porque los casos como (2) están incluidos.

(2) Programa de actos de Nuevo León

No hace falta decir que aquí *de* y *nuevo* no forman una locución. En este caso, las dos palabras están alineadas por pura casualidad. En adelante, llamaremos “ruidos” a este tipo de secuencias de palabras. La posibilidad de que Concordance también contenga ruidos dificulta la medición de la frecuencia del uso de locuciones.

En el ámbito de la enseñanza del inglés, Martínez (2011) aborda, en su tesis doctoral, esta cuestión y extrae de un corpus las UPV más frecuentes en inglés y presenta una lista de las UPV más importantes para los aprendices (*PHASE List*). Martínez (2011) extrajo los n-grams (de 2-4 grams) de *British National Corpus* por orden de frecuencia y, posteriormente, especificó las UPV eliminando combinaciones libres. Después, analizó manualmente cada una de las UPV con Concordance y contó el número de ruidos entre los casos analizados. El autor estima la frecuencia real de cada UPV a partir del número de los ruidos y mostró una lista de las UPV más frecuentes. En este artículo seguiremos y modificaremos esa metodología para componer una lista de las locuciones más frecuentes en español.

1.3. Objetivo de este estudio

Siguiendo la metodología de Martínez (2011) intentamos elaborar una lista de las locuciones más frecuentes en español para los estudiantes de nivel principiante e intermedio. Según nuestra lista de las palabras singulares más frecuentes que presentamos en la sección 1., *preocupante* es la 5,000ª palabra. Basándonos en este criterio, abordamos como pregunta de investigación: ¿cuáles locuciones son más frecuentes que la palabra *preocupante*?

1.3.1. Locución y UPV

Como se ha mencionado en la primera parte del artículo, UPV es un término genérico para las locuciones, colocaciones, y etc. Entre ellos, analizaremos solo las locuciones puesto que como mencionaremos después, la especificación de la frecuencia de las locuciones se realiza de manera

totalmente distinta de la de las colocaciones.

La NGLE define la locución como siguiente;

Se llaman LOCUCIONES los grupos de palabras LEXICALIZADOS (en el sentido de ya formados e incluidos en el diccionario) que constituyen una sola pieza léxica y ejercen la misma función sintáctica que la categoría que les da nombre. Así, las LOCUCIONES ADJETIVAS equivalen, con diversos grados de proximidad, a los adjetivos (*de cuidado* viene a significar 'peligroso'); las LOCUCIONES PREPOSICIONALES equivalen a las preposiciones (*a bordo de*, *se acerca a sobre*, o *a en* en ciertos contextos); las LOCUCIONES ADVERBIALES equivalen a los adverbios (*a rajatabla* se aproxima a *estrictamente*), y las LOCUCIONES CONJUNTIVAS equivalen a las conjunciones (*si bien* alterna con *aunque*).

(NGLE 1.10a)

Su rasgo característico es la fijación de las palabras constituyentes, lo que significa que no es posible sustituir una palabra por otra sin perder su significado idiomático. Por ejemplo, la NGLE (34.11j) cita los casos de *tomarle el pelo* (locución) y *tomarle el cabello* (combinación libre) en los que la segunda expresión no tiene sentido de burlar a alguien, a pesar de que el significado de cada palabra es casi igual. Además, siendo una locución, la pronominalización de *pelo* tampoco es posible como en **tomárselo*.

Eliminamos las colocaciones de este trabajo debido a la diferencia de fijación. Las colocaciones, en general, permiten la sustitución (*tomar café* puede parafrasearse como *beber café*) y pronominalización por su bajo grado de fijación (*tomarlo*). Debido a este menor grado de fijación, el método de analizar la frecuencia de las colocaciones es totalmente diferente del de las locuciones. Por esta razón, nos centraremos en las locuciones.

2. Metodología

En este trabajo especificaremos las locuciones más frecuentes siguiendo la metodología de Martínez (2011), en otras palabras, primero extraeremos los n-grams más frecuentes del corpus y de ellos sacamos las locuciones. El corpus utilizado en este estudio es esTenTen, publicado en Sketch Engine y comprende 19,600,041,678 tokens. El corpus está formado por variados tipos de documentos de la web como página web de empresa, documentos administrativos, Wikipedia, blogs personales, etc. Debido a su extensión, consideramos este corpus uno de los más robustos para describir las tendencias generales del uso de la locución en el español moderno.

Los corpus publicados en Sketch Engine cuentan con una función analítica N-grams. Esta función permite extraer 2-6-grams en el corpus por orden de frecuencia. En primer lugar, utilizamos

esta función para crear una lista de los n-grams más frecuentes en el corpus.⁵ A continuación presentamos una lista parcial de n-grams.

N-GRAMS Spanish Web 2018 (esTenTen18) Q i

2-6-grams, lemma (items: 39,585,429 , total frequency: 1,781,404,576)

Word	Frequency ?	Word	Frequency ?
1 de el	15,148,080 ...	11 de uno	2,197,166 ...
2 en el	10,253,507 ...	12 ser uno	2,068,516 ...
3 a el	5,473,832 ...	13 para el	1,887,045 ...
4 el que	3,756,623 ...	14 todo el	1,756,539 ...
5 con el	3,333,976 ...	15 en uno	1,494,909 ...
6 y el	3,153,014 ...	16 que ser	1,310,707 ...
7 que el	3,047,745 ...	17 de su	1,240,369 ...
8 por el	2,948,625 ...	18 de este	1,236,470 ...
9 ser el	2,383,170 ...	19 que no	1,204,169 ...
10 que se	2,370,610 ...	20 con uno	1,182,995 ...

Figura 1. Análisis con N-grams

De esta muestra se desprende que la gran mayoría de los n-grams no son locuciones, sino combinaciones libres: *de el*, *en el*, *a el*, etc.⁶ Entre las secuencias mostradas en la imagen, solo *el que* se considera una locución. Por lo tanto, es necesario realizar la tarea de extracción manual de las locuciones a partir de este enorme número de n-grams.

2.1. Criterio para determinar una locución

Aunque existen varios estudios previos sobre la definición de la locución, la decisión de si considerar una secuencia como locución depende de cada lexicógrafo. Es el investigador quien determina si una combinación de palabras es realmente una locución. Este problema es inevitable en la investigación de la UPV y es uno de los factores que dificultan la investigación en esta área. En primer lugar, utilizamos un diccionario como criterio para distinguir entre una secuencia libre y una

⁵ En el estatus original, solo se dispone una lista de 1000 ngrams más frecuentes. Para obtener una lista exhaustiva se requiere un contrato.

⁶ Las entidades en las listas realizadas por los corpus de Sketch Engine están lematizados. Por lo tanto, *de el* en la lista es lema y en él todas las formas *del*, *de la*, *de los* y *de las* están incluidas.

locución: si una secuencia aparece como una locución en el DLE o en el WSE, entonces se trata como tal. Algunas de las secuencias que analizamos durante el estudio parecían ser locuciones, pero no se tratan así en los diccionarios. Para definir tales secuencias, analizamos si cumplen los criterios para ser locuciones establecidos por Martínez (2011:68-69): 1. *Is the expression a Morpheme Equivalent Unit?* 2. *Is the expression semantically transparent?* 3. *Is the expression potentially "deceptively transparent"?* De esta manera, incluimos en la lista solamente aquellas secuencias que cumplieren con al menos una de estas condiciones para ser entendidas como locuciones. Esto dio lugar al establecimiento de dos categorías sintácticas nuevas, producto de esta investigación que no se encontraban en los diccionarios: auxiliar y afijal. La primera funciona como equivalente a un auxiliar singular, la segunda como un afijo.

Las locuciones auxiliares corresponden a *tener que*, *ir a*, etc. A estas siguen los infinitivos para expresar obligación o futuro. Estos significados no son la suma de los significados de los constituyentes y cumplen las condiciones de ser locución. Lo que llamamos “locuciones auxiliares” suelen tratarse como auxiliares junto con los auxiliares singulares como *poder* y *querer* en los estudios de las perífrasis verbales. Por ejemplo, Gómez Torrego (1999) defiende lo siguiente:

Una perífrasis verbal es la unión de dos o más verbos que constituyen un solo 'núcleo' del predicado. El primer verbo, llamado 'auxiliar', comporta las informaciones morfológicas de número y persona, y se conjuga en todas (o en parte de) las formas o tiempos de la conjugación. El segundo verbo, llamado 'principal' o 'auxiliado', debe aparecer en infinitivo, gerundio o participio, es decir, en una forma no personal.

(Gómez Torrego 1999: 3325)

En esta obra junto con los auxiliares singulares como *poder*, las combinaciones como *tener que* se tratan como auxiliares. De este modo, la categoría de locución auxiliar no entra en contradicción con los estudios anteriores sobre las perífrasis verbales. En la categoría que hemos llamado “locución afijal”, entran las secuencias *de manera*, *de forma* y *de carácter*. Estas secuencias nunca se utilizan solas. Siempre deben ser modificadas por un adjetivo. Véanse los siguientes ejemplos:

- (3) Se requiere disponibilidad para incorporarse de manera inmediata.
- (4) Para ellos, no existe el silencio y el ruido llega a alterar de forma considerable su calidad de vida.
- (5) Sin embargo, estas divisiones son de carácter meramente táctico.

La construcción *de manera/forma* + adjetivo es casi sinónimo de un adverbio con un afijo

-mente: ej.: *de manera inmediata e inmediatamente*. Así, *de forma/manera* son equivalentes a *-mente* y pueden considerarse que funcionan como un afijo equivalente de dos palabras. En cuanto a *de carácter*, sirve como un afijo para formar adjetivos añadiendo un matiz perifrástico, o formal: *tácito* y *de carácter tácito*. De esta manera, las tres secuencias cumplen el “core criteria” de Martínez (2011): ser equivalente de un morfema y las tratamos como locuciones.^{7,8}

2.2. Frecuencia estimada

El siguiente paso es identificar la frecuencia concreta de estas locuciones en el corpus. La lista de n-grams se extrajo de un corpus reducido a 1/16, que es diferente de la frecuencia de uso en el corpus entero. La identificación de la frecuencia real en el corpus se obtiene por analizar cada locución con Concordance.

Después eliminamos los ruidos, que producen la dificultad de la extracción automática de las UPV y, en particular, la dificultad para identificar la frecuencia de uso. El ruido se refiere aquí a la inclusión de secuencias como *proyecto de Nuevo León* al analizar la frecuencia de la locución *de nuevo*, como se describe en la sección 1.2.

Para eliminar este ruido, tomamos los medios de observación manual de los ejemplos. En primer lugar, extrajimos al azar 100 ejemplos de cada locución con Concordance y contamos manualmente cuántos de ellos se utilizan como locución. De esta manera se aclara la tasa de locución. A continuación, se multiplica la frecuencia de las locuciones de la concordancia por la tasa de locución para obtener la frecuencia estimada, sin el ruido del corpus. En el caso de *de nuevo*, por ejemplo, hay 3,849,309 ocurrencias en el corpus. Sin embargo, entre los cien casos aleatoriamente extraídos, *de nuevo* como locución solo se encuentra en 43 casos. Entonces 0.43 es la tasa de locución de *de nuevo*. Multiplicando el número de ocurrencia 3,849,309 por la tasa de locución 0.43, se estima la frecuencia de uso de la locución de nuevo: 1,655,203 veces. El objetivo de este estudio es presentar una lista de locuciones con su frecuencia de uso, y nótese que la frecuencia de uso adjunta a la lista es la frecuencia de uso estimada.

3. Lista simplificada de las locuciones más frecuentes

Aquí presentamos una lista de las locuciones con una frecuencia de uso estimada superior a la de la palabra singular número 5,000 en frecuencia.

⁷ En lugar de adjetivos, los artículos y los demostrativos también pueden modificar *de manera/forma*. Y, por supuesto, se antepone al sustantivo: *de esta manera/de una forma*, etc. Debido a esta posibilidad de insertar determinantes, el grado de fijación de estas secuencias es más bajo que el de las locuciones puras. Además, nótese que la secuencia “*de + determinante + manera*” se diferencia de “*de manera + adjetivo*”, ya que nuestro análisis se basa en la extracción automática de N-grams y estas frases se tratan como diferentes por parte del corpus desde el principio.

⁸ Aunque los autores no emplean el término “afijal”, en el diccionario de locuciones, Takahashi & Ito (2013) incluyen entradas de *de forma* y *de manera*.

1	el (la, lo, los, las) que	63,992,369	80	a la vez	1,146,098	159	de manera que	447,642
2	ir a	12,933,567	81	a principio (al principio)	1,137,639	160	en contacto con	447,210
3	el (la, lo, los, las) cual	11,177,151	82	punto de vista	1,136,204	161	al servicio de	441,602
4	tener que	10,588,670	83	puesto que	1,099,659	162	calidad de vida	441,073
5	para que	10,504,927	84	a favor	1,089,374	163	en todo caso	431,043
6	ya que	8,445,631	85	a fin de	1,080,004	164	por completo	430,858
7	a través de	7,319,562	86	medio ambiente	1,062,339	165	al lado de	422,752
8	después de	6,514,740	87	con el fin de	1,050,052	166	en virtud de	415,066
9	dentro de	6,060,836	88	sin que	1,039,525	167	sí que	412,302
10	sin embargo	6,040,240	89	en torno a	1,038,626	168	en principio	411,659
11	antes de	5,524,276	90	en contra de	1,037,768	169	a pie	408,873
12	así como	5,191,536	91	por otra parte	1,036,439	170	a mano	408,821
13	a partir de	4,316,702	92	cada vez más	1,032,932	171	en el ámbito de	403,882
14	un poco	4,306,411	93	con respecto a	1,003,483	172	atreverse a	393,490
15	por ejemplo	4,171,475	94	a cargo de	989,867	173	terminar de	391,018
16	además de	4,141,639	95	por encima	955,816	174	a causa de	384,335
17	de forma	3,791,913	96	en el marco de	937,735	175	inferior a	382,657
18	frente a	3,533,516	97	por último	917,500	176	a largo plazo	382,347
19	cada uno	3,529,226	98	otra vez	905,895	177	por lo general	381,566
20	volver a	3,517,912	99	de carácter	872,799	178	por delante	375,260
21	de manera	3,371,791	100	de color	871,665	179	en venta	359,725
22	gracias a	3,199,460	101	en lugar de	871,535	180	mientras tanto	356,399
23	es decir	3,127,683	102	debajo de	871,433	181	por ahora	353,159
24	junto a	3,090,147	103	detrás de	869,502	182	ahora mismo	352,586
25	sobre todo	3,005,802	104	dar a conocer	862,643	183	en serio	348,720
26	por eso	2,984,709	105	por encima de	831,171	184	de emergencia(s)	348,431
27	empezar a	2,924,374	106	dado que	830,494	185	terminar con	346,708
28	mientras que	2,912,661	107	todo el mundo	829,855	186	valer la pena	338,718
29	a pesar de	2,816,811	108	superior a	829,231	187	para nada	337,252
30	comenzar a	2,749,809	109	venir a	825,992	188	ponerse en contacto	331,758
31	llegar a	2,711,085	110	en particular	810,335	189	a cambio de	330,203
32	así que	2,687,692	111	en medio	804,628	190	a modo de	319,285

33	por parte de	2,611,956	112	en función de	786,057	191	en el interior de	317,342
34	por qué	2,591,510	113	a menudo	776,080	192	en cualquier caso	317,023
35	llevar a cabo	2,513,521	114	en cambio	765,481	193	a base de	314,825
36	a nivel	2,437,666	115	por lo menos	759,543	194	para siempre	313,728
37	cerca de	2,375,576	116	esta vez	754,618	195	a diario	313,286
38	en cuanto a	2,338,204	117	hacer falta	727,021	196	algo así	308,337
39	hasta que	2,203,336	118	menor de	726,267	197	tanto como	302,966
40	de acuerdo	2,172,771	119	una vez que	709,018	198	estar a punto de	296,824
41	aquel (aquella/os/as) que	2,171,377	120	aún más	708,034	199	de conformidad con	293,861
42	dejar de	2,156,984	121	hoy en día	705,153	200	hacer frente	290,587
43	fuera de	2,119,064	122	optar por	701,472	201	en favor	283,907
44	de seguridad	2,049,404	123	en línea (en línea + adj)	688,389	202	de repente	282,723
45	de hecho	1,843,317	124	más bien	684,872	203	de vuelta	282,712
46	en forma	1,803,722	125	siempre que	653,061	204	dudar en	282,148
47	a lo largo de	1,770,451	126	poco a poco	645,138	205	tener miedo	281,203
48	alrededor de	1,754,471	127	en cuanto	638,283	206	ser de	279,721
49	acerca de	1,734,582	128	a principio de	635,076	207	en común	275,958
50	haber de	1,715,601	129	en vez de	634,703	208	por falta de	270,290
51	en general	1,676,925	130	nada más	626,861	209	siempre y cuando	269,656
52	como si	1,668,773	131	poner en marcha	620,919	210	en nombre de	268,710
53	de nuevo	1,655,203	132	de esta forma	618,569	211	al aire libre	265,663
54	pensar en	1,613,555	133	como consecuencia	605,487	212	de todas formas	265,137
55	por tanto	1,608,174	134	más o menos	592,225	213	de estilo	261,820
56	junto con	1,601,619	135	conforme a	589,170	214	encima de	259,309
57	de acuerdo con	1,577,796	136	tener que ver con	579,017	215	tener razón	257,344
58	tal como (tales como)	1,564,878	137	cada vez que	570,379	216	de forma que	253,066
59	por lo tanto	1,515,085	138	la verdad ser que	569,809	217	con la finalidad	250,492
60	por X parte	1,484,060	139	en fin	566,466	218	en colaboración con	250,376
61	tener en cuenta	1,452,948	140	por un lado	543,665	219	de vez en cuando	247,489
62	en caso de (en casos de)	1,398,024	141	aun así	541,676	220	por escrito	246,999

63	no obstante	1,395,478	142	en consecuencia	531,588	221	un abrazo	246,133
64	en donde	1,366,993	143	a diferencia de	528,904	222	tener sentido	240,519
65	en realidad	1,354,018	144	de momento	525,484	223	de ahí que	240,276
66	a final de	1,353,638	145	un saludo	518,339	224	pues bien	239,121
67	acabar de	1,348,264	146	dar lugar	515,230	225	a menos que	238,144
68	por favor	1,344,999	147	de moda	506,748	226	en color	237,865
69	formar parte	1,338,223	148	en concreto	503,625	227	merecer la pena	236,860
70	a la hora de	1,287,569	149	desde luego	502,053	228	poner a disposición	226,168
71	por supuesto	1,269,625	150	por fin	499,268	229	de todos modos	221,631
72	cumplir con	1,263,257	151	tender a	497,781	230	una y otra vez	218,897
73	a continuación	1,187,126	152	ahora que	476,799	231	ante todo	218,513
74	por medio	1,185,997	153	por debajo	474,103	232	de alquiler	218,135
75	de calidad (de la calidad)	1,163,137	154	como resultado	474,062	233	punto de partida	217,698
76	por otro lado	1,162,854	155	antes que	471,076	234	como mínimo	213,462
77	desde que	1,156,285	156	delante de	470,159	235	hoy día	201,121
78	tal vez	1,155,354	157	a cambio	461,099			
79	tener lugar	1,149,286	158	de modo que	460,774			

4. Lista completa de las locuciones más frecuentes

Para facilitar la lectura, añadimos la categoría sintáctica de cada locución en la lista completa, una frase de ejemplo tomada del corpus y el “nivel de frecuencia”, que servirá como índice marcador de gradación. Por ejemplo, *el que* aparece como -50, lo que indica que se utiliza con más frecuencia que la 50ª palabra más frecuente en español. Aquí ponemos una muestra de la lista completa.

#	Locución	Categoría	Significado	Frecuencia estimada	Nivel de frecuencia	Ejemplos
1	el que (la que, los que, las que, lo que)	loc pron	the one who/that what (relative pronoun)	63.992.369	-50	Pudimos saber que su pareja, con la que había convivido más de doce años, no le estaba siendo infiel.
2	ir a	loc aux	be going to	12.933.567	-100	Permanezca conectado con lo que le importa a usted.
3	el cual (la cual, los cuales, las cuales, lo cual)	loc pron	the one who/that what (relative pronoun)	11.177.151	-200	En principio no voy a repetir porque tengo varias mascarillas por gastar pero no porque no me haya gustado.
4	tener que	loc aux	have to	10.588.670	-200	La superficie de Jalapa es de 642.91 Kilómetros Cuadrados, el cual representa el 2.6% del territorio del estado.
5	para que	loc conj	so that	10.504.927	-200	Tienes que establecer una contraseña para la conexión.
6	ya que	loc conj	because, since, now that	8.445.631	-200	En algunos casos es necesario instalar cookies para que el navegador no olvide su decisión de no aceptación de las mismas.
7	a través de	loc prep	through	7.319.562	-200	La situación es confusa y contradictoria ya que el país presenta indicadores de crecimiento, tanto económicos como sociales.
8	después de	loc prep	after	6.514.740	-500	Con toda la información que nos proporcione a través del formulario es suficiente para poder determinar el valor del servicio.
9	dentro de	loc prep	inside, within	6.060.836	-500	Por lo tanto, inmediatamente después de este proceso necesita lavarse las manos, los dedos y las uñas.
10	sin embargo	loc conj	however, nevertheless	6.040.240	-500	Dentro de un par de semana recibirá una copia, y la añadiré a la galería.
11	antes de	loc prep	before	5.524.276	-500	Sin embargo, la familia no tiene los fondos económicos necesarios para viajar y hacer frente al problema de su hija.
12	así como	loc conj, adv	just like, just as	5.191.536	-500	Antes de la conferencia tendrá lugar la presentación de la Publicación Memoria de Proyectos 09/10.

La lista completa puede descargarse en nuestra página web: <https://onl.sc/p52YLNr>.

Como se puede observar en la lista, algunas locuciones tienen una frecuencia de uso extremadamente alta. Por ejemplo, la frecuencia de las 75 locuciones más frecuentes supera a la de la palabra singular más frecuente en el puesto 1,500. Es importante recordar que la dependencia de las palabras más frecuentes del español, en particular de las 1,500 primeras, es especialmente alta. Aunque el Instituto Cervantes no lo señala explícitamente, la capacidad de manejar un vocabulario de este tamaño se considera una tarea de nivel A1 y A2. “El vocabulario importante” puede variar según los intereses de los estudiantes y los objetivos de la enseñanza pero consideramos que las locuciones de este nivel de frecuencia de uso deben ser adquiridas en una etapa inicial, independientemente de los intereses y los objetivos. Como defienden Christiansen y Arnon (2017), las locuciones en español no son rarezas lingüísticas sino elementos básicos que constituyen partes importantes del idioma.

5. Conclusión y trabajo en el futuro

En este estudio, hemos desarrollado una lista de las locuciones más frecuentes a través del análisis de corpus y hemos demostrado su importancia didáctica. Nuestro trabajo en el futuro consistirá en ampliar y profundizar la lista. Actualmente se estima que el tamaño de la lista solo cubre hasta el nivel B2, por lo que será necesario ampliarla para satisfacer las necesidades de los estudiantes de niveles superiores. Además, en este estudio hemos descartado factores como las diferencias regionales y los registros, que también son esenciales para manejar las locuciones presentadas. Será necesario llevar a cabo futuras investigaciones sobre las regiones y los registros en los que estas locuciones se utilizan con mayor frecuencia.

La pregunta que hemos abordado en este estudio es “¿qué locuciones deben presentarse a

los alumnos?” El próximo paso será abordar “¿cómo las enseñamos?”

En cuanto a la enseñanza para el nivel inicial, es fructífero incorporar la enseñanza de la locución en la de gramática, puesto que la mayoría de las 75 locuciones más frecuentes son funcionales; pronominal, preposicional, conjuntiva, auxiliar y afijal. Aparte de las locuciones de estas clases, algunas de las locuciones adverbiales también muestran la cierta funcionalidad. A modo de ejemplo, algunas funcionan como marcadores del discurso (*por ejemplo, sobre todo, por eso*, etc.).

De ello, se puede deducir que el papel principal de las locuciones en español es completar la gramática y, por lo tanto, su enseñanza debe ser incorporada en la enseñanza de la gramática. Considerando que la investigación y producción académica en la enseñanza de ELE presta cada vez más atención a las locuciones léxicas nominales y verbales que señalamos en 1.2., esperamos que este estudio contribuya a esclarecer la importancia didáctica de las locuciones gramaticales.

Aparte de lo anterior, también consideramos que los estudios contrastivos entre estas locuciones y sus correspondientes en la lengua materna de los alumnos son especialmente importantes, ya que como Yamashita & Jiang (2010) y Laufer & Waldman (2011) señalan, el mayor factor de los errores del uso de UPV suele ser la transferencia negativa desde la lengua materna. Este punto coincide con nuestra experiencia como profesores y hablantes no nativos. Por ello es provechoso que los profesores tengan en cuenta las similitudes y diferencias entre estas locuciones y las correspondientes en la lengua materna de los estudiantes de antemano. De esa forma sería posible diseñar estrategias de enseñanza que disminuyan tal transferencia negativa.

Bibliografía

Alonso Ramos, M. (2004) : *Las construcciones con verbo de apoyo*, Visor Libros.

Altenberg, B. (1998) : «On the phraseology of spoken English: The evidence from recurrent word-combinations», in A. P. Cowie (ed.), *Phraseology Theory Analysis and Applications*, 101–122, Oxford University Press.

Christiansen, M. H., & Arnon, I. (2017) : «More Than Words: The Role of Multiword Sequences in Language Learning and Use», *Topics in Cognitive Science*, 9(3), 542–551.

Davies, M., & Face, T. (2006) : «Vocabulary coverage in Spanish textbooks: how representative is it?», *Selected Proceedings of the 9th Hispanic Linguistics Symposium*, 132–143.

Davies, M., & Hayward Davies, K. (2018) : *A Frequency Dictionary of Spanish Core Vocabulary for Learners*, (second edition), Routledge.

Erman, B., & Warren, B. (2000) : The idiom principle and the open choice principle, *Text*, 20(1), 29–62.

García Fernandez, L. (2006) : *Diccionario de perífrasis verbales*, Gredos.

Gómez Torrego, L. (1999) : «Los verbos auxiliares. Las perífrasis de infinitivo», en I. Bosque y V. Demonte (dirs.), *Gramática descriptiva de la lengua española*, 3323–3389, Espasa Calpe.

- Koike, K. (2001) : *Colocaciones léxicas en el español actual: estudio formal y léxicosemántico*, Universidad de Alcalá/Universidad de Takushoku.
- Koprowski, M. (2005) : «Investigating the usefulness of lexical phrases in contemporary coursebooks», *ELT Journal*, 59(4), 322–332.
- Laufer, B., & Waldman, T. (2011) : «Verb-noun collocations in second language writing : A corpus analysis of learners' English», *Language Learning*, 61(2), 647–672.
- Leal Riol, M. J. (2011) : *La enseñanza de la fraseología en español como lengua extranjera Estudio comparativo dirigido a estudiantes anglófonos*. Secretariado de publicaciones e intercambio editorial Universidad de Valladolid.
- Martinez, R. (2011) : *The Development of a Corpus-informed List of Formulaic Sequences for Language Pedagogy*, Doctoral Thesis, University of Nottingham.
- Martinez, R. (2013) : «A framework for the inclusion of multi-word expressions in ELT», *ELT Journal*, 67(2), 184–198.
- Martinez, R., & Murphy, V. A. (2011) : «Effect of frequency and idiomaticity on second language reading comprehension», *TESOL Quarterly*, 45(2), 267–290.
- Mendivil Giró, J. L. (1990) : «El concepto de “locución verbal” y su tratamiento léxico», *Cuadernos de Investigación Filológica*, 16, 5–30.
- Meunier, F., & Granger, S. (Eds.) : (2008). *Phraseology in Foreign Language Learning and Teaching*, John Benjamins Publishing Company.
- Moreno, Teva, I. (2012): *Las secuencias formulaicas en la adquisición de español L2*, Doctoral thesis, Stockholm University.
- Nation, I. S. P. (2001) : *Learning Vocabulary in Another Language*, Cambridge University Press.
- Real Academia Española/Asociación de Academias de la Lengua Española. (2009) : *La nueva gramática de la lengua española*, Espasa Calpe.
- Schmitt, N., & Schmitt, D. (2020) : *Vocabulary in Language Teaching*, Cambridge University Press.
- Siyanova-Chanturia, A., & Spina, S. (2020) : «Multi-word expressions in second language writing: A large-scale longitudinal learner corpus study», *Language Learning*, 70(2), 420–463.
- Yamashita, J., & Jiang, N. (2010) : «L1 influence on the acquisition of L2 collocations: Japanese ESL users and EFL learners acquiring English collocations», *TESOL Quarterly*, 44(4), 647–668.

Corpus, diccionarios y materiales

- Bregstein, B. (2016) : *Complete Spanish Step-by-Step*, McGraw Hill Education.
- Encinar, Á. (2012) : *Uso interactivo del vocabulario (Nivel A-B1)*, Edelsa.
- Lexical Computing (s.f.) : *Sketch Engine. Spanish web 2018* (esTen Ten).
https://app.sketchengine.eu/#dashboard?corpname=preloaded%2Festenten18_fl5_1, [Última consulta 21/09/2022.].

Muñoz-Basols, Javier, Pérez Sinusía, Y y David, M. (2014) : *Spanish Idioms in Practice*, Routledge.

Pérez Cinco, W. (2000): *Manual práctico de la preposición española*, Verbum,

De Prada, M, Salazar, D y Molero CM. (2012) : *Uso interactivo del vocabulario (Nivel B2-C2)*, Edelsa.

Real Academia Española: (2014) : *Diccionario de la lengua española*, 23ª edición, Espasa. <https://dle.rae.es/>,
[Última consulta 21/09/2022].

Weibel, P. (2004) : *The Big Red Book of Spanish Idioms*, McGraw Hill.

WordReference.com. *The WordReference English-Spanish Dictionary*,
<https://www.wordreference.com/es/translation.asp>, [Última consulta 21/09/2022.].

高橋覚二 伊藤ゆかり (2013) 例文で覚えるスペイン語熟語集 白水社 (Takahashi & Ito 2013).
(Universidad de Kyushu)