ONOMÁZEIN
Revista de lingüística, filología y traducción

PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE
FACULTAD DE LETRAS

Frecuencia y corrección colocacional en la producción escrita de aprendices de español

Frequency and collocational correction in the writing of learners of Spanish

Marcos García Salido
Universidade da Coruña
España

ONOMÁZEIN 38 (diciembre de 2017): 22-46
DOI: 10.7764/onomazein.38.10

Marcos García Salido: Departamento de Gallego-Portugués, Francés y Lingüística, Facultad de Filología, Universidade da Coruña, España. | Correo electrónico: marcos.garcias@udc.es

Fecha de recepción: agosto de 2016
Fecha de aceptación: enero de 2017

Resumen

El objetivo del presente trabajo es determinar si existe alguna relación entre la corrección de las colocaciones producidas por aprendices de español en textos escritos y la asociación que sus miembros presentan en un corpus representativo de esta lengua. Para ello se partirá de una muestra de colocaciones extraídas de un corpus de aprendices, las que han sido identificadas desde presupuestos fraseológicos, y se estudiará su frecuencia en un corpus representativo del español. A partir de estos datos, se comprobará si existe alguna relación entre la asociación de los miembros de una colocación en términos de frecuencia de coaparición e información mutua y la probabilidad con la que aparece producida correctamente en los textos de aprendices.

Palabras clave: aprendizaje de español; léxico; colocaciones; frecuencia; información mutua.

Abstract

This study aims to establish whether there is any relation between the correctness of collocations present in the writings of learners of Spanish and the association their members display in a representative corpus of this language. In order to do that, starting from a sample of collocations extracted from a learner corpus where they have previously been identified by means of phraseological criteria, I study their frequency in a representative corpus of Spanish. Based on these data, I try to establish if the association between the constituents of a collocation, as measured in terms of frequency of co-occurrence and mutual information, has any influence on the probability wherewith learners produce such collocation correctly.

Keywords: learning of Spanish; vocabulary; collocations; frequency; mutual information.

1. Introducción

Desde la segunda mitad del pasado siglo las colocaciones léxicas han venido suscitando un interés creciente en la investigación lingüística (Firth, 1957 [1951], Mel'cuk 1960 *apud* Weinreich, 1969; Halliday, 1966; Cowie, 1981), si bien desde concepciones que divergen en varios aspectos. En este trabajo entenderé por *colocación* una combinación del tipo *dar un paseo, fumador empedernido, paquete de tabaco*, etc., en la que una de las unidades léxicas que la forman se selecciona en función de la otra. Así, en los ejemplos anteriores, el nombre *paseo* selecciona al verbo *dar* y no, por ejemplo, *hacer*; el nombre *fumador* al adjetivo *empedernido* y no, por ejemplo, el sintagma *en cadena* (*cf.* ing. *chain smoker*), y el nombre *tabaco* selecciona al nombre *paquete* y no, por ejemplo, *estuche*. Siguiendo a Hausmann (1989), denominaré *base* al elemento que dirige la selección y *colocativo* al elemento seleccionado (términos que también acabarían adoptando Alonso Ramos (2004) o Mel'čuk (2012)).

En lo que concierne a la lingüística española en concreto, el fenómeno ha venido recibiendo una atención continuada desde la década de 1990. Es pionero, en este sentido, el trabajo de Alonso Ramos (1993). Este fenómeno reviste, además, un especial interés desde la perspectiva de la lingüística aplicada y, más en concreto, desde el aprendizaje de lenguas segundas y lenguas extranjeras por diversos motivos. En primer lugar, se supone que contar con un repertorio relativamente amplio de colocaciones —o un conjunto de secuencias que se solaparía hasta cierto punto con lo que aquí se entiende por tal— disminuye el coste de procesamiento en la producción lingüística (Pawley y Syder, 1983; Wray, 2002). En segundo lugar, porque ser capaz de producir este tipo de combinaciones también parece favorecer la comprensión por parte del oyente/lector (Wray, *ibid.*) o, al contrario, la presencia de combinaciones que se desvían del repertorio colocacional de un hablante nativo puede perjudicarla (Howarth, 1998: 177). Además, una serie de estudios (Altenberg y Granger, 2001; Gilquin, 2007, o Vincze y otros (en prensa), por citar una selección) sugiere que, incluso para aprendices de niveles avanzados, las colocaciones siguen siendo un aspecto problemático.

Probablemente debido a su importancia para la adquisición de una lengua y a la constatación de las dificultades que comportan para un hablante no nativo, por un lado, así como a la reciente disponibilidad de corpus de aprendices de español, por otro, en los últimos tiempos han proliferado una serie de estudios centrados en las colocaciones en el ámbito del español como lengua extranjera o segunda (en adelante, ELE/2). Algunos de estos estudios tienen una vertiente más didáctica (Buckingham, 2008; Higueras, 2006; Ferrando Aramo, 2012), mientras que otros se centran en analizar el uso de las colocaciones por parte de los aprendices (Alonso Ramos y otros, 2010a y 2010b; Pérez Serrano, 2014; Uriel Domínguez, 2014; Vincze, 2015, que también trata el aspecto anterior; o Vincze y otros, en prensa). Finalmente, pueden encontrarse algunas propuestas sobre cómo determinar la competencia colocacional de los aprendices (Orol González, 2015; Pérez Serrano, 2015).

El presente estudio pretende ser una nueva contribución al estudio del uso de las colocaciones que hacen los aprendices del español, pero desde una perspectiva que todavía no ha sido abordada en la investigación sobre esta lengua; se pondrán en relación los datos del input que reciben los aprendices con su producción colocacional. Más concretamente, el propósito de este trabajo es verificar hasta qué punto ciertas medidas de la asociación entre los miembros de una determinada colocación (derivadas de su frecuencia en un corpus) están relacionadas con la probabilidad de que dicha colocación sea producida de forma correcta por los aprendices. Para ello se utilizarán datos provenientes de un corpus de aprendices junto a otros extraídos de un corpus representativo del español europeo. Se combina así elementos de dos perspectivas diferentes en cuanto a la noción de colocación: por un lado, la concepción de una tradición fraseológica de la que se toma la noción de colocación definida más arriba y para la cual la frecuencia de aparición de una determinada combinación es irrelevante para tratarla como colocación o combinación libre (*cf.*, por ejemplo, Alonso Ramos, 1994-1995: 14); por otro, ciertas medidas de asociación que, en general, responden a una forma de entender las colocaciones, según la cual estas serían combinaciones de palabras que ocurren en un corpus con una frecuencia mayor de lo que sería esperable si las palabras de dicho corpus se distribuyesen al azar (la idea se encuentra ya en Halliday (1966: 156) y será adoptada por autores como Sinclair (1987: 70)). El primer enfoque se ha utilizado para delimitar la muestra de colocaciones estudiada, y algunas de las medidas propuestas por ciertos autores pertenecientes al segundo enfoque, para determinar el grado de asociación que existe entre los constituyentes de las colocaciones de la muestra.

El artículo se estructura como sigue. En la sección que viene a continuación se revisa una serie de estudios que se han servido de medidas de asociación para dar cuenta de las diferencias que se observan entre hablantes nativos y no nativos con respecto a la producción y al procesamiento de combinaciones léxicas, y se concretan las preguntas que pretende responder el presente trabajo. A continuación se describen los datos estudiados y los métodos usados para analizarlos, en concreto, los criterios para identificar colocaciones y errores colocacionales en textos de aprendices, así como el procedimiento de asignación de frecuencia a dichas colocaciones. Tras presentar los resultados del análisis, se presenta una discusión de los mismos. El trabajo se cierra exponiendo las conclusiones del estudio y señalando qué cuestiones quedan pendientes para investigaciones futuras.

2. Antecedentes y objetivo

Una serie de estudios recientes ha utilizado medidas de asociación basadas en la frecuencia que presentan determinadas combinaciones léxicas en un corpus bien para analizar las producciones escritas de aprendices y compararlas con producciones de nativos o con las de aprendices de distintos niveles (Durrant y Schmitt, 2009; Bestgen y Granger, 2014; Granger y Bestgen, 2014; Lorenz, 1999), bien para determinar la relación de dichas medidas con el procesamiento

de las combinaciones en cuestión (Ellis y otros, 2008). Empezaré por dar cuenta de las medidas empleadas en los trabajos citados para, a continuación, revisar brevemente sus resultados.

Existe un amplio repertorio de medidas basadas en la frecuencia que determinan el grado de asociación de una determinada combinación de palabras[1]. Tres son especialmente relevantes para los propósitos del presente trabajo: (i) la *frecuencia de coaparición* de las palabras que forman una determinada combinación, (ii) la llamada *t-score* de dicha combinación y (iii) la *información mutua* (*mutual information* o MI, por sus siglas en inglés).

La más simple de las tres es la frecuencia de coaparición: se trata simplemente del cómputo de veces que una determinada combinación se da en un corpus concreto. Esta medida interviene en el cálculo de las otras dos junto con la *frecuencia esperada* de la combinación en cuestión. La frecuencia esperada se obtiene multiplicando el número de ocurrencias de cada uno de los constituyentes de la colocación (f_a, f_b) y dividiéndolos por el número de ocurrencias (esto es, el tamaño total) del corpus manejado (N)[2]. Se obtiene así la frecuencia que se esperaría si las dos palabras que forman la combinación estuviesen distribuidas al azar en el corpus en cuestión.

(1) frecuencia esperada = $f_a \times f_b / N$

A partir de estas dos medidas (la frecuencia de coaparición efectivamente observada y la esperada), la *t-score* mide la diferencia entre ellas y la probabilidad de que dicha diferencia se debe al azar. Se calcula según la fórmula siguiente, donde O es la frecuencia observada y E la esperada:

(2) t-score = $O - E / \sqrt{O}$

Normalmente, se asume un valor de 2 como umbral de significatividad para esta medida. Es decir, una *t-score* igual o mayor que 2 sería evidencia suficiente para considerar que la combi-

1 *Vid*. Evert (2008) para un panorama más completo del que se ofrece aquí.

2 Evert (2008) advierte de que el cálculo de cada uno de estos datos depende del método con el que se extraen los candidatos a colocación en cada caso. Así, por ejemplo, si, en lugar de *tokens* a una determinada distancia, se extraen relaciones de dependencia del tipo *x→y*, en una combinación como *a→b* la frecuencia de *a* sería el número de veces que aparece como núcleo del tipo de dependencia considerado (*a→y*), la frecuencia de *b*, las veces que aparece como dependiente (*x→b*) y el total de la muestra, el número de dependencias del tipo *x→y* extraídas. En la práctica, sin embargo, parece que no siempre se atiende a la relación entre el método de extracción y el método de cálculo de frecuencias. Así, por ejemplo, los *word sketches* de *Sketch Engine* se basan en una lista de candidatos a colocación constituidos por la coocurrencia de dos lemas en una determinada cadena de etiquetas de clase de palabra (por ejemplo, verbo+sustantivo) y sus medidas de asociación se calculan a partir de la frecuencia de los dos lemas y del tamaño total del corpus y no del número de cadenas verbo+sustantivo presentes en él (Lexical Computing, 2015).

nación que tiene esa puntuación es significativamente más frecuente de lo que se esperaría por azar (y al revés, valores iguales o inferiores a -2 indicarían combinaciones significativamente menos frecuentes de lo esperable por azar). Aparte de proporcionar este umbral de "colocabilidad", esta medida de asociación presenta un grado de correlación notable con la frecuencia de coaparición, hasta tal punto que en algunos estudios revisados en esta sección se interpreta como un índice de frecuencia colocacional.

La información mutua responde a una lógica distinta. Según Durrant (2014: 455), mide hasta qué punto uno de los miembros de la colocación predice la presencia del otro. Se calcula según la fórmula siguiente:

(3) $MI = \log_2(O/E)$

Valores altos de MI indican que una proporción considerable de las ocurrencias de uno de los constituyentes de la combinación tiene lugar en contextos donde el otro también está presente. El valor que normalmente se adopta como umbral de colocabilidad es 3. Como en el caso anterior, valores negativos indicarían que las palabras en combinación en cierto modo se repelen, mientras que valores próximos a 0 indicarían una proporción 1:1 entre la frecuencia observada y la esperable por azar. Al contrario que la *t-score*, valores altos de MI no tienen por qué corresponderse con colocaciones muy frecuentes. Es más, se ha señalado que la MI privilegia colocaciones infrecuentes (*cf.*, por ejemplo, Evert, 2008).

Uno de los primeros estudios en analizar la fraseología de los aprendices de una lengua por medio de medidas de asociación fue Lorenz (1999). Se centraba en el análisis de la intensificación de adjetivos en aprendices del inglés y, en contra de lo que ha acabado por ser habitual, el autor obtuvo las medidas de asociación de las propias muestras de aprendices y hablantes nativos estudiadas en lugar de extraerlas de un corpus de referencia. Lorenz encontró un patrón de sobreutilización de intensificadores de frecuencia alta en los aprendices. Sin embargo, en términos de información mutua, la media de las combinaciones empleadas por los aprendices era menor que la media de las usadas por los nativos (Lorenz, 1999: 185). En el caso de los nativos, el autor interpreta la información mutua como una medida de la "idiomaticidad" de una determinada combinación (1999: 184), mientras que en los aprendices defiende cierta cautela al respecto, pues una MI alta puede ser el resultado de combinar de manera errónea dos formas poco frecuentes.

Ellis y otros (2008) combinan métodos de lingüística de corpus (de donde extraen la información relativa a la frecuencia de las combinaciones que estudian) con un enfoque psicolingüístico. Realizan varios experimentos en los que toman diferentes medidas de la rapidez de reacción de hablantes nativos y no nativos a combinaciones léxicas con distintas frecuencias y valores de información mutua. De forma recurrente, la frecuencia influye de forma significativa en los tiempos de reacción de los no nativos, mientras que el parámetro determinante en el

caso de los hablantes nativos es el valor de información mutua de la combinación en cuestión. Los autores concluyen que los aprendices, algunos de ellos tras diez años de instrucción en inglés, están empezando a reconocer las combinaciones más frecuentes de esa lengua, mientras que para los nativos la frecuencia ha dejado de ser tan relevante, debido a su alto grado de exposición a datos discursivos, y han extraído de su input información relativa a la mayor o menor probabilidad de que una determinada forma se dé en el contexto de otra (Ellis y otros, 2008: 391).

Volviendo a estudios basados exclusivamente en datos de corpus, Durrant y Schmitt (2009) (y previamente Durrant (2008) con datos y resultados similares) comparan colocaciones constituidas por premodificador + nombre en escritos de aprendices de inglés y hablantes nativos en cuanto a sus valores de *t-score* e información mutua en un corpus de referencia de esa lengua. La comparación de los repertorios de nativos y aprendices se lleva a cabo tanto en *types* como en *tokens*. En primer lugar, dividen su muestra en combinaciones por encima de los umbrales convencionales de las medidas usadas, por un lado, y en combinaciones con menos de 5 ocurrencias en su corpus de referencia, por otro. Las combinaciones infrecuentes representan una proporción mayor en el caso de los nativos. En el caso de las colocaciones (esto es, combinaciones con *t-score* ≥ 2 y MI ≥ 3) recurrentes (más de cinco ocurrencias), encuentran que los aprendices sobreutilizan las que tienen valores altos de *t-score* e infrautilizan las que presentan valores altos (≥ 7) de información mutua.

Bestgen y Granger (2014) y Granger y Bestgen (2014) adoptan unos presupuestos muy similares a los anteriores: se centran en secuencias de dos palabras adyacentes (bigramas) y utilizan las mismas medidas de asociación. Ahora bien, amplían el espectro de categorías consideradas al analizar todos los bigramas de su corpus sin limitarse a combinaciones del tipo premodificador + nombre. En el caso de Granger y Bestgen (2014), además del conjunto total de bigramas de los corpus estudiados, estudian individualmente las combinaciones de categorías premodificador + nombre, nombre + nombre, adjetivo + nombre y adverbio + adjetivo. Otra novedad de estos trabajos es que, en lugar de comparar el uso de los aprendices con el de los nativos, llevan a cabo un estudio seudolongitudinal (Granger y Bestgen) de aprendices de distintos niveles y otro (Bestgen y Granger) que combina la perspectiva longitudinal, estudiando muestras de los mismos aprendices recogidas en distintas fases de su desarrollo, con la seudolongitudinal. En Granger y Bestgen (2014) se detecta un aumento significativo de la proporción de bigramas con valores altos de MI a la vez que una disminución también significativa de la proporción de bigramas con *t-score* alta en los aprendices más avanzados. En Bestgen y Granger (2014) se constata en el análisis longitudinal que la media de *t-score* de cada texto desciende de forma significativa conforme el nivel de los aprendices sube y en el seudolongitudinal que las calificaciones dadas por una serie de evaluadores se correlacionan positivamente con el valor de información mutua.

Los trabajos revisados parecen apuntar todos en una dirección similar: las combinaciones léxicas más frecuentes en la lengua general son también más frecuentes en las produc-

ciones de aprendices que en las de nativos y aprendices más avanzados y son más fáciles de procesar para hablantes no nativos que las menos frecuentes. Las que presentan una MI alta, por su parte, son menos frecuentes en las producciones escritas de los hablantes no nativos y a estos les resultan más difíciles de procesar que a los nativos.

Teniendo en cuenta lo anterior, el objetivo de este trabajo es determinar si la asociación entre los constituyentes de las colocaciones de la muestra que se estudiará aquí (determinada por medio de medidas basadas en la frecuencia) está en relación también con su corrección. Más concretamente, la investigación tratará de responder a las siguientes preguntas:

a) ¿Cómo influye la frecuencia de coaparición de los constituyentes de una colocación en la probabilidad de que los aprendices la produzcan correctamente?

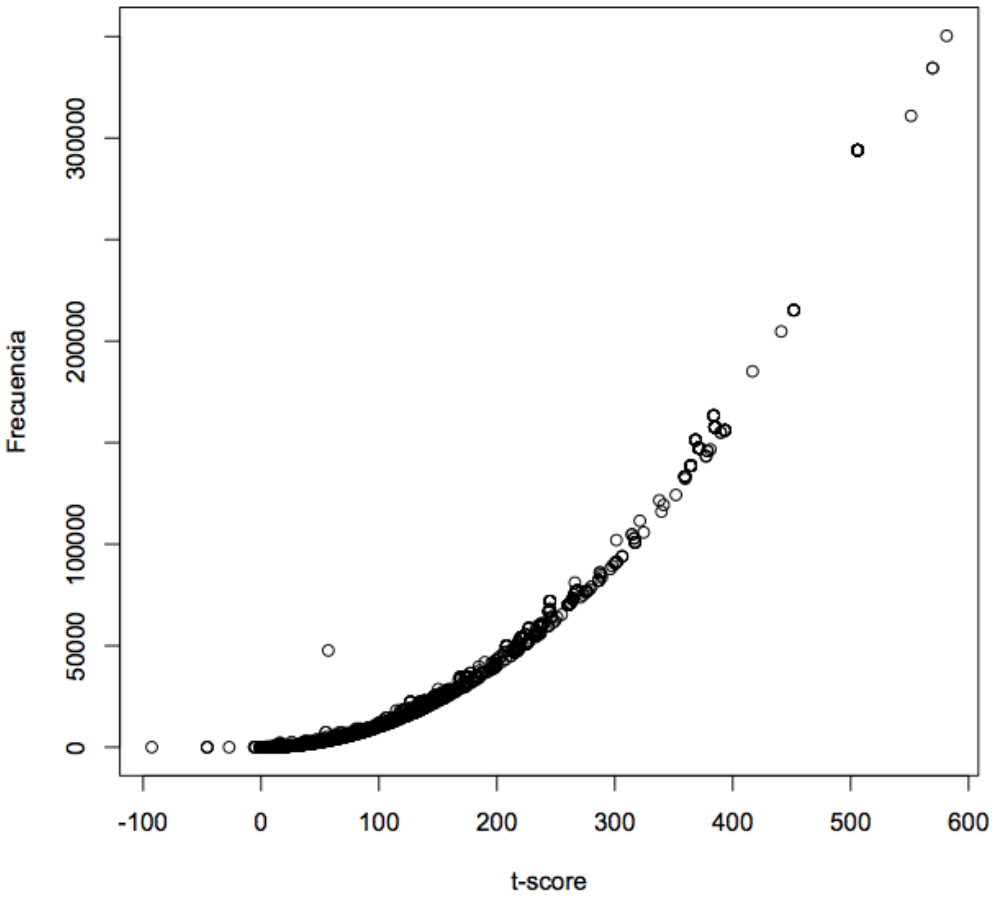
b) ¿Cómo influye la información mutua de la colocación en el mismo parámetro?

Como se desprende de las preguntas anteriores, la investigación se centrará en solo dos medidas de asociación de las tres discutidas: frecuencia e información mutua. Se deja, por tanto, de lado la *t-score*. Los motivos para haber tomado esta decisión son dos. En primer lugar, los estudios reseñados que se sirven de esta última medida la interpretan directamente como un índice de frecuencia colocacional. Es cierto que la *t-score* aporta información adicional además de la frecuencia: un umbral de significatividad que en los trabajos reseñados se toma como un indicio de qué combinaciones son colocaciones y cuáles no. Ya que en el presente trabajo se ha determinado con otros criterios el carácter colocacional de una determinada combinación (*vid. infra*), el plus informativo que ofrece la *t-score* aquí no es tal. En segundo lugar, la correlación que existe entre frecuencia y *t-score* puede ser problemática para modelos de regresión como el que se empleará más adelante. La correlación entre estas dos dimensiones en la muestra de colocaciones que se analizará puede apreciarse gráficamente en la figura 1. Tiene un índice de correlación muy elevado (τ_b=0,97), lo cual indica que se trata una correlación casi perfecta (por el contrario, la correlación entre MI y frecuencia, si es que puede hablarse de tal, arroja un índice muy próximo a 0: τ_b=0,02).

3. Muestra y metodología

La muestra que se analiza en este trabajo está compuesta por las colocaciones identificadas manualmente en una parte del corpus *CEDEL2* (Lozano, 2009; Lozano y Mendikoetxea, 2013). Este fragmento de *CEDEL2* se compone de 100 textos de otros tantos aprendices que alcanzaron una puntuación de entre el 77% y el 100% en la prueba de nivel administrada a los informantes en la fase de compilación del corpus. El estudio no es pues longitudinal o seudolongitudinal, sino que se ciñe a un nivel de aprendizaje bastante concreto que podríamos situar entre el intermedio-avanzado y el avanzado. El conjunto es una mezcla de textos

FIGURA 1
Correlación entre frecuencia y *t-score* en colocaciones de aprendices



fundamentalmente argumentativos y narrativos y su extensión media es de 464,2 palabras (desv. est. = 113,83). En total, el subcorpus manejado consta de 46420 palabras.

Las colocaciones fueron identificadas manualmente por tres hablantes nativos que anotaron el corpus siguiendo los criterios de la lexicología explicativa y combinatoria (véase Mel'čuk y otros, 1995). Para esta corriente, una colocación es una combinación de unidades léxicas no libre, puesto que solo uno de los elementos, la *base*, se elige exclusivamente por su significado, mientras que la selección del otro, el *colocativo*, está condicionada por la identidad léxica del primero (Mel'čuk, 2012). Así se explica que nombres de una semántica similar, como *paseo* o *excursión*, que denotan ambos un tipo de 'desplazamiento con fines recreativos', impongan una selección diferente de verbos (*dar un paseo*, *hacer una excursión*).

Además de identificarlas según este criterio, los anotadores determinaron si la colocación era correcta o no y clasificaron los errores de acuerdo con una tipología presentada en Alonso Ramos y otros (2010a y 2010b), que daba cuenta de la localización del error (esto

es, de si afectaba a la base o al colocativo), lo describía y formulaba hipótesis acerca de su causa (básicamente, si era fruto de una transferencia de la L1 o se debía a rasgos de la lengua meta). En cuanto a la descripción, pueden establecerse dos grandes grupos de errores: gramaticales y léxicos. Los primeros se aplican a casos tales como el uso incorrecto de artículos (**tener derecho de…*, por *tener el derecho de…*), preposiciones (**hablar al teléfono*, por *hablar por teléfono*), clíticos, incorrecciones relativas a género y número de la base y el colocativo, entre otros. Los segundos son fundamentalmente casos en los que la base o el colocativo empleados no son los adecuados en la colocación en cuestión, aunque también responden a este tipo de error el uso de una combinación con apariencia de colocación donde el español hubiera usado una única unidad léxica (**poner apasionado* en lugar de *apasionar*), el uso de unidades léxicas donde sería esperable una colocación (**misinterpretaciones* en lugar de *malas interpretaciones*) o el empleo de colocaciones existentes en español empleadas con un sentido inadecuado (*dar la gana* por *tener ganas*).

Dos de los anotadores analizaron el fragmento de *CEDEL2* descrito de manera independiente. Un tercer anotador se encargó de combinar las anotaciones resultantes y resolver los casos en los que había discrepancias entre los dos primeros mediante los siguientes procedimientos: (i) consultando el *Corpus de Referencia del español actual, CREA* (si la combinación se documentaba al menos cinco veces, se consideraba correcta); (ii) en caso de que lo anterior no fuera suficiente, se consultaba a tres hablantes nativos sobre la corrección de la colocación; (iii) finalmente, los casos que no se resolvían mediante los dos procedimientos anteriores, se discutían en sesiones semanales (*vid.* Vincze y otros (2011) para más detalles sobre el proceso de anotación).

A partir de esta anotación manual, para este trabajo he extraído únicamente las colocaciones que respondían a los cuatro tipos de estructura más recurrentes:

a) sujeto + verbo (p. ej., *sale el sol, sube la deuda,* etc.).
b) verbo + objeto directo o preposicional (p. ej., *dar la bienvenida* o *asistir a la universidad*).
c) nombre *de* nombre: generalmente se trata de un nombre cuantificador que sintácticamente domina al nombre cuantificado (p. ej., *paquete de tabaco*).
d) nombre + adjetivo (p. ej., *día festivo, larga distancia*, etc.).

La muestra consta de un total de 1679 colocaciones, de las cuales 1315 se han considerado correctas y 364 incorrectas. A las colocaciones que componen esta muestra se les han asignado datos de frecuencia extraídos del corpus *esTenTen11* de español europeo (Kilgarriff y Renau, 2013) en su versión anotada mediante TreeTagger. Se ha elegido la parte europea del *esTenTen11* (a) porque el destino de los aprendices de la muestra que realizaron una estancia en un país hispanohablante era mayoritariamente España (Vincze y otros, en prensa), (b) porque era la variedad de los anotadores y, finalmente, (c) porque, independientemente de que

esto sea problemático, la variedad europea todavía parece en la práctica la referencia en el ámbito de la enseñanza del español (*cf.* Alonso Ramos, en prensa, y las referencias allí citadas)

La frecuencia de cada uno de los miembros de la colocación fue obtenida consultando sus respectivos lemas, mientras que la frecuencia de la colocación se determinó a partir de las coocurrencias de dichos lemas en una configuración sintáctica que se correspondiera con la de la colocación buscada. Con tal fin se usaron reglas similares a las que usa el Sketch Engine en sus *word-sketches* (Kilgarriff y otros, 2014), aprovechando el etiquetado morfológico del corpus, pero en la adaptación que proponen Vincze y Alonso Ramos (2013). La frecuencia de coocurrencia no se estableció, así pues, a partir de la copresencia de formas correspondientes a dos lemas en una distancia determinada, sino de su copresencia en una cadena de ciertas categorías gramaticales. La frecuencia esperada de cada colocación se calculó a partir de la frecuencia de sus constituyentes según la fórmula que figura en (1) y el valor de información mutua asignada a cada una según la fórmula que aparece en (3).

En el caso de las colocaciones incorrectas, los valores asignados son los relativos a la corrección propuesta por los anotadores. Esta decisión se justifica, en primer lugar, porque el objetivo de este trabajo es verificar la influencia que la frecuencia de aparición de una colocación en el posible input de los aprendices tiene en su producción. En este sentido, una colocación incorrecta puede no formar parte en absoluto de la lengua a la que los aprendices están expuestos, sino que supone una desviación con respecto a una determinada expresión más convencional que los correctores proponen como la forma meta (p. ej., **hacer una marca [diacrítica]* sería una desviación con respecto al más idiomático *poner un acento [diacrítico]*). En segundo lugar, si usamos la frecuencia de las colocaciones erróneas sabiendo que en muchos casos son inexistentes en la producción nativa, un análisis de frecuencia no tiene sentido, ya que el resultado probable de antemano es que las colocaciones erróneas tengan frecuencias muy bajas en conjunto.

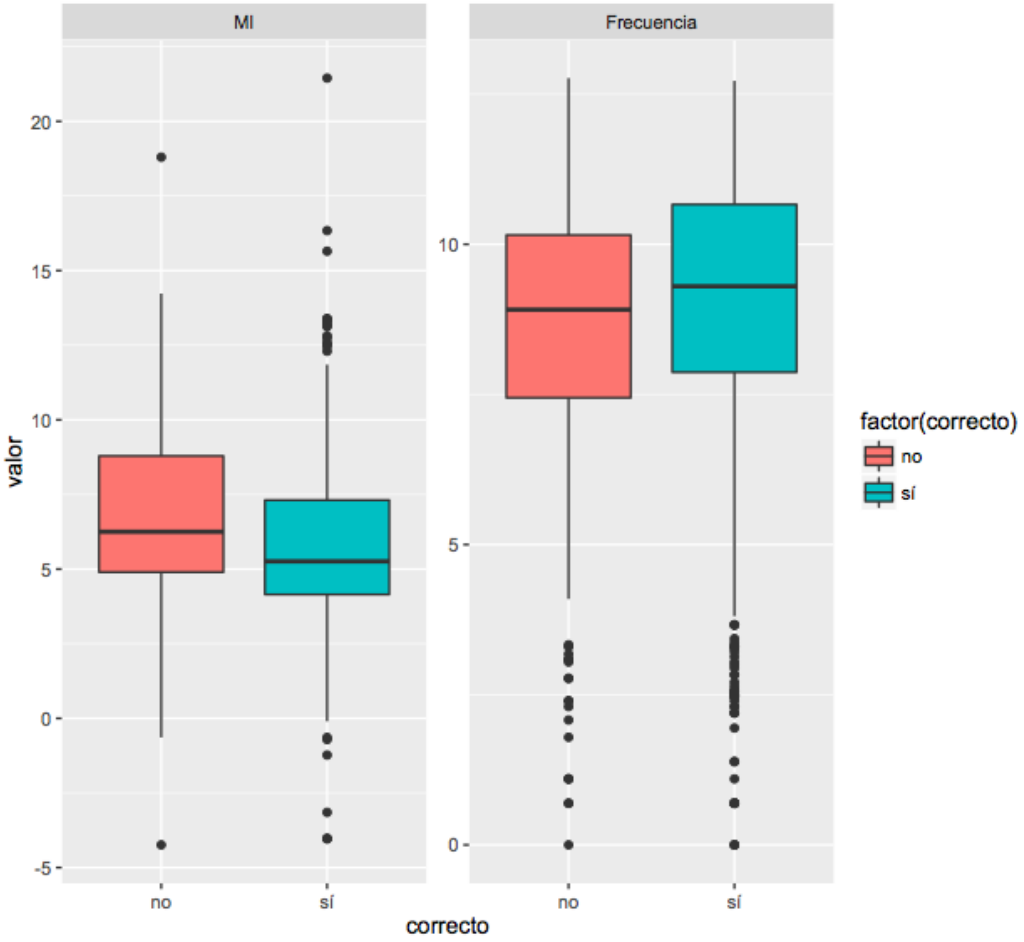
Puesto que se pretende comprobar la influencia de estas dos medidas de asociación en la corrección de las colocaciones producidas por los aprendices, se ha optado por emplear un modelo de regresión logística de efectos mixtos (*cf.* Baayen, 2008; Gries, 2015). Por una parte, la regresión logística es adecuada para tratar variables dependientes dicotómicas como la que nos ocupa (correcto/incorrecto). Por otra, un modelo mixto permite dar cuenta del peso que tienen factores aleatorios, como los informantes incluidos en la muestra o las propias colocaciones, que suponen un subgrupo fortuito de todas las colocaciones posibles del español.

4. Resultados

Antes de pasar a exponer los resultados de la regresión propiamente dichos, es interesante revisar las tendencias centrales de colocaciones correctas e incorrectas en cuanto a información mutua y frecuencia de coaparición (figura 2).

FIGURA 2

MI y frecuencia en colocaciones correctas e incorrectas



Puede apreciarse que, en general, se dan tendencias contrarias con respecto a los valores de una y otra medida en las colocaciones correctas y en las incorrectas. Las colocaciones incorrectas se corresponden con colocaciones que en el corpus que hemos usado como referencia muestran valores centrales (mediana, rango intercuartílico) de MI más altos que las correctas. En lo que toca a la frecuencia, la situación es la inversa: las colocaciones correctas presentan valores más altos de coocurrencia que las versiones corregidas de las erróneas.

Como se indicaba en la sección anterior, para comprobar el efecto conjunto de estas dos variables (frecuencia e información mutua) en la probabilidad de obtener una colocación correcta en la producción de aprendices, se ha usado un modelo mixto de regresión logística[3].

3 El *software* empleado para ajustar dicho modelo es el paquete lme4 de R.

Los resultados correspondientes a los factores fijos —esto es, frecuencia, información mutua y la interacción de ambos— pueden verse en el cuadro 1.

CUADRO 1
Factores fijos en el modelo mixto generalizado

	COEF.	ERROR ESTÁNDAR	VALOR Z	P (>|Z|)
(Intersección)	0,769	0,643	1,197	0,231
MI	0,093	0,105	0,895	0,371
Frecuencia	0,265	0,082	3,243	0,001**
MI:Frec	-0,034	0,013	-2,544	0,011*

Se puede apreciar que, entre los factores fijos, solo la frecuencia y la interacción entre información mutua y frecuencia resultan estadísticamente significativos. La frecuencia de una colocación presenta una correlación positiva con la probabilidad de que esta sea correcta en la muestra manejada. En cuanto a la interacción entre frecuencia e información mutua, su coeficiente negativo indica que estas dos dimensiones tienen efectos opuestos en la variable de respuesta.

Para comprobar la correlación entre las probabilidades predichas por el modelo y los valores observados en la muestra se han calculado los índices D_{xy} de Somers y C, siguiendo a Baayen (2008: 204, 281). Ambos índices arrojan valores altos (C=0,97; D_{xy}=0,94), con lo que se puede afirmar que hay una correspondencia notable entre las probabilidades predichas por el modelo para el parámetro colocación correcta/incorrecta y los valores efectivamente observados en la muestra.

Además, para determinar la proporción de varianza explicada por los factores fijos y los factores aleatorios, Gries (2015) sugiere adoptar el método propuesto por Nakagawa y Schielzeth (2013). Se obtiene así la proporción de la varianza explicada por los factores fijos (en nuestro caso, $R^2_{marginal}$= 0,043) y el total del modelo ($R^2_{condicional}$=0,475). Se constata de esta manera que el modelo empleado da cuenta de menos de la mitad de la varianza, y la proporción explicada por los efectos fijos (frecuencia, información mutua y la interacción de ambas) es todavía más pequeña (no llega al 5%).

5. Discusión

En general, las tendencias observadas en cuanto a la frecuencia y a la información mutua de las colocaciones correctas e incorrectas de la muestra parecen estar en consonancia con los resultados de los estudios previos reseñados en el apartado 2. Así, puede afirmarse que las

colocaciones erróneas producidas por aprendices son desviaciones con respecto a colocaciones que tienden a presentar valores de información mutua más altos que las correctas. Esto podría tomarse como evidencia de que las colocaciones con MI alta son más problemáticas para los aprendices. Por el contrario, si atendemos a la frecuencia de coaparición de los miembros de las colocaciones estudiadas, encontramos que las correctas tienden a presentar valores más altos que las versiones "canónicas" o corregidas de las incorrectas.

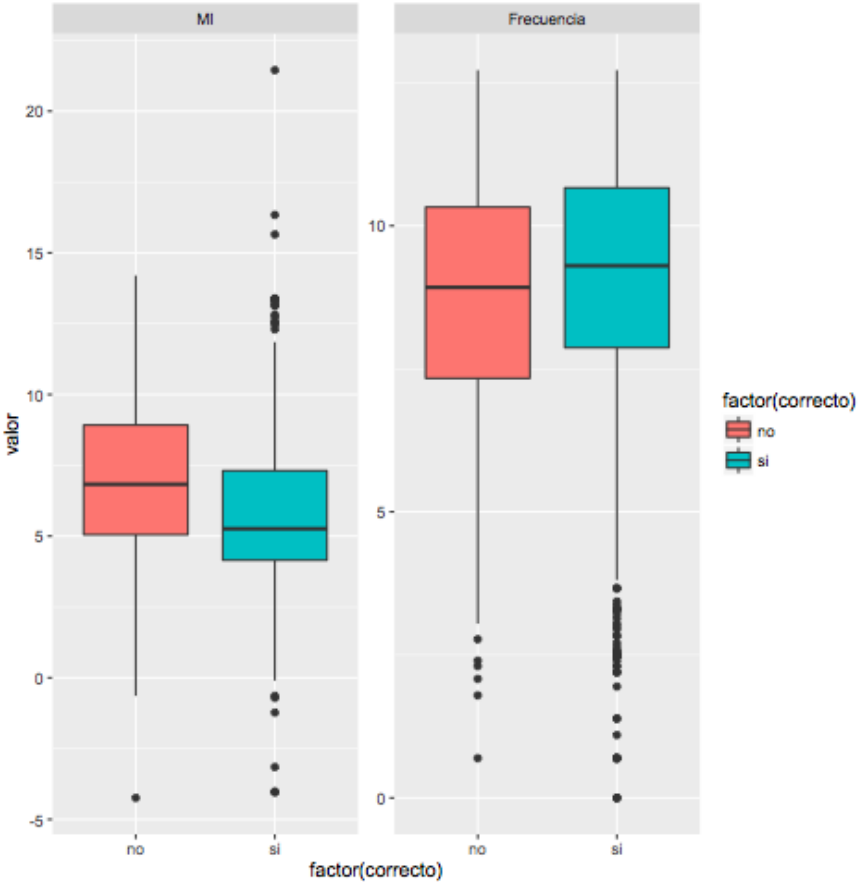
Como se apunta más arriba, las colocaciones incorrectas de la muestra contenían tanto errores de tipo léxico como errores de tipo gramatical. Podría ser objetable la decisión de agrupar ambos tipos de error, si se argumentase que cuando los aprendices producen colocaciones del tipo **interrumpir una ley* es porque no han tenido la suficiente exposición a secuencias como *violar una ley, quebrantar una ley*, etc., mientras que, cuando producen una colocación del tipo *tener *el sexo*, lo hacen porque desconocen el funcionamiento de un patrón gramatical independiente de secuencias léxicas concretas. En ambos casos, sin embargo, la tendencia en cuanto a los dos parámetros estudiados es similar, como se muestra en las figuras 3 y 4, donde se consideran por separado colocaciones afectadas exclusivamente por errores de tipo gramatical (n=159) y aquellas que contienen algún tipo de error léxico (n=205).

Para entender mejor la incidencia de los parámetros estudiados en la producción de los aprendices es útil la revisión cualitativa de algunas colocaciones con distintos valores de frecuencia y MI. Revisemos en primer lugar esta última medida, que, como se apuntaba, tiende a ser más alta en el caso de colocaciones erróneas. Algunos ejemplos de errores en colocaciones de MI alta son **interrumpir una ley* en lugar de *violar una ley* (MI=11,75), **el hielo descongela* por *el hielo se derrite* (MI=11,57) o **marca diacrítica* en lugar de *acento diacrítico* (MI=13,73). Probablemente el último sea uno de los casos más representativos de colocaciones con MI alta: *diacrítico* es un adjetivo que previsiblemente aparecerá en el contexto de *acento* en una proporción considerable de sus ocurrencias, ya que la lista de candidatos a ser modificados por él no parece muy larga (aparte de *acento, tilde, signo*…). Curiosamente, en este caso el aprendiz comete un error al elegir la base, que es el elemento no restringido. En los otros dos casos el error afecta al colocativo.

Podría decirse que, en el caso de los ejemplos de colocaciones con MI alta examinados, el sentido del colocativo está de algún modo previsto por la base, lo que quizá pueda relacionarse con la observación hecha por Ellis y otros (2008), de acuerdo con la cual las combinaciones de MI alta son especialmente coherentes. Ya hemos visto la restringida lista de nombres de los que se puede predicar un carácter 'diacrítico'. En cuanto a *violar una ley*, cabe notar que el nombre expresa un sentido que implica un cumplimiento y el verbo es la forma convencional de expresar que tal implicación no se da. De modo similar, el sentido de *derretirse* parece aplicable solo a objetos sólidos susceptibles de licuarse, lo que en el caso del hielo resulta esperable. Ahora bien, sus altos valores de MI probablemente no se deban únicamente a esta "concordancia" semántica. Es necesario tener en cuenta además que existe una ulterior res-

FIGURA 3

Frecuencia y MI en colocaciones correctas e incorrectas (solo errores léxicos)

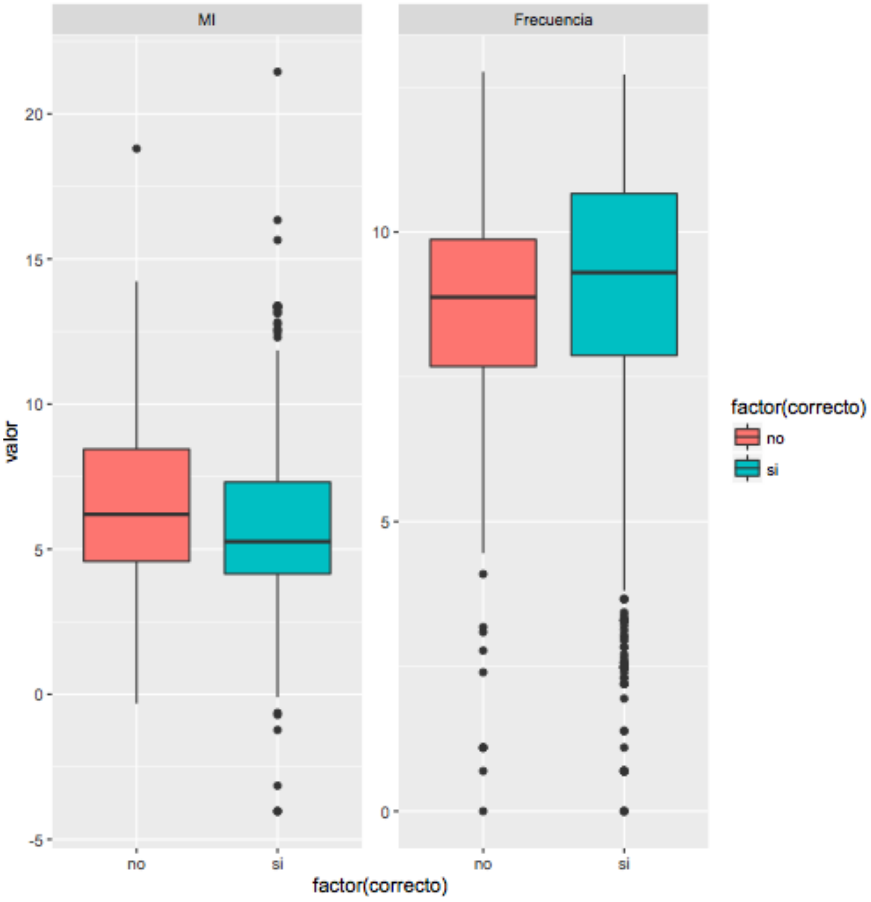


tricción en cuanto a las unidades léxicas que pueden expresar los significados en cuestión dada una determinada base. De este último aspecto se deriva que estas producciones en concreto se hayan considerado erróneas: los aprendices expresan combinaciones de sentidos coherentes, pero mediante elecciones léxicas que para un hablante nativo resultan inadecuadas. Si efectivamente un valor alto de información mutua indica que al menos uno de los miembros de la colocación tiene unas posibilidades combinatorias más restringidas que los constituyentes de colocaciones con valores más bajos, es lógico pensar que las colocaciones de MI alta sean más propicias al error.

En lo que atañe a la frecuencia, ya hemos visto que en general la corrección colocacional está asociada a valores altos. Ejemplos de colocaciones con frecuencia alta en la muestra son *tener derecho, tener x año(s)* o *tener (un) problema(s)*. Estas tres colocaciones figuran entre las seis más recurrentes en el corpus y no tienen un valor especialmente llamativo de información mutua (el mayor —5,02— corresponde a *tener [un] problema[s]*). El hecho de que tengan valores de información mutua más bajos sugiere que se trata de colocaciones for-

FIGURA 4

Frecuencia y MI en colocaciones correctas e incorrectas (solo errores gramaticales)



madas por palabras que no están tan ligadas al contexto concreto de la colocación como las anteriores, sino que su versatilidad distribucional es considerablemente mayor. Esto último probablemente tiene que ver con que sus componentes sean formas frecuentes del español. A este respecto es llamativo el caso de *tener*, el tercer verbo más frecuente de nuestro corpus de referencia tras *ser* y *haber*. Las combinaciones en cuestión no son menos coherentes que las anteriores desde el punto de vista semántico[4]: en los tres casos estamos ante nombres

4 El hecho de que no puedan establecerse diferencias claras en cuanto a la coherencia de colocaciones de MI alta y baja en la muestra estudiada aquí, en contra de lo constatado por Ellis y otros (2008), está en relación con la naturaleza de dicha muestra. Esta incluye únicamente colocaciones identificadas mediante los criterios de la lexicología explicativa y combinatoria (Mel'čuk y otros, 1995). Cabe esperar, pues, que todos los colocativos expresen sentidos "relevantes" o previstos por su base. Ellis y otros (2008), por el contrario, estudian bigramas que no necesariamente tienen que constituir una colocación en sentido fraseológico o siquiera un sintagma.

relacionales unidos a su primer actante semántico mediante un verbo que expresa precisamente una relación. Las diferencias, por tanto, hay que buscarlas en la alta frecuencia de las propias combinaciones en el español general, en su alta recurrencia en el corpus de aprendices, lo que indica cierta seguridad en su uso por parte de estos hablantes[5], y, quizá, en la alta frecuencia de sus constituyentes tomados individualmente, lo cual sugiere una mayor versatilidad contextual.

Con respecto al hecho de que se observen las mismas tendencias en cuanto a los valores de información mutua y frecuencia de las colocaciones de la muestra, aun cuando se consideren de forma separada errores léxicos y errores gramaticales, cabe hacer varias consideraciones. En primer lugar, se ha señalado que las unidades fraseológicas, en la medida en que son combinaciones prefabricadas, liberan al hablante de la aplicación de una serie de reglas en teoría necesarias, si las combinaciones en cuestión se produjesen *ex novo* (*cf.*, por ejemplo, Hakuta (1976: 333) o Nattinger y de Carrico (1992: 27-28))[6]. Así, el hablante tendría únicamente que emitir una secuencia almacenada y más o menos lista para usar. En segundo lugar, varios trabajos señalan el comportamiento idiosincrático de ciertos rasgos gramaticales en construcciones de tipo fraseológico, por ejemplo, el uso de determinantes (Alonso Ramos, 2004: 197 y ss.; García Salido, 2016: 371; o Laca, 1999: 918). Teniendo esto en cuenta, tiene sentido que las colocaciones que presentan más dificultades para los aprendices (es decir, colocaciones infrecuentes en el input y con valores altos de información mutua) lo hagan tanto desde el punto de vista léxico como del gramatical, porque, por una parte, suponen una hipótesis en cuanto a la combinación de dos unidades léxicas por parte del aprendiz, que puede o no coincidir con la solución convencional para los nativos y, por otra, conllevan un esfuerzo mayor que puede afectar al resultado de la producción. Así, una secuencia que no forma parte del repertorio de los aprendices tiene que construirse (i) buscando las formas léxicas correspondientes y (ii) aplicando los mecanismos necesarios para su combinación, en lugar

5 En este sentido cabe notar que los ejemplos comentados de *tener* se dan en construcciones con verbo de apoyo (*vid.* Alonso Ramos, 2004) y que estas construcciones son incluso algo más frecuentes en los textos de aprendices que en los de nativos (García Salido, 2014), aunque presentan una variedad menor —es decir, los aprendices repiten más las mismas combinaciones de lemas—. Esto podría indicar que este tipo de colocaciones es una especie de forma comodín o *teddy bears* léxicos (Hasselgren, 1994) para los aprendices.

6 Se ha discutido si el efecto de poseer un repertorio de secuencias prefabricadas es el mismo en el desarrollo de patrones estructurales en la L1 y en la L2 (para una revisión de las evidencias a favor y en contra pueden consultarse Durrant (2008: 45-57), Vincze (2015: 68-71) o Wray (2002: 205-206), y las referencias allí citadas). Independientemente de que contar con un repertorio de secuencias prefabricadas sirva o no a los aprendices de una L2 para abstraer y aplicar patrones formales más generales, el hecho de que tales secuencias incluyan rasgos gramaticales podría explicar por qué las más difíciles de adquirir desde un punto de vista léxico sean también más problemáticas en cuanto a su gramática.

de (iii) simplemente emitiendo una secuencia ya almacenada en su memoria, con el coste de procesamiento adicional que la suma de (i) y (ii) supone frente a (iii).

Además de las tendencias observadas, mediante el modelo de regresión aplicado hemos podido constatar que tanto la frecuencia de la colocación como su MI inciden significativamente en la probabilidad de corrección de las colocaciones producidas por los aprendices —en el caso de la última, solo cuando interactúa con la primera, moderando su efecto—. Ahora bien, el citado modelo, que solo tiene en cuenta frecuencia e MI, tiene escaso poder predictivo y solo arroja factores significativos cuando se aplica a la totalidad de la muestra (y no, por tanto, a los conjuntos compuestos de colocaciones correctas más colocaciones con errores solo gramaticales o de correctas más colocaciones con errores léxicos). Es más, si se compara la proporción de varianza explicada por el modelo en general con la proporción atribuible al efecto de frecuencia y MI, cabe concluir que los factores aleatorios —es decir, los propios aprendices y las colocaciones que aparecen en la muestra— tienen mucho mayor peso que la frecuencia, la MI y la interacción de ambas.

El pequeño efecto de las medidas de asociación en la dicotomía correcto/incorrecto posiblemente se debe a que en el dominio de una determinada colocación por parte de un aprendiz concreto intervienen más factores que la frecuencia que esta presenta en su input o la probabilidad de que un miembro de la colocación aparezca en el contexto del otro. Habría que contar además con las asociaciones que el aprendiz pueda hacer con colocaciones de su L1, sus intereses por determinados temas, las colocaciones que se trabajan de manera explícita durante su instrucción, etc. Así pues, parece una tarea imposible predecir la probabilidad con que un aprendiz produce una colocación correcta a partir exclusivamente de un determinado umbral de frecuencia o información mutua. En este sentido, también cabe observar que los valores de estas dos medidas presentan un considerable grado de solapamiento tanto en el grupo de colocaciones incorrectas como en el de colocaciones correctas. Por ello, que tales factores muestren cierta significatividad en relación con la producción de los aprendices puede considerarse ya, en cierta medida, satisfactorio. Ahora bien, serían deseables futuras investigaciones que, junto con las medidas estudiadas, consideren factores adicionales que aquí no se han podido tratar para verificar la significatividad de la frecuencia y la información mutua. Esto nos lleva directamente a las limitaciones del presente estudio.

La primera de estas limitaciones tiene que ver con las variables que un estudio de corpus permite controlar: los aprendices que han contribuido al corpus tienen seguramente historiales de aprendizaje diferentes y, si bien la muestra está limitada a los que tienen un dominio del español hasta cierto punto homogéneo, resulta imposible incluir información relativa al tipo de instrucción que cada informante ha recibido. Esto contrasta, por ejemplo, con lo que ocurre con diseños experimentales que miden la competencia en cuanto a un determinado fenómeno lingüístico de un determinado grupo antes y después de ser expuestos al mismo tipo de instrucción. Tal limitación es consecuencia del propio enfoque adoptado, que, en con-

trapartida, ofrece otras ventajas, como el acceso a una muestra más representativa, el análisis de una producción realizada en condiciones menos artificiales, etc. (*cf.* Granger, 1998: 4-5).

Otro factor que no se ha podido tener en cuenta aquí y que se ha mostrado fundamental en la explicación de errores colocacionales, especialmente los léxicos, es la influencia de la L1 de los aprendices (*cf.*, por ejemplo, Nesselhauf (2005: 181) o, para el caso del español, Vincze y otros (en prensa)). Esto no ha sido posible, en primer lugar, porque la anotación del corpus de aprendices manejado solo incluye información relativa a la congruencia de una colocación con respecto a la L1 en el caso de colocaciones incorrectas. Además, para establecer una correspondencia entre el input y la producción de los aprendices, como se ha hecho con la frecuencia y la MI, sería necesario no solo extender la información relativa a la congruencia con la L1 a las colocaciones correctas, sino modificar la ya aplicada a las incorrectas indicando la congruencia respecto a la colocación meta (esto es, a la versión corregida) y no al error efectivamente producido por el aprendiz.

Lo anterior implicaría apoyarse nuevamente en las hipótesis de los anotadores en cuanto a la correspondencia entre la producción de los aprendices y su equivalente correcto. Esto en sí mismo tiene consecuencias con respecto a los datos que se han manejado aquí, lo cual también podría suscitar críticas. Así, por ejemplo, en el caso de **interrumpir una ley* citado más arriba, los anotadores optaron por la corrección *violar una ley*, pero otras opciones de corrección como *quebrantar una ley, incumplir una ley*, etc., parecen también válidas. Optar por una opción u otra repercute en los resultados del estudio, en la medida en que las distintas alternativas presentan valores de frecuencia e información mutua distintos. Ahora bien, en muchos de los casos de errores léxicos hay indicios adicionales que sugieren una corrección en particular: parece muy probable, por ejemplo, que una secuencia como **agregar un problema* esté construida sobre el modelo de *agravar un problema* tanto por el sentido con el que está usada como por la similitud fónica entre los dos verbos. Asimismo, en los casos de errores gramaticales como *tener *el sexo* en lugar de *tener sexo*, la labor interpretativa en la corrección ha sido mínima. En todo caso, como se defiende más arriba, si el propósito es establecer la relación del input con la producción correcta o incorrecta de los aprendices, son necesarias hipótesis con respecto a las correspondencias entre uno y otra.

Por último, podría ponerse en cuestión hasta qué punto es representativo del input que han recibido los aprendices el corpus tomado como referencia y del que se han extraído los datos de frecuencia de las colocaciones estudiadas. Esta es una objeción que puede extenderse a la totalidad de los trabajos que usan datos de corpus de referencia y los ponen en relación con la producción de hablantes no nativos, y en última instancia, a cualquier estudio de corpus. Como señala, por ejemplo, Hoey (2005: 14), no está justificado identificar un determinado corpus de referencia con los datos lingüísticos a los que un hablante concreto ha estado expuesto. Ahora bien, es razonable pensar que dicho corpus presentará relaciones entre elementos lingüísticos de una manera en cierta medida proporcional a la que pueden

experimentar los hablantes de la lengua en cuestión. Es decir, si en el corpus de referencia *violar una ley* tiene una frecuencia menor que *tener padre* y *tener* unas posibilidades combinatorias mucho mayores que *violar*, lo esperable es que esto se verifique también en el input de la mayoría de hablantes de español. En el caso de los aprendices, cabe asumir una situación similar, aunque su exposición probablemente haya sido menor. Esto es así porque (i) la mayoría de los aprendices de la muestra han pasado temporadas en un país hispanohablante (mayoritariamente España; *cf.* Vincze y otros (en prensa)) y (ii), previsiblemente, en clase de español tendrán hasta cierto punto acceso al tipo de lenguaje que el corpus representa, en la medida en que se les proporcionen materiales auténticos.

6. Conclusiones

El presente estudio ha querido investigar cómo influyen la frecuencia y el grado de asociación entre los elementos de una colocación en la corrección con la que la emplean los aprendices de español. Resultados de diversos estudios previos, centrados en aprendices de inglés, nos hacían partir de la hipótesis de que la frecuencia sería un elemento facilitador, mientras que las colocaciones con información mutua más alta presentarían un grado de dificultad mayor y, por tanto, una mayor probabilidad de aparecer de forma incorrecta en la producción de los aprendices. Los resultados del estudio van efectivamente en esta dirección: la frecuencia presenta una correlación positiva con la probabilidad de acierto y la información mutua, aunque no presenta un efecto significativo por sí sola, modera el efecto de la frecuencia.

Ahora bien, a pesar de su significatividad, el efecto de estos dos factores es pequeño, con lo cual resulta imposible predecir si una colocación va a ser producida correctamente o no basándose exclusivamente en ellos. Esto probablemente sea una consecuencia del hecho de que en la competencia colocacional de los aprendices intervengan más factores que la frecuencia o el grado de asociación entre los miembros de las colocaciones de la lengua meta. La incidencia de otros factores distintos de la asociación medida en términos de frecuencia, tales como la congruencia de las colocaciones producidas con respecto a sus equivalentes en la L1 del aprendiz, es, pues, una vía que futuros estudios deberían explorar. Como se ha señalado en la discusión, no obstante, no es posible controlar todos los factores que intervienen en la adquisición de colocaciones —o de vocabulario en general— en un estudio de corpus.

Los resultados proporcionan, pues, cierta evidencia de que las medidas de asociación derivadas de la frecuencia inciden en la producción colocacional de los aprendices. En inglés existen ya una serie estudios que apuntan en la misma dirección. Encontrar nuevos elementos que justifiquen postular este tipo de relación en el aprendizaje de español tendría una serie de implicaciones, por ejemplo, con respecto a la selección del léxico en programas de ELE/2. Es una idea hasta cierto punto asumida que la frecuencia determina en un grado considerable la utilidad del léxico en general y que debería ser empleada como criterio para deter-

minar qué vocabulario (incluidas secuencias pluriverbales) ha de tratarse en la enseñanza de una lengua y en qué orden (Ferrando Aramo, 2012: 360-361; Martinez, 2013; Nation, 2001: 13-15, 329). Si tenemos que la frecuencia es, además, un elemento facilitador y que la información mutua tiene el efecto contrario, contamos también con elementos de juicio para identificar colocaciones previsiblemente difíciles para los aprendices sobre las que se debería hacer hincapié, al menos en ciertos niveles. Así, es posible que la utilidad pese más en niveles iniciales y se decida no incluir en ellos colocaciones infrecuentes. Ahora bien, en niveles más avanzados, podría ser razonable tratar colocaciones no especialmente frecuentes, pero con una MI alta, lo cual indicaría, por un lado, que son combinaciones prominentes en el repertorio colocacional de los hablantes nativos (*cf.* Ellis y otros, 2008) y, por otro, que son previsiblemente difíciles para los no nativos.

Por último, la existencia de una relación entre las dimensiones consideradas y la dificultad colocacional abre puertas a aplicaciones de evaluación automática del léxico de los aprendices, tal como proponen Granger y Bestgen (2014: 248). Si efectivamente valores más altos de MI y valores bajos de frecuencia son indicio de colocaciones más difíciles, los textos que muestren estos rasgos deberían corresponderse con aprendices de niveles más avanzados. Antes de embarcarse en el desarrollo de este tipo de aplicaciones para el español, no obstante, se antojan necesarios estudios (seudo)longitudinales que confirmen la relación y el poder predictivo de las medidas de asociación que se han discutido aquí.

7. Bibliografía citada

Alonso Ramos, Margarita, 1993: *Las funciones léxicas en el modelo lexicográfico de I. Mel'čuk*. Tesis doctoral, UNED.

Alonso Ramos, Margarita, 1994-1995: "Hacia una definición del concepto de colocación: de J.R. Firth a I.A. Mel'čuk", *Revista de Lexicografía* 1, 9-28.

Alonso Ramos, Margarita, 2004: *Las Construcciones con Verbo de Apoyo*, Madrid: Visor Libros.

Alonso Ramos, Margarita, en prensa: "Looking at Spanish learner corpus research: Achievements and challenges" en Margarita Alonso Ramos (ed.): *Spanish Learner Corpus Research: Current trends and Future Perspectives*, Amsterdam/Philadelphia: John Benjamins.

Alonso Ramos, Margarita, Leo Wanner, Nancy Vázquez Veiga, Orsolya Vincze, Estela Mosqueira Suárez y Sabela Prieto González, 2010a: "Tagging collocations for learners" en Sylviane Granger y Magali Paquot (eds.): *Elexicography in the 21st Century: New Challenges, New Applications. Proceedings of eLex2009*, Cahiers du cental 7, Louvain-la Neuve: Presses Universitaires de Louvain, 375-380

Alonso Ramos, Margarita, Leo Wanner, Orsolya Vincze, Gerard Casamayor, Nancy Vázquez Veiga, Estela Mosqueira Suárez y Sabela Prieto González, 2010b: "Towards a motivated annotation schema

of collocation errors in learner corpora" en Nicoletta Calzolari y otros (eds.): *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10), 19-21 May 2010, Valletta, Malta*, European Language Resources Association (ELRA), 3209-3214.

Altenberg, Bengt, y Sylviane Granger, 2001: "The Grammatical and Lexical Patterning of MAKE in Native and Non-Native Student Writing", *Applied Linguistics* 22, 173-195.

Baayen, Harald, 2008: *Analyzing linguistic data with R*, Cambridge: Cambridge University Press.

Bestgen, Yves, y Sylviane Granger, 2014: "Quantifying the development of phraseological competence in L2 English writing: An automated approach", *Journal of Second Language Writing* 26, 28-41.

Buckingham, Louisa, 2008: "Spanish verb support constructions from a learner perspective", *Elia: Estudios de Lingüística Inglesa Aplicada* 8, 151-179.

Cowie, Anthony P., 1981: "The treatment of collocations and idioms in Learners' dictionaries", *Applied Linguistics* 2(3), 223-235.

Durrant, Philip, 2008: *High frequency collocations and second language learning*. Tesis doctoral, University of Nottingham.

Durrant, Philip, 2014: "Corpus frequency and second language learners' knowledge of collocations", *International Journal of Corpus Linguistics* 19 (4), 443-477.

Durrant, Philip, y Norbert Schmitt, 2009: "To what extent do native and non-native speakers maker use of collocations?", *IRAL* 47, 157-177.

Ellis, Nick, Rita Simpson-Vlach y Carson Maynard, 2008: "Formulaic Language in Native and Second Language speakers: Psycholinguistics, Corpus Linguistics, and TESOL", *TESOL Quarterly* 42 (3), 375-396.

Evert, Stefan, 2008: "Corpora and collocations" en Anke Lüdeling y Merja Kytö (eds.): *Corpus Linguistics. An International Handbook*, Berlín: Mouton de Gruyter [versión extendida disponible en http://purl.org/stefan.evert/PUB/Evert2007HSK_extended_manuscript.pdf, fecha de consulta: 15 de abril de 2016].

Ferrando Aramo, Verónica, 2012: *Aspectos teóricos y metodológicos para la compilación de un diccionario combinatorio destinado a estudiantes de E/LE*. Tesis doctoral, Universitat Rovira i Virgili.

Firth, John R., 1957 [1951]: "Modes of Meaning" en *Papers in Linguistics 1934-1951*, Londres: Oxford University Press, 190-216.

García Salido, Marcos, 2014: "O uso de construcións con verbos soporte en aprendices de español como lingua estranxeira e en falantes nativos", *Cadernos de Fraseoloxía Galega* 16, 181-198.

García Salido, Marcos, 2016: "Error Analysis of Support Verb Constructions in Written Spanish Learner Corpora", *Modern Language Journal*, 100(1), 362-376.

Gilquin, Gaëtanelle, 2007: "To Err Is Not All: What Corpus and Elicitation Can Reveal About the Use of Collocations By Learners", *Zeitschrift Für Anglistik Und Amerikanistik* 55 (3), 273-291.

Granger, Sylviane, e Yves Bestgen, 2014: "The use of collocations by intermediate vs. advanced non-native writers: A bi-gram based study", *IRAL* 52(3), 229-252.

Granger, Sylviane, 1998: "The computer learner corpus: a versatile new source of data for SLA research" en Sylviane Granger (ed.): *Learner English on Computer* Londres / Nueva York: Longman, 1-19.

Gries, Stefan, 2015: "The most under-used statistical method in corpus linguistics: multilevel (and mixed-effect) models", *Corpora* 10 (1), 95-125.

Hakuta, Kenji, 1976: "Becoming bilingual: a case study of a Japanese child learning English", *Language Learning* 26, 321-351.

Halliday, M. A. K., 1966: "Lexis as a linguistic level" en C. E. Bazell, J. C. Catford, M. A. K. Halliday y R. H. Robins (eds.): *In Memory of J.R. Firth*, Londres: Longman, 148-162.

Hasselgren, Angela, 1994: "Lexical teddy bears and advanced learners: A study into the ways Norwegian students cope with English vocabulary", *International Journal of Applied Linguistics* 4 (2), 23-260.

Hausmann, Franz Josef, 1989: "Le dictionnaire de collocations" en Franz Josef Hausmann, Oskar Reichmann, Hans Ernst Wiegand y Ladislav Zgusta (eds.): *Wörterbücher : ein internationales Handbuch zur Lexikographie. Dictionaries. Dictionnaires*, Berlín: Mouton de Gruyter, 1010-1019.

Higueras, Marta, 2006: *Las colocaciones y su enseñanza en la clase de ELE*, Madrid: Arco Libros.

Hoey, Michael, 2005: *Lexical priming. A new theory of words and language*, Londres: Routledge.

Howarth, Peter A., 1998: "The phraseology of learners' academic writing" en Anthony P. Cowie (ed.): *Phraseology, Theory, Analysis and Applications*, Oxford: Oxford University Press, 161-186.

Kilgarriff, Adam, e Irene Renau, 2013: "esTenTen, a Vast Web Corpus of Peninsular and American Spanish", *Procedia* 95, 12-19.

Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý y Vít Suchomel, 2014: "The Sketch engine: ten years on", *Lexicography* 1 (1), 7-36.

Laca, Brenda, 1999: "Presencia y Ausencia de Determinante" en Ignacio Bosque y Violeta Demonte (eds.): *Gramática descriptiva de la lengua española*, Madrid: Espasa-Calpe, 891-928.

Lexical Computing, 2015: "Statistics used in the Sketch-Engine" [disponible en línea en https://www.sketchengine.co.uk/wp-content/uploads/ske-stat.pdf, fecha de consulta: 1 de junio de 2016].

Lorenz, Gunter, 1999: *Adjective Intensification - Learners versus Native Speakers: A Corpus Study of Argumentative Writing*, Amsterdam: Rodopi.

Lozano, Cristóbal, 2009: "CEDEL2: Corpus Escrito del Español L2" en Carmen M. Bretones Callejas y otros (eds.): *Applied Linguistics Now: Understanding Language and Mind / La lingüística aplicada hoy: Comprendiendo el lenguaje y la mente*, Almería: Universidad de Almería, 197-212.

Lozano, Cristobal, y Amaya Mendikoetxea, 2013: "Learner corpora and second language acquisition: The design and collection of CEDEL" en Ana Díaz-Negrillo, Nicolas Ballier y Paul Thompson (eds.): *Automatic Treatment and Analysis of Learner Corpus Data*, Amsterdam/Philadelphia: John Benjamins, 65-100.

Martinez, Ron, 2013: "A framework for the inclusion of multi-word expressions in ELT", *ELT Journal* 67, 184-198.

Mel'čuk, Igor, 1960: "O terminax 'ustojćivost' i 'idiomatićnost'", *Voprosy jazykoznanija* 4, 73-80.

Mel'čuk, Igor, 2012: "Phraseology in the language, in the dictionary, and in the computer", *Yearbook of Phraseology* 3 (1), 31-56.

Mel'cuk, Igor, André Clas y Alain Polguere, 1995: *Introduction à la lexicologie explicative et combinatoire*, Louvain-la-Neuve: Duculot.

Nation, Paul, 2001: *Learning vocabulary in another language*, Cambridge: Cambridge University Press.

Nakagawa, Shinichi, y Holger Schielzeth, 2013: "A general and simple method for obtaining R2 from generalized linear mixed-effects models", *Methods in Ecology and Evolution* 4 (2), 133-142.

Nattinger, James R., y Jeanett S. de Carrico, 1992: *Lexical phrases and Language Teaching*, Oxford: Oxford University Press.

Nesselhauf, Nadia, 2005: *Collocations in a Learner Corpus*, Amsterdam/Philadelphia: John Benjamins.

Orol González, Ana, 2015: "Construcción de una lista de colocaciones para medir la competencia colocacional", *E-Aesla* 1 [disponible en http://cvc.cervantes.es/lengua/eaesla/pdf/01/09.pdf, fecha de consulta: 24 de mayo de 2016].

Pawley, Andrew, y Frances H. Syder, 1983: "Two puzzles for linguistic theory: nativelike selection and nativelike fluency" en J. C. Richards y R. W. Schmidt (eds.): *Language and communication*, Harlow: Longman, 191-226.

Pérez Serrano, Mercedes, 2014: "Análisis de errores colocacionales en un corpus de aprendientes de ELE", *MarcoELE: Revista de Didáctica de Español Como Lengua Extranjera* 19 [disponible en http://marcoele.com/analisis-de-errores-colocacionales-en-un-corpus-de-aprendientes-de-ele, fecha de consulta: 1 de septiembre de 2015].

Pérez Serrano, Mercedes, 2015: *Un enfoque léxico a prueba: Efectos de la instrucción en un aprendizaje de las colocaciones léxicas*. Tesis doctoral, Universidad de Salamanca.

Sinclair, John M. (ed.), 1987: *Looking Up - An account of the COBUILD Project in lexical computing*, Londres: Harper-Collins.

Uriel Domínguez, Meritxell, 2014: *Las colocaciones en un corpus de aprendices valón y flamenco*. Tesis de Máster, Universitat de Barcelona / Universitat Pompeu Fabra.

Vinzce, Orsolya, 2015: *Learning multiword expression from corpora and dictionaries*. Tesis doctoral, Universidade da Coruña.

Vinzce, Orsolya, y Margarita Alonso Ramos, 2013: "Incorporating frequency information in a collocation dictionary: Establishing a methodology", *Procedia - Social and Behavioral Sciences* 96, 241-248.

Vinzce, Orsolya, Margarita Alonso Ramos y Estela Mosqueira, 2011: "Exploiting a learner corpus for the development of a CALL environment for learning Spanish collocations" en Iztok Kosem y Karmen Kosem (eds.): *Electronic lexicography in the 21st century: New applications for new users. Proceedings of eLex 2011*, Ljubljana: Trojina, Institute for Applied Slovene Studies, 280-285.

Vinzce, Orsolya, Marcos García Salido, Ana Orol y Margarita Alonso Ramos, en prensa: "A corpus study of Spanish as a Foreign Language learners' collocation production" en Margarita Alonso Ramos (ed.): *Spanish Learner Corpus Research: Current trends and Future Perspectives*, Amsterdam/Philadelphia: John Benjamins.

Weinreich, Uriel, 1969: "Problems in the analysis of idioms" en Jaan Puhvel (ed.): *Substance and Structure of Language*, Berkeley / Los Angeles: University of California Press, 23-81.

Wray, Alison, 2002: *Formulaic Language and the Lexicon*, Cambridge: Cambridge University Press.