Revista Signos. Estudios de Lingüística ISSN 0718-0934
© 2015 PUCV, Chile DOI: 10.4067/S0718-09342015000200002 48(88) 174-196

Hacia una extracción automática de colocaciones en definiciones de verbos de un diccionario explicativo en español

Toward an Automatic Extraction of Collocations in Verb Definitions from a Spanish Explanatory Dictionary

Noé Alejandro Castro-Sánchez
Centro Nacional de Investigación y Desarrollo Tecnológico
México
ncastro@cenidet.edu.mx

Irasema Cruz Domínguez
Instituto de Investigaciones Filológicas
Universidad Nacional Autónoma de México
México
irasema.cd@gmail.com

Grigori Sidorov
Centro de Investigación en Computación
Instituto Politécnico Nacional
México
sidorov@cic.ipn.mx

Alicia Martínez Rebollar
Centro Nacional de Investigación y Desarrollo Tecnológico
México
amartinez@cenidet.edu.mx

Recibido: 12-VIII-2013 / **Aceptado:** 11-VII-2014

Resumen

En este artículo presentamos un método para identificar colocaciones de manera automática en definiciones de verbos extraídas del diccionario explicativo de la Real Academia Española (RAE) con el fin de probar que las colocaciones pueden identificarse aplicando heurísticas sencillas que consideran solo criterios semánticos en contextos textuales bien estructurados, como es el caso de las definiciones lexicográficas. Los candidatos a colocaciones se caracterizan porque están situados al inicio de las definiciones y tienen como particularidad que la base de la colocación candidata pertenece a la familia léxica del verbo definido (1.347 casos). La evaluación de las combinaciones de palabras obtenidas se realizó de manera semiautomática, considerando criterios estadísticos y sintáctico-semánticos. Ésta arrojó como resultado que el 61% de las combinaciones de palabras extraídas de esta manera son colocaciones, logrando alcanzar una cobertura del 36%.

Palabras Clave: Colocaciones, unidades fraseológicas, diccionario explicativo, extracción automática de colocaciones.

Abstract

In this paper we present a method for identifying collocations in an automatic way in verb definitions extracted from the explanatory dictionary of the Royal Spanish Academy, in order to test that collocations can be identified by applying simple heuristics considering only semantic criteria in well-structured textual contexts, as lexicographic definitions are presented. The method identifies candidates for collocations located at the beginning of the definitions that have a special feature: the base of the candidate collocation belongs to the lexical family of the defined verb (1,347 cases). The evaluation of the obtained word combinations was performed both manually and automatically following various statistical and syntactic-semantic criteria. The results of our experiment show that 61% of the extracted verb combinations are collocations, obtaining a recall of 36%.

Key Words: Collocations, phraseological units, explanatory dictionaries, automatic extraction of collocations.

INTRODUCCIÓN

Los hablantes usan expresiones y combinaciones lingüísticas que tienen como fin codificar su entorno. Desde siempre, han usado tanto estructuras determinadas por las reglas de su lengua, como construcciones prefabricadas. Así, a finales del siglo XX se acuña el término 'fraseología', (sub)disciplina que estudia construcciones lingüísticas denominadas 'unidades fraseológicas' (UFs), estas se caracterizan por combinar dos o más palabras cuyas propiedades principales son la fijación y la idiomaticidad (Školníková, 2010).

Dentro de las UFs, se han identificado diferentes tipos de combinaciones de palabras, por lo que se han hecho diferentes propuestas de clasificación. Sin embargo, el principal problema al que se enfrentan las propuestas es que los grupos identificados tienen un comportamiento heterogéneo y los límites de cada una son difusos. Ante este panorama, buscamos delimitar la noción de colocación que nos permita tener determinadas características para facilitar su identificación. Por ello, en este trabajo nos enmarcamos en la propuesta de Corpas (1996) y Koike (2001).

Esta autora propone la existencia de tres tipos de UFs: Colocaciones, Locuciones y Enunciados Fraseológicos. El primero refiere a las combinaciones prefabricadas en la norma que se caracteriza por cierta fijación interna. Estas son las construcciones de las que nos ocupamos en esta investigación y, por tanto, profundizaremos más adelante. El segundo refiere a unidades fraseológicas del sistema lingüístico que presentan una mayor fijación interna y externa, también presentan mayor unidad de significado y, generalmente, funciona como elemento oracional. El último, se caracteriza por remitir a enunciados completos en el acto de habla y, además, por su fijación interna y externa. De acuerdo con la clasificación de esta autora, el primero pertenece esencialmente a una fijación de la norma, el segundo a una fijación del sistema y el tercero a una fijación del habla.

Así, el uso y conocimiento de las colocaciones refleja la competencia lingüística de los hablantes, resaltando su importancia e impacto en diferentes áreas, por mencionar algunas, la enseñanza de segundos idiomas, la lexicografía, la traducción automática, así como la extracción automática de términos.

Este trabajo tiene como objetivo presentar un método para identificar colocaciones de manera automática a partir del procesamiento de las definiciones de verbos en un diccionario explicativo con la finalidad de probar que las colocaciones pueden identificarse aplicando heurísticas sencillas considerando solo criterios semánticos en contextos textuales bien estructurados, como es el caso de las definiciones lexicográficas. Específicamente, el método extrae combinaciones de palabras situadas al inicio de las definiciones, presentándolas como candidatos a colocaciones siempre y cuando exista una relación léxica entre el verbo definido y la base de la colocación. Los resultados obtenidos fueron evaluados considerando los criterios sintáctico-semánticos de Koike (2001) y Školníková (2010), y demuestran que en la mayoría de los casos, las combinaciones extraídas se comportan como colocaciones.

El artículo se organiza de la siguiente manera: en la sección 1 abordamos los principales formalismos que explican y estudian el concepto de colocación; en la sección 2, describimos los datos de nuestros experimentos, los criterios que seguimos para evaluarlos y el método que utilizamos y; finalmente, en la sección 3, comentamos los resultados obtenidos.

1. Concepto de colocación

El término 'colocación' ha sido estudiado desde diversos enfoques (Corpas, 1996; Bosque, 2001; Koike, 2001; Alonso, 2003; García-Page Sánchez, 2005; Sánchez Rufat, 2010; Školníková, 2010) sin lograr aún una definición ampliamente aceptada. Siguiendo a Corpas (1996) y Koike (2001), usamos este término para designar aquellas combinaciones de dos palabras con propiedades específicas, como en (1a), lo que lo sitúa en un punto intermedio entre las combinaciones fijas (como frases idiomáticas y refranes), como (1b) y combinaciones libres, como en (1c).

(1)
a. Colocación:
Rendir homenaje
b. Combinación fija:
Tirar la casa por la ventana
c. Combinación libre:
Comprar libros

En esta unión de palabras, uno de los componentes de la colocación, la denominada 'base,' aporta todo o casi todo el significado del conjunto, y elige a la segunda palabra, nombrada 'colocativo', que 'selecciona en éste una acepción especial, frecuentemente de carácter abstracto o figurativo' (Alonso, 2003), como se ejemplifica en (2).

(2) **Componentes del colocativo:**

a. Rendir$_{\text{colocativo}}$ homenaje$_{\text{base}}$

b. Elevar$_{\text{colocativo}}$ la autoestima$_{\text{base}}$

Así, en las colocaciones anteriores podemos considerar al nominal ('homenaje' y 'autoestima') la base y a los verbos ('rendir' y 'elevar') como colocativos de la colocación.

De acuerdo con Koike (2001) y en la misma línea Školníková (2010), existen seis características determinantes para distinguir las colocaciones:

i) La co-ocurrencia frecuente de dos unidades léxicas: si bien este rasgo fundamental no es un rasgo solo de las colocaciones, como veremos en la siguiente sección, entre otras cosas porque la co-aparición de los lexemas puede estar determinada por el significado de ambos lexemas, independientemente de que aparezcan juntos. Además, en la mayoría de las colocaciones AB hay una preferencia léxica del lexema A para aparecer con el lexema B, pero no a la inversa. Así, en la colocación 'tener hambre', el sustantivo 'hambre' suele colocarse con el verbo tener, pero el verbo 'tener' no tiende a combinarse con el sustantivo 'hambre'.

ii) Las restricciones combinatorias de los lexemas en la norma: este es uno de los mayores diferenciadores entre una colocación y un sintagma libre. Al igual que la co-ocurrencia de dos unidades léxicas, uno de los constituyentes muestra más restricciones que el otro. Así, en la colocación 'coger la gripe', el sustantivo 'la gripe' se coloca principalmente con el verbo coger y sus sinónimos. Sin embargo, el verbo 'coger' se combina con una amplia lista de sustantivos.

iii) La composicionalidad formal: las colocaciones presentan cierta flexibilidad combinatoria, tanto morfológica como sintáctica, puesto que admiten algunas modificaciones en sus componentes, que tomaremos como criterios para identificar una colocación en la sección 2.2. Pese a esta flexibilidad, las combinaciones libres son más flexibles combinatoria, morfológica y sintácticamente que las colocaciones, y a su vez, las colocaciones son más flexibles que las locuciones.

iv) El vínculo de dos lexemas: la relación semántica de las colocaciones se establece entre los significados léxicos, no entre dos unidades léxicas. Por ello, existe la posibilidad de alternancia de la categoría gramatical de los componentes de una colocación.

v) La relación típica entre sus componentes: las colocaciones presentan una relación semántica típica de sus constituyentes. Así hay una relación típica en 'cargar una pistola', pero no en 'lavar' u 'olvidar una pistola', puesto que el sustantivo 'pistola', potencialmente, solo establece una relación típica en calidad de arma de fuego. Esta relación suele estar presente en la definición lexicográfica.

vi) La precisión semántica de la combinación: la colocación refleja un concepto

inequívoco para los hablantes nativos, por lo que desempeña una función fundamental en el acto comunicativo. Por ello, las colocaciones (especialmente las de sustantivo-verbo). aparecen en las acepciones de las entradas de los diccionarios para definir una unidad léxica simple. Así, en la primera acepción de la unidad 'asustar', aparece la colocación 'dar un susto' (dar < una persona > un susto [a otra persona]).

Estas características son fundamentales para poder discriminar entre un tipo de unidad fraseológica y otra. No obstante, la línea divisoria entre una construcción y otra es muy difusa y, por lo tanto, la identificación de una colocación a veces resulta caótica.

1.1. Principales enfoques en el estudio de las colocaciones

La conceptualización y caracterización de las colocaciones puede ser atendida básicamente desde dos enfoques: el estadístico y el semántico o fraseológico. En el primero de ellos se establece que las palabras en una colocación co-ocurren de manera más frecuente que sus respectivas frecuencias y, además, que puede predecirse la longitud del texto que separa a los miembros de la colocación (distancia colocacional), esto es, el número de palabras, tanto a la derecha como a la izquierda, que separan a la base del colocativo (Jones & Sinclair, 1974). Este enfoque ha visto su influencia en la lexicografía con la creación de dos diccionarios de colocaciones para el idioma inglés que están basados en corpus: *A Dictionary of English Collocations* (Sinclair, 1995) y *COBUILD-English Collocations on CD-ROM* (Kjellmer, 1998). Aún con estas contribuciones, este enfoque ha sido ampliamente cuestionado subrayando que los datos ofrecidos al investigador finalmente deben discriminarse atendiendo criterios en su mayoría de tipo semántico. Entre los principales argumentos usados en contra de este enfoque (Corpas, 2001), se encuentran los siguientes: (i) hay combinaciones muy frecuentes que no presentan un grado de estabilidad suficiente para ser consideradas colocaciones; (ii) hay colocaciones muy estables cuyos colocados son palabras poco frecuentes, por lo que no aparecen en un corpus dado; (iii) hay colocaciones cuyos elementos aparecen muy distanciados en el discurso, por lo que no pueden ser extraídos de forma automática; (iv) la frecuencia estadística no puede dar cuenta de la prominencia cognitiva de algunas colocaciones muy establecidas y típicas de una lengua; (v) los programas de gestión de corpus no están diseñados para detectar colocaciones en el nivel lexemático, solo en el nivel de la palabra gráfica; por último, (vi) el enfoque estadístico no dispone de instrumentos para el análisis semántico de una determinada colocación.

El segundo enfoque, el semántico o fraseológico, no considera a las colocaciones como meras asociaciones de palabras cuantificables estadísticamente, sino como unidades fraseológicas, es decir, combinaciones de palabras que presentan cierto grado de fijación combinatoria. Son de especial notoriedad los trabajos surgidos en la escuela británica, donde se distinguen y clasifican las distintas unidades multiléxicas

que existen, siguiendo dos criterios fundamentales: la transparencia semántica y la conmutabilidad (Howard, 1994). Con el primero se indica si el significado que pueden adquirir las palabras que conforman una combinación puede ser literal (o el más frecuente, que estaría representado con la primera acepción utilizada para definir la palabra en un diccionario) o secundario (que se correspondería con cualquier otra acepción). El segundo criterio alude a la posibilidad de sustituir uno de los elementos de la combinación sin observar cambio de significado en el resto, o en la naturaleza de la construcción.

Otro de los trabajos que ha profundizado en el estudio de las colocaciones bajo el enfoque fraseológico, se encuentra en la 'Teoría Significado <=> Texto', donde se establece que la elección del colocativo por la base se da por un mecanismo denominado 'Función Léxica' (FL). La palabra 'función' se presenta en su sentido matemático: existe una correspondencia 'f' que asocia una unidad léxica L, denominada argumento de la función, con un conjunto de unidades léxicas f(L), denominadas valores. En el argumento se aplica el sentido 'f', y el valor representa a un conjunto de unidades léxicas o expresiones libres que pueden expresar el sentido 'f' en lugar de L o junto a L (Mel'čuk, 2006).

Una de las ventajas de la FL, es que permite clasificar de manera sistemática todas las colocaciones existentes mediante la distribución de combinaciones según el significado general que expresan y que representa su denominador común (Moreno, 2009), como en (3):

(3)
a. Dar un paseo
b. Tomar una decisión

En (3) expresan la misma idea, que consiste en 'efectuar' o 'realizar algo', utilizando colocativos diferentes (Kolesnikova, 2011). En el primer caso, para transmitir la idea de 'efectuar un paseo' (es decir, 'pasear') el argumento se combina con el verbo 'dar'. En el segundo caso, para 'efectuar la decisión' (o lo mismo, 'decidir') se combina el argumento con el verbo 'tomar'. La idea que ambas colocaciones transmiten es considerada como una función denominada 'Oper' y su notación se representa como se muestra en (4):

(4)
a. Oper(paseo) = dar.
b. Oper(decisión) = tomar.

Este enfoque ha sido adoptado para la lengua española y ha sido utilizado, por ejemplo, en la elaboración del 'Diccionario de Colocaciones del Español', centrado en nombres de sentimientos (Alonso, 2003, 2006).

El contraste que existe entre el método que proponemos en este artículo y los enfoques antes mencionados, es el siguiente: no adoptamos un enfoque estadístico porque no trabajamos con corpus de textos, en donde se presta especial atención al procesamiento de frecuencias de palabras, sino que trabajamos con un diccionario explicativo, de forma que aprovechamos diversas ventajas que se pueden obtener de estos repositorios, como el hecho de que la información tiene una estructura homogénea (en la macroestructura encontramos una ordenación de los materiales léxicos que se definen, también llamados 'entradas', y en el plano de la microestructura, una disposición constante de los elementos informativos que acompañan a cada entrada, la manera en que se representan y el orden en que aparecen), y la existencia de una relación léxica, que podemos identificar, entre la entrada y la base de la colocación.

1.2. Tipología de las colocaciones

Atendiendo a los componentes que conforman las colocaciones, se han propuesto dos tipos de éstas: las simples, formadas por dos unidades léxicas simples, y las compuestas, formadas por una unidad léxica y otra fraseológica compleja, es decir, construida por más de una palabra (locución) (García-Page Sánchez, 2005). En la siguiente tipología (Koike, 2001), mostramos las combinaciones que pueden observarse tanto en las colocaciones simples como en las colocaciones complejas.

Tabla 1. Tipología de las colocaciones (Basados en Koike, 2001).

Tipo de colocación	Combinación categorial
Simple	Sustantivo + verbo (SV) 'rumiar' (la vaca). Sustantivo + adjetivo (SA) 'lluvia torrencial'. Sustantivo + preposición + sustantivo (SPS) 'rebanada de pan'. Verbo + adverbio (VR) 'cerrar herméticamente'. Adverbio + adjetivo (RA) 'sobradamente conocido'. Verbo + adjetivo (VA) 'salir mal parado'.
Compleja	Verbo + locución nominal: 'dar un golpe de Estado'. Locución verbal + sustantivo: 'llevar a cabo un proyecto'. Sustantivo + locución adjetival: 'dinero constante y sonante'. Verbo + locución adverbial: 'pagar a tocateja'. Adjetivo + locución adverbial: 'loco de remate'.

De acuerdo con Koike (2005: 183), la diferencia entre estos dos grupos reside en que:

> "las colocaciones complejas difieren de las colocaciones simples en su estructura formal y en la distribución cuantitativa. Mientras que en las colocaciones simples las combinaciones sustantivo-verbo son las más representativas en su número, en las complejas las formadas por un verbo y una locución adverbial constituyen el grupo más numeroso".

Es de relevancia señalar que en las colocaciones donde el colocativo es un verbo y la base un adjetivo o sustantivo, el verbo es elegido como portador de las determinaciones gramaticales de tiempo, persona, número y modo. Este hecho asegura la correcta inserción de la construcción en el nivel sintáctico-oracional, y la dota de sus propias características funcionales, pero con pérdida, quizá no total, de su valor léxico. Así, el verbo actúa siempre como auxiliar del componente nominal (Koike, 2001).

1.3. Esfuerzos en la identificación automática de colocaciones

El número de intentos por identificar colocaciones de manera automática no resulta ser muy extenso, y por obvias razones, hay una clara tendencia hacia su tratamiento estadístico. Los trabajos involucrados en esta tarea buscan identificar colocaciones según la co-ocurrencia frecuente de varias palabras a corta distancia unas de otras en un corpus textual, con una frecuencia de co-aparición superior a lo que el azar permitiría predecir (Pazos, 2005).

Uno de los primeros trabajos que puede encontrarse en la literatura es el expuesto por Berry-Roghe (1972). El objetivo de éste es obtener una lista de elementos sintagmáticos (colocativos) que co-ocurran de manera significativa con un determinado elemento léxico dentro de una distancia lineal específica. Las pruebas se ejecutaron sobre un corpus de poco más de 71.000 palabras, de ellas se extrajeron combinaciones relacionadas con el término 'habitación'.

Posteriormente, Pazos (2005) realiza 5 experimentos con mejoras sucesivas, basándose en lo planteado por el estado del arte (frecuencias de bigramas y cálculo de 'puntuación z' (*z-score*), 'puntuación t' (*t-score*) y fórmula de *Dunning*), de manera que se comprobó que los patrones y comportamientos descritos en otras lenguas se producían también en el caso del español. En estos experimentos, se observó que al incrementar el tamaño del corpus y al realizar la lematización y etiquetado gramatical de los textos, los resultados mejoraban (al reducir el texto a sus formas canónicas, se podían integrar variantes que por aparecer pocas veces, se excluían).

El corpus final de prueba se conformó por poco más de 1.000.000 palabras. En éste se consideraron las combinaciones 'SVV', 'SA', 'AS' y 'VR', estableciendo co-

apariciones con recurrencia mayor a 2. La combinación más beneficiada fue 'AS' con un 30% de combinaciones fraseológicas identificadas de las 1.060 extraídas en total, en su la mayoría colocaciones. Se concluye que el uso de frecuencias y otras medidas de asociación, aunque sí constituyen una condición, no resultan suficientes para distinguir una colocación de la que no es.

Santana, Pérez, Sánchez y Gutiérrez (2011) realizaron la extracción de colocaciones de términos económicos utilizando un corpus de textos variados con un tamaño aproximado de 300.000.000 palabras, del cual se obtuvieron poco más de 14.000.000 combinaciones con frecuencias mayores a 2. Se utilizaron las técnicas estadísticas de frecuencia relativa, 'puntuación z', 'puntuación t', información mutua y la fórmula de *Dunning*. Se encontró que para un término económico fijado, la frecuencia relativa, 'puntuación z' y la fórmula de *Dunning* lograban incrementar el catálogo con nuevas colocaciones, la mayoría relacionadas con la economía.

Una variante de estos métodos es el presentado por Gelbuk y Kolesnikova (2011), donde buscan predecir colocaciones utilizando algoritmos de aprendizaje automático supervisado, basándose en patrones semánticos que se especificaron bajo el formalismo de 'Funciones Léxicas' (FL). Se utilizó como lista de entrenamiento 1.000 pares de palabras verbo-sustantivo más frecuentes del *Spanish Web Corpus* de *Sketch Engine*. Todas las combinaciones extraídas de esta lista fueron etiquetadas con sus correspondientes sentidos de palabra de *Spanish Wordnet* y aquellas que correspondían a colocaciones fueron marcadas manualmente con sus respectivas FL. Para cada combinación de palabras, se generó una representación binaria de características, que se conformaron por los hiperónimos de las palabras de la lista y una característica categorial, la cual se marcaba como negativa ('no') en caso de que la FL asociada no fuera previamente clasificada. Después de evaluar diferentes clasificadores, no se encontró alguno que lograra detectar todas las FL. Al final, lograron alcanzar un promedio de 74% en 'medida-F', mejor que el 66% del trabajo de Wanner, Bohnet y Giereth (2006) contra el que se compararon.

Bajo esta misma línea, en otro trabajo, Kolesnikova y Gelbukh (2012) comparan la eficiencia del enfoque estadístico contra uno basado en reglas. El primero lo utilizan para entrenar métodos de aprendizaje automático y en el segundo utilizan sentencias condicionales (operador *if*) para modelar representaciones de hiperónimos. Los resultados obtenidos muestran que los métodos basados en reglas superan significativamente a los métodos estadísticos.

En general, podemos observar que los trabajos realizados hasta el momento, toman como recurso primario de procesamiento un corpus de texto para aplicarle técnicas estadísticas. Es aquí donde se destaca una diferencia importante de estos trabajos con nuestro método: enfocamos nuestra atención en el procesamiento de un diccionario explicativo, donde la información se muestra de manera más estructurada pero las estadísticas no podrían aportar mucha información.

2. Experimento realizado

2.1. Datos

En esta sección damos cuenta de cómo es organizada la información en el diccionario que procesamos y cómo ha surgido la idea de identificar colocaciones con las características que definimos al inicio.

El diccionario utilizado es el Diccionario de la Real Academia de la Lengua Española (Real Academia Española (2011) en adelante DRAE). Éste se estructura por secciones textuales denominadas 'artículos', los cuales están dispuestos ordenadamente y se conforman por dos elementos: i) 'entradas' o 'unidades léxicas' y ii) por la información que las define o describe. Las unidades léxicas se dividen en dos grandes sectores: 'palabras de contenido léxico' (ej. sustantivos, adjetivos, verbos y adverbios) y 'palabras funcionales' (ej. preposiciones, pronombres, etc.). Atendiendo estos dos sectores, se reconoce la definición lexicográfica de dos maneras: como 'definición propia' o 'perifrástica', encargada de expresar el significado de las entradas en cuanto a su contenido léxico-semántico, y la 'definición impropia' o 'funcional', utilizada para describir o explicar el funcionamiento y empleo de palabras funcionales.

La estructura de las definiciones propias suele seguir la norma establecida por la llamada 'definición aristotélica', la cual está conformada por un enunciado encabezado por un término genérico (*genus*) o hiperónimo inmediato, seguido de una diferencia específica, o conjunto de rasgos y características que diferencian el término definido de otros que se agrupan bajo el mismo hiperónimo.

El análisis de las definiciones lexicográficas de verbos en el DRAE nos permite identificar tres tipos de elementos empleados al inicio de las definiciones (los cuales destacamos con un subrayado). El primero considera el uso de verbos que de manera individual expresan un significado concreto, como en (5). El segundo, el empleo de perífrasis verbal, ejemplificado en (6). El último caso, mostrando el uso de colocaciones, señalado en (7).

(5) **Con verbos**:

a. Con un solo verbo:
Susurrar: Hablar quedo, produciendo un murmullo

b. Con dos o más verbos enlazados por conjunciones:
Mascar: Partir y triturar algo con la dentadura

c. Con dos o más verbos enlazados por disyunciones:
Abarrotar: Apretar o fortalecer con barrotes algo.

(6) **Con perífrasis verbal**:

a. Rendir: Tener que admitir algo

b. Aclarar: Volver a lavar la ropa con agua sola después de jabonada

c. Concebir: Comenzar a sentir alguna pasión o afecto

(7) **Con colocaciones**:

a. Invalidar: Hacer inválido, nulo o de ningún valor algo

b. Amenizar: Hacer ameno algo

c. Inutilizar. Hacer inútil, vano o nulo algo

Puede advertirse que si se pretende identificar el hiperónimo de las definiciones antes mostradas, resultaría imposible para los casos (6) y (7), pues ambos no constituyen definiciones de tipo aristotélico. Sin embargo, lograr que un proceso automático identifique qué definiciones inician con perífrasis verbal, como en (6), no resulta tan complicado, pues gran parte de la solución vendría dada por la previa indicación de los auxiliares tradicionalmente reconocidos (Topor, 2005), en los ejemplos en (6) se tendría: 'tener que', 'volver a' y 'comenzar a', y de qué manera se combinan con las formas verbales adicionales que le dan el significado léxico. Esta facilidad de reconocimiento automático no ocurre para los casos de las colocaciones en (7): se sabe de antemano qué disposición pueden seguir las palabras según sus categorías gramaticales dentro de las colocaciones, pero no podemos hablar de la identificación de formas verbales determinadas que se usen como colocativos.

2.2. Criterios para evaluar combinaciones de palabras

Entre las características presentadas en la definición de la colocación, basamos el análisis y la evaluación de las colocaciones en criterios y pruebas establecidos en Koike (2001) y Školníková (2010), donde se proponen las características propias del comportamiento de la colocación en (i) el plano sintáctico, (ii) en el plano semántico, (iii) en el plano sintagmático y (iv) en el plano paradigmático:

(i) Pruebas sintácticas: debido a la flexibilidad sintáctica de algunas colocaciones, éstas aceptan el intercambio de alguno de sus constituyentes o incluso algunas transformaciones estructurales como las siguientes:
   a. La modificación adjetival: 'Ganar un premio' > 'Ganó un premio importante'. [ganar un premio].
   b. La pronominalización: 'Pagar la multa' > 'La multa ha prescrito y ya no

tengo que pagarla'. [pagar la multa].

c. La nominalización: 'Derogar leyes' > 'La derogación de las leyes de la dictadura es normal en una democracia'. [derogar leyes].
d. La pasivización: 'Inhumar el cadáver' > 'El cadáver fue inhumado en el cementerio del pueblo'. [inhumar el cadáver].
e. El uso atributivo y predicativo de algunos adjetivos: 'Invierno crudo' > 'El invierno es crudo'.
f. La inclusión de cuantificadores: 'Cuchillo agudo' > 'El cuchillo es muy agudo'. [cuchillo agudo].
g. La relativización: 'Seguidor asiduo' > 'El que sigue asiduamente'.
h. Las colocaciones de tipo sustantivo + verbo permiten ser modificadas por una escala de verbos auxiliares: 'Guardar silencio' > 'Tener que guardar silencio' > 'Poder guardar silencio'.

(ii) Pruebas semánticas: la relación semántica que guardan los constituyentes de la colocación puede estar determinada por los siguientes aspectos:

a. La especialización semántica en las colocaciones: la composicionalidad del significado es fundamental en las colocaciones, ya que su significado es muy deducible de los significados de sus colocados. Sin embargo, las colocaciones presentan diversos grados de especialización semántica y, por eso, no resulta ser tan inequívocos. Generalmente, las colocaciones de tipo sustantivo + verbo y sustantivo + adjetivo, indican que el sustantivo es autónomo y el verbo o el adjetivo van a especializarse semánticamente.
b. La neutralización semántica: este es el resultado de la especialización semántica. En la base de una colocación, por ejemplo sustantivo + verbo o sustantivo + adjetivo, cuando el sustantivo se combina con más de un verbo forma colocaciones sinónimas: 'dar/ emitir/ pegar/ lanzar/ soltar un grito'. Esto es, los significados de los verbos [o adjetivos] quedan neutralizados semánticamente al funcionar como sinónimos.

(iii) En el plano sintagmático, se establece que las colocaciones presentan diferentes combinaciones. Sin embargo, si dos colocaciones presentan un elemento en común se refiere a una colocación concatenada:

a. Colocaciones concatenadas con el verbo en común[1]:
Colocación sustantivo$_{sto}$ + verbo y verbo + sustantivo$_{CD}$ = sustantivo$_{sto}$ + verbo + sustantivo$_{CD}$: 'Las abejas liban las flores / Las abejas libar y libar las flores'.
b. Colocaciones concatenadas con el sustantivo común:
    1. Dos colocaciones de verbo + sustantivo$_{CD}$: como 'Solo monto caballos ensillados' / 'Montar el caballo' y 'Ensillar el caballo'.
    2. Colocación sustantivo + verbo y sustantivo + adjetivo: como

'Ha cometido un craso error, creyendo que podía confiar en él' / 'Cometer un error' y 'Craso error'.

3. Colocaciones con sustantivo + verbo y sustantivo + preposición + sustantivo: 'Devané la madeja de lana' / 'Devanar la madeja y madeja de lana'.

De acuerdo con Školníková (2010), el tercer tipo de la colocación concatenada (verbo + sustantivo, sustantivo + preposición + sustantivo) parece ser el más corriente, gracias a las condiciones sintácticas.

(iv) En el plano paradigmático, las colocaciones se configuran en dos tipos de relaciones :

a. Colocación derivada, esta puede tener sus formas correspondientes en otras estructuras sintácticas y puede cambiar su categoría gramatical gracias al significado léxico de sus constituyentes, algunos ejemplos ilustrativos son: sustantivo + preposición + sustantivo > verbo + sustantivo ('una rebanada de pan' > 'rebanar el pan'); adverbio + adjetivo > verbo + locución adverbial ('sobradamente conocido' > 'conocer de sobra'), etc.
b. Colocación no derivada, esta no presenta esta posibilidad: 'Trabar amistad' > '*amistad tratable'.

Como vemos en esta serie de pruebas y restricciones para delimitar una colocación, existen diferencias, tanto formales como conceptuales. Sabemos que la definición e identificación de colocaciones puede sobrepasar esta serie de restricciones. Sin embargo, nos concentraremos en analizar aquellas que cumplen con estos criterios como primer paso para para identificar y discriminar las colocaciones.

2.3. Método usado

El uso de colocaciones suele extenderse a diferentes partes de la definición, como en notas explicativas, ejemplos, en el contorno (sin y con demarcación) y fórmulas introductorias restrictivas (Ruíz, 2007) (Serra, 2012); sin embargo, hemos observado que muchas definiciones de verbos inician también con una colocación, es decir, tomando la posición de *genus* o hiperónimo inmediato, considerando que las definiciones lexicográficas de los verbos se adaptan en su mayoría al tipo de definición aristotélica (Battaner & Torner, 2008). Bajo este planteamiento, nuestro trabajo se centra en identificar de manera automática combinaciones de palabras que tengan como restricción las siguientes características: (i) la distribución de las palabras en las combinaciones debe darse de la siguiente manera[2]: verbo + sustantivo (VS), verbo + preposición + sustantivo (VPS), verbo + adverbio (VR) y verbo + adjetivo (VA), y (ii) tienen que ser empleadas al inicio de definiciones de verbos.

El experimento consistió en tomar como candidatas a colocaciones las combinaciones de palabras cuya base (de la posible colocación) pertenezca a la

familia léxica del verbo definido. La evaluación del experimento se realizó de manera semiautomática, y se explicará en el siguiente capítulo.

En la siguiente tabla se muestran tres definiciones donde se observa el cumplimiento de las restricciones antes planteadas:

Tabla 2. Ejemplo de colocaciones encontradas al inicio de definiciones.

Entrada	Definición	Colocación	Combinación
Burlar	Hacer burla de alguien o algo	Hacer burla	VS
Abofetear	Dar de bofetadas	Dar de bofetadas	VPS
Agriar	Poner agrio algo	Poner agrio	VA

El proceso que seguimos para identificar y extraer las combinaciones de palabras que consideramos candidatas a ser colocaciones se divide en dos fases:

(i) La fase de preprocesamiento consistió en realizar el etiquetado gramatical de las definiciones de los verbos, utilizando la herramienta Freeling (Padró & Stanilovsky, 2012), y la identificación y agrupación de las palabras en familias léxicas, a través de la implementación de una heurística desarrollada por los autores de este trabajo, la cual agrupa las palabras en torno a raíces obtenidas por la eliminación de afijos (tomados de una lista previamente creada) de las formas lematizadas de las palabras.

(ii) La fase de procesamiento en la que se verificó el cumplimiento de los criterios en los que nos basamos es la siguiente: se seleccionaron las combinaciones de palabras cuyas categorías gramaticales correspondieran con las combinaciones de las colocaciones que nos interesaba identificar y, además, se consideró que la base (sustantivo o adjetivo) de la combinación candidata a colocación perteneciera a la familia léxica del verbo definido.

Es de interés mencionar que en varias definiciones observamos el uso del contorno de la definición entre lo que podría ser el colocativo y la base de la colocación. Por esta razón, previo a la fase de elección de combinaciones de palabras, se realizó la remoción del contorno en todas las definiciones. En (8) incluimos una definición que muestra una colocación (indicada en negrita) y el contorno introducido entre el colocativo y la base (señalado entre comilla simple):

(8) Santificar. 1. tr. **Hacer** 'a alguien' **santo** por medio de la gracia.

El contorno que removimos de las definiciones son las palabras 'algo', 'alguien', 'cosa', 'persona' y 'lugar'. La elección y eliminación de estos elementos obedece al hecho de que son comúnmente identificados en la lexicografía como contornos de las definiciones (Serra, 2009), y que al llegar a ser insertados entre el colocativo y la

base de colocaciones usadas en las definiciones, impide que nuestro método logre identificar estas combinaciones de palabras (como se observa en el ejemplo (8), donde el contorno 'a alguien' separa los elementos de la colocación). Además, los pronombres indefinidos que decidimos eliminar, frente a otros elementos utilizados también como contornos (ej. sustantivos), muestran una mayor frecuencia de uso en las definiciones del diccionario de la DRAE (Castro-Sánchez & Sidorov, 2010).

Finalmente, en algunos casos donde alguna de las palabras removidas es un constituyente de la colocación, como en ('ocultar algo'), no lo consideramos colocación, puesto que únicamente será candidato a colocación cuando el pronombre indefinido 'algo' sea sustituido por un referente, por ejemplo un nominal, como en ('ocultar la verdad'). Otros casos pueden presentarse en los verbos empleados dentro del ámbito del trabajo, por ejemplo, 'contratar a una persona', o 'despedir a una persona', construcciones que no se suelen escuchar normalmente, sino solo cuando el contorno 'persona' se sustituye por personas concretas, como 'contratar a Juan' o 'despedir a Pedro' (Travalia, 2006). Esto es, las palabras elegidas que constituyen el contorno parecen no hacer referencia ninguna combinación frecuente, sino hasta que son sustituidos por un elemento léxico que generalmente acompaña al colocativo en cuestión.

3. Evaluación de resultados

3.1. Resultados

Realizamos la identificación y extracción de un total de 1.347 combinaciones de palabras candidatas a colocaciones, de acuerdo con el cumplimiento de las restricciones que señalamos previamente en el apartado 2.2.

La frecuencia de estas combinaciones dada por la categoría gramatical de sus constituyentes[3] se muestra en la siguiente Tabla.

Tabla 3. Frecuencias de las combinaciones categoriales extraídas.

Combinación	Frecuencia	Ejemplo de candidato de colocación
VS	823	Adquirir conciencia
VA	76	Hacer ameno
VR	10	Dejar seguro
VPS	438	Dar de puñaladas

Del total de combinaciones obtenidas con este método, se extrajo una muestra aleatoria del 75% (1.010 combinaciones) que se evaluó manualmente de acuerdo con los

criterios descritos en la sección 2.2. La evaluación aplicada a estas 1.010 definiciones, requirió aproximadamente 20 horas de trabajo ejecutado por una persona.

De esta muestra, se encontró que el 61,3% (619) son colocaciones, algunas de las cuales mostramos en la Tabla 4. El resto de los candidatos a colocaciones, se distribuyen de la siguiente manera: el 2,8% (28) corresponde a locuciones, y el 35,9% (363) se identificó como combinaciones libres de palabras.

Tabla 4. Muestra de colocaciones obtenidas.

Categoría	Entrada en el diccionario	Colocación
VA	Amenizar	Hacer ameno
VA	Tersar	Poner terso
VA	Aligerar	Hacer ligero
VN	Diplomar	Conceder un diploma
VS	Sancionar	Aplicar una sanción
VS	Excepcionar	Alegar excepción
VS	Alegrar	Causar alegría
VMS	Enseñorear	Hacerse señor
VMPS	Espejar	Mirarse al espejo
VMPS	Pretextar	Valerse de un pretexto
VMPS	Amancebarse	Unirse en amancebamiento
VPS	Desgreñar	Andar a la greña
VR	Glotonear	Comer glotonamente
VR	Interinar	Desempeñar interinamente

Dado que nuestro método no analiza un corpus y, por lo tanto, no se basa en los resultados de un estudio estadístico, consideramos que compararlo con los trabajos del estado del arte, que sí realizan análisis de corpus, no reflejaría el impacto real de

nuestros resultados. En lugar de ello, optamos por establecer una línea base con la cual compararnos, quedando definida como se muestra en (9), y que se evaluó según lo descrito en la sección 2.2:

(9) Considerar como colocación toda combinación de palabras con la que inician las definiciones de los verbos.

Una vez que realizamos las pruebas planteadas en (9), se extrajeron 2.000 definiciones y se supuso que todas iniciaban con una colocación. Se evaluaron y se encontró que en solo 316 definiciones se demostraba este planteamiento. Esto resulta en un 15,8% de eficiencia, el cual está muy por debajo del 61,3% que alcanzamos con nuestro método.

Esta misma muestra de 2.000 definiciones, se utilizó para calcular la cobertura. Se encontró que en las 316 definiciones que iniciaban con una colocación, en 114 se cumplía que la base perteneciera a la familia léxica del verbo definido, lo que arrojó un porcentaje del 36%.

3.2. Evaluación

La evaluación de los potenciales candidatos a colocaciones se realizó de manera manual y consistió en aplicar a cada ejemplo de la muestra los criterios y pruebas (restricción de combinación léxica, prueba semántica, sintáctica y comportamiento en el plano sintagmático y paradigmático). De acuerdo con estos, el 61,3% de los ejemplos de la muestra corresponden a una colocación.

Sin embargo, además de estas pruebas aplicadas de forma manual, también verificamos que la colocación existiera y estuviera en uso, de manera que no fuera una construcción artificial. Para ello, se realizó una búsqueda de cada colocación en el motor de búsqueda *Google*. Con ello, nos cercioramos del uso y además obtuvimos la frecuencia de aparición de cada colocación, al menos, en un corpus electrónico. Esto demuestra que el uso real de las colocaciones está incluso almacenado en la *web*.

Ahora, en la evaluación de las colocaciones, se presentaron comportamientos que se pueden clasificar de acuerdo con (i) las características del grupo verbal al que perteneces, (ii) con la semántica de cada uno de sus componentes y (iii) de acuerdo con la frecuencia de aparición.

(i) Los verbos de régimen preposicional, como en (10), en general, pueden ser considerados buenos candidatos para ser colocaciones.

(10) Atar con juncos.

En este ejemplo, el verbo de régimen preposicional 'atar con' normalmente está acompañado por una frase preposicional. Esta complemento preposicional puede codificar diferentes roles temáticos: un paciente ('atar la caja'), un

instrumento ('atar con un listón'), un modo o manera (atar con fuerza). De manera que cada uno de estos complementos tiene características y funciones específicas. Por ejemplo, en el dominio de los instrumentos, palabras como 'listón', 'cuerda', 'mecate', etc., suelen co-aparecer frecuentemente con el verbo 'atar', puesto son dos palabras relacionadas tanto en la configuración sintáctica, como semántica.

(ii) No obstante, también encontramos verbos de régimen preposicional cuya combinación con el otro elemento léxico, no es típico ni hay especializad semántica entre los constituyentes, como en (11).

(11) dar de color.

Esto puede ocurrir por diversos factores:

a) El complemento no corresponde a la semántica del verbo. Esto es, no existe una relación semántica típica entre el verbo y el complemento.

En el ejemplo (11), el verbo 'dar' pertenece al campo de transferencia; entonces, se espera que aparezcan complementos como objetos concretos o, en sus usos metafóricos, podría recibir otros complementos más abstractos ('dar de topes'). Pero en este caso, el adjetivo 'color' no presenta una correspondencia entre un uso primario ni metafórico.

b) La relación que establece un verbo con el complemento no es tan común, por lo que al realizar la búsqueda en el *Google*, este buscador arroja un número muy reducido de co-apariciones, como en (12) comparado con los que están relacionados típicamente.

(12) Sacar con socaliña

(iii) Finalmente, en la muestra se presenta casos en los que algunos candidatos cumplen con todos los criterios y pruebas establecidas, pero no suelen tener una frecuencia de aparición alta en el buscador *Google*, como en (13). Por el contrario, hay ejemplos de construcciones que no cumplen con todos los criterios aplicados; sin embargo, suelen aparecer con mucha frecuencia en el buscador, como el ejemplo de (14).

(13) Hablar en diálogos

(14) Poner en libertad

O bien, son locuciones que no presentan flexibilidad como las colocaciones, ya que de acuerdo con Corpas (1996), la colocación muestra cierta flexibilidad combinatoria, mientras que las locuciones carecen de éstas.

En general, los comportamientos de las colocaciones identificadas aquí, presentan

problemas y dificultades de naturaleza semántica-sintáctica. Sin embargo, también impacta la frecuencia con la que se presente. De manera que la identificación de las colocaciones está muy vinculada con estos dos criterios, tanto los criterios lingüísticos como estadísticos para evaluar la naturaleza de estas combinaciones.

CONCLUSIONES Y TRABAJO FUTURO

En este trabajo presentamos un método que permite extraer colocaciones de las definiciones de algunos verbos de un diccionario explicativo. El método propone como colocaciones combinaciones de palabras situadas al inicio de las definiciones, que tienen como única restricción que la base de la colocación candidata se relacione léxicamente con el verbo definido. A pesar de que nuestro método está basado en esta idea sencilla, los resultados indican que la mayor parte de estas combinaciones de palabras son efectivamente colocaciones, puesto que tanto estadística, como sintáctica y semánticamente mostraban una vinculación estrecha que se diferenciaba de la flexibilidad de las combinaciones libres y las restricciones de las locuciones.

Cabe destacar que nuestro método no realizó el procesamiento de un corpus para evaluar los resultados con base en un estudio estadístico, por lo que consideramos que la comparación con los trabajos del estado del arte, que en cambio sí realizan análisis de corpus, no reflejaría el impacto real de nuestros resultados.

En el futuro, extenderemos este análisis a aquellas definiciones donde la palabra perteneciente a la familia léxica del verbo definido es utilizada en la diferencia específica (Abanicar. 1. tr. 'Hacer aire' con el 'abanico'. U. m. c. prnl.)

En la identificación de familias léxicas, tarea que resultó ser el elemento principal para establecer la restricción única que consideramos para discernir entre una colocación con otro tipo de construcción sintáctica, solo consideramos la obtención de la raíz a partir de la supresión de afijos encontrados en las palabras, e ignoramos procesos ortográficos que sin duda contribuirían a la recuperación de una mayor cantidad de candidatos a colocaciones. Por ejemplo, en (15)

(15) Pacificar. 1. tr. Establecer la paz donde había guerra o discordia.

Donde puede apreciarse que la supresión del sufijo en la entrada 'pacificar', sin el tratamiento ortográfico adecuado, imposibilita relacionarlo con el término 'paz' bajo una misma familia léxica.

Todas estas opciones de procesamiento nos muestran la posibilidad de extraer colocaciones de un contexto textual bien estructurado, como es el caso de la definición, con heurísticas relativamente sencillas de implementar.

REFERENCIAS BIBLIOGRÁFICAS

Alonso, M. (2003). Hacia un Diccionario de Colocaciones del español y su codificación. En M. A. Martí (Ed.), *Lexicografía computacional y semántica* (pp. 11-34). Barcelona: Edicions de La universitat de Barcelona.

Alonso, M. (2006). Glosas para las colocaciones en el Diccionario de Colocaciones del Español. En M. Alonso Ramos (Ed.), *Diccionario y Fraseología* (pp. 59-88). A Coruña: Universidade da Coruña.

Battaner, M. & Torner, S. (2008). La polisemia verbal que muestra la lexicografía. En D. Azorín (Ed.), *Actas del II Congreso Internacional de Lexicografía Hispánica* (pp. 204-216). Alicante: Universidad de Alicante.

Berry-Roghe, G. (1972). *The Computation of Collocations and their Relevance in Lexical Studies* [en linea]. Disponible en: http://www.chilton-computing.org.uk/acl/applications/cocoa/p010.htm

Bosque, I. (2001). Sobre el concepto de colocación y sus límites. *Lingüística Española Actual, 23*(1), 9-40.

Castro-Sánchez, N. & Sidorov, G. (2010). Analysis of definitions of verbs in an explanatory dictionary for automatic extraction of actants based on detection of patterns. En C. Hopfe, Y. Rezgui, E. Métais, A. Preece & H. Li (Eds.),

Lecture notes in computer science (pp. 233-239). Berlin: Springer-Verlag.

Corpas, G. (1996). *Manual de fraseología española*. Madrid: Gredos.

Corpas, G. (2001). Apuntes para el estudio de la colocación. *Lingüística Española Actual, 23*(1), 41-57.

García-Page Sánchez, M. (2005). Colocaciones simples y complejas: Diferencias estructurales. En R. Amela & G. Wotjak (Eds.), *Fraseología contrastiva: Con ejemplos tomados del alemán, español, francés e italiano* (pp. 145-168). Murcia: Universidad de Murcia.

Gelbuk, A. & Kolesnikova, O. (2011). Supervised learning for semantic classification of Spanish collocations. En J. Martínez-Trinidad, J. Carrasco-Ochoa & J. Clitter (Eds.), *Lecture Notes in Computer Science* (pp. 362-371). Berlin: Springer-Verlag.

Howard, P. (1994). *A computer-assisted study of collocations in academic prose, with special reference to grammatical structure and stylistic value*. Tesis doctoral, Universidad de Leeds, Leeds, West Yorkshire, Inglaterra.

Jones, S. & Sinclair, J. (1974). English lexical collocations. A study in computational linguistics. *Cahiers de Lexicology, 24*(1), 15-61.

Kjellmer, G. (1998). A dictionary of English collocations, based on the Brown Corpus. *International Journal of Corpus Linguistics*, *3*(2), 338-348.

Koike, K. M. (2001). *Colocaciones léxicas en el español actual: Estudio formal y léxico semántico*. Universidad de Alcalá de Henares, España.

Koike, K. M. (2005). Colocaciones complejas en el español actual. En R. Pérez, E. Trives & G. Wotjak (Eds.), *Fraseología contrastiva: Con ejemplos tomados del alemán, español, francés e italiano* (pp. 169-184). Murcia: Univerdidad de Murcia.

Kolesnikova, O. (2011). *Automatic extraction of lexical functions*. Tesis doctoral, Instituto Politécnico Nacional, Ciudad de México, México.

Kolesnikova, O. & Gelbukh, A. (2012). Semantic relations between collocations: A Spanish case study. *Revista Signos. Estudios de Lingüística*, *45*(78), 44-59.

Mel'čuk, I. (2006). Colocaciones en el diccionario. En M. Alonso Ramos (Ed.), *Diccionarios y Fraseología* (pp. 11-43). La Coruña: Universidad de Coruña.

Moreno, M. (2009). *Recopilación, desarrollo pedagógico y evaluación de un banco de colocaciones frecuentes de la lengua inglesa a través de la lingüística de corpus y computacional*. Tesis doctoral, Universidad de Granada, España.

Pazos, J. (2005). *Detección automatizada de fraseologismos*. Tesis doctoral, Universidad de

Granada, España.

Padró, L. & Stanilovsky, E. (2012). FreeLing 3.0: Towards Wider Multilinguality. En N. Calzolari, K. Choukri, T. Declerck, M. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk & S. Piperidis (Eds.), *Proceedings of the Language Resources and Evaluation Conference LREC 2012* (pp. 2473-2479). Istanbul: ELRA.

Real Academia Española. (2011). *Diccionario de la lengua española* [en linea]. Disponible en: http://www.rae.es/rae.html

Ruíz, A. (2007). La noción de colocación en las partes introductorias de algunos diccionarios monolingües del español. *Revista de Lexicografía*, 13, 139-182.

Sánchez Rufat, A. (2010). Apuntes sobre las combinaciones léxicas y el concepto de colocación. *Anuario de Estudios Filológicos*, 33, 291-306.

Santana, O., Pérez, J., Sánchez, I. & Gutiérrez, V. (2011). Extracción automática de colocaciones terminológicas en un corpus extenso de lengua general. *Procesamiento del Lenguaje Natural,* 47, 145-152.

Serra, S. (2009). Las restricciones de selección en los diccionarios generales de la lengua española. *Boletín de Filología*, *44*(2), 187-213.

Serra, S. (2012). *Gramática y diccionario. Contornos, solidaridades léxicas y colocaciones en lexicografía española contemporánea*. Tesis doctoral, Universidad Complutense de Madrid, España.

Sinclair, J. (1995). *Collins Cobuild English Collocations on CD-ROM*. Londres: HarperCollins.

Školníková, P. (2010). *Las colocaciones léxicas en el español actual*. Tesis doctoral, Universidad de Masaryk de Brno, República Checa.

Topor, M. (2005). Criterios identificadores de las perífrasis verbales del español. *Sintagma: Revista de lingüística*, 17, 51-69.

Travalia, C. (2006). Las colocaciones implícitas. *Estudios de Lingüística Universidad de Alicante*, 20, 317-334.

Wanner, L., Bohnet, B. & Giereth, M. (2006). What is beyond Collocations? Insights from Machine. *12th EURALEX International Congress,* 1071-1084.

NOTAS

1 Las abreviaturas son Sto= Sujeto y CD=Complemento Directo, (Koike, 2001: 152).

2 Las etiquetas utilizadas para denotar la categoría gramatical de los constituyentes se basa en las etiquetas propuestas por el grupo *Eagles* para la anotación morfosintáctica de lexicones y corpus

para todas las lenguas europeas.

3 Ignoramos el uso de clíticos y determinantes para representar las etiquetas de los constituyentes de las colocaciones. De esta manera, por ejemplo, las combinaciones del tipo VDN se tratan como VN.