PROGRAMA DE DOCTORADO

SIMULACIÓN NUMÉRICA EN CIENCIAS Y TECNOLOGÍA

TESIS DOCTORAL

EXTRACCIÓN AUTOMÁTICA

DE

NEXOS LÉXICOS

AUTORA: ISABEL SÁNCHEZ BERRIEL

LAS PALMAS DE GRAN CANARIA, MAYO DE 2015

PROGRAMA DE DOCTORADO

SIMULACIÓN NUMÉRICA EN CIENCIAS Y TECNOLOGÍA

DEPARTAMENTO DE MATEMÁTICAS

TESIS DOCTORAL

EXTRACCIÓN AUTOMÁTICA

DE

NEXOS LÉXICOS

AUTORA: ISABEL SÁNCHEZ BERRIEL

DIRECTOR:

DR. OCTAVIO SANTANA SUÁREZ

CO-DIRECTOR:

DR. JOSÉ RAFAEL PÉREZ AGUIAR

LAS PALMAS DE GRAN CANARIA, MAYO DE 2015

Agradecimientos

Este trabajo de tesis supone el punto de arribo gracias a los esfuerzos realizados por un numeroso grupo de personas que han persistido en apoyar la trayectoria seguida desde que emprendí mis primeros pasos con las colocaciones hasta que comencé a "vislumbrar la luz al final del túnel". A todos ellos quiero agradecer su aportación, cada uno en su especialidad y vocación universitaria.

En primer lugar quiero referirme, al Dr. Octavio Santana Suárez, director de esta tesis, quien no sólo abrió la puerta de su ciencia e interés por el tema, sino que la ha mantenido de par en par sin ambages, alimentando la obstinación, el tesón y la ilusión de las que me he armado para recorrer este camino; quisiera agradecerle el esfuerzo, todo el tiempo dedicado, el conocimiento compartido y los recursos facilitados sin los que no hubiese podido llevar a buen puerto este proyecto. Al Dr. José Pérez Aguiar le agradezco sinceramente todo el apoyo como co-director, y cada una de sus orientaciones que han contribuido a la consecución de este trabajo.

Reservo un lugar para mis compañeros de la Universidad de La Laguna: Luzma, Jesús y Macu, que alguna que otra vez les ha tocado sustituirme en mis estancias en Las Palmas. Inestimable ha sido la ayuda de José Luis Roda, que siempre que lo he necesitado ha aportado la tecnología que tantas palabras parecían devorar. A Félix García, mi maestro, mil gracias por además ayudar en todo lo que estaba en su mano para que yo pudiera desempeñar esta tarea.

Un recuerdo especial para mis amigos Bea y Robert, que con sus risas y todo lo demás siempre han estado ahí.

Por último a todas esas personas que, como en el cuento del zapatero y los duendes, silenciosamente han propiciado cada uno de los pasos dados. Mis padres, mis hijas, Cristo, Vicky, Rafa, Sayo, mi familia, todos ellos mis duendes.

Resumen

En el presente trabajo se presentan análisis y soluciones al problema de la extracción automática de colocaciones léxicas. Con independencia de la lengua de que se trate, a partir de la propuesta de Sinclair –estudia las frecuencias de aparición en corpus textuales–, las técnicas resuelven el problema desde el punto de vista estadístico: una combinación de palabras se considera colocación si su aparición conjunta queda motivada en lugar de provenir del puro azar. Aunque este enfoque sea demasiado simplista para el lingüista, aporta la herramienta básica para resolver el problema desde la perspectiva computacional. Bajo la premisa de implementar una solución que requiera el menor número posible de recursos léxicos, se conjuga el conocimiento lingüístico con un planteamiento estadístico. Las conclusiones y aportaciones que se derivan dan respuesta a la extracción de colocaciones de un corpus textual sea cual sea su volumen.

Se maneja un corpus textual de 11 000 textos literarios y no literarios con un número aproximado de 300 000 000 palabras de la lengua española. A partir de esta amplia muestra del español, se procede a la extracción de colocaciones mediante técnicas basadas en las frecuencias de palabras; se analizan los resultados que se obtienen y se contrastan con las combinaciones del Diccionario Combinatorio del Español Contemporáneo, Redes (DCECR). Se dan soluciones a la inestabilidad que producen las marcadas diferencias entre las frecuencias de uso de las distintas palabras en el corpus, que se ve acentuada por el uso de un corpus con tan amplio volumen. Por una parte, se diseña una metodología de análisis centrada en la palabra, y por otra se proponen nuevos indicadores, basados en las técnicas de detección de *outliers*, que tratan de captar las características de uso preferentes de las uniones presentes en este tipo de combinaciones. Las propuestas se orientan de modo que sean válidas en corpus menos extensos: se utiliza como corpus reducido las novelas de D. Benito Pérez Galdós.

Se resuelve también el problema de extracción automática de nexos semánticos motivados por la selección de argumentos por parte de los predicados; se enfoca el estudio hacia la determinación de clases semánticas entre los elementos que combinan con una determinada palabra, se lleva a cabo un repaso de las técnicas que permiten extraer tales grupos de forma automática. En este caso, se opta por adaptar el indicador *asociación de selección* (Resnik, 1997), que mide el vínculo entre un grupo de argumentos y su predicado, utilizando para evaluarlos diccionarios ideológicos del español como clasificaciones de palabras presentes en el corpus. Las pruebas se realizaron con el Diccionario Ideológico Vox (DIV); en concreto, el trabajo se centra en los casos: *verbo + grupos de sustantivos*, *adjetivo + grupos de sustantivos* y *adverbio + grupos de verbos*.

Por último, se aporta una aplicación software que permite consultar las propiedades combinatorias de una palabra de la lengua española a través de Internet, así como la extracción automática de las colocaciones en textos en español de forma amigable para el usuario.

Índice

1. Introducción

1.1. Introducción

El término colocación es utilizado en lingüística para designar un tipo de combinaciones de palabras usadas con relativa frecuencia en una determinada lengua que, sin llegar a ser combinaciones libres, tampoco constituyen locuciones o expresiones idiomáticas. Ejemplos de colocaciones en español:

tener apetito
afrontar riesgos
dar (una) voltereta
poner (un) telegrama
sufrir (un) desengaño
vomitar insultos
accidente aparatoso
agenda apretada
catarro impresionante
charla animada
comida suculenta
enemigo acérrimo
noticia terrible
precio razonable
registro civil
vacaciones escolares
izar (la) bandera
enhebrar (la) aguja
arribar (un) barco
cardar (la) lana
cariarse (los) dientes
ensillar (un) caballo
banco de peces
acceso de tos
rebanada de pan
comer frugalmente
asesinar brutalmente
competir duramente
disgustarse seriamente

En este capítulo se presentan distintos elementos que definen y caracterizan las colocaciones léxicas en español. El objetivo del mismo radica en la recopilación de información

sobre el tema como punto de partida para la construcción de un reconocedor automático de colocaciones. El estudio se estructura de la forma que se expone a continuación. En los dos primeros puntos se hace una revisión de las diversas corrientes sobre el tema: el enfoque estadístico, y la interpretación léxico-semántica del fenómeno, respectivamente. En el tercer punto se tratan las características de las colocaciones en español extraídas de los trabajos recopilados. A continuación se recogen las diversas clasificaciones de colocaciones propuestas en la bibliografía, tanto desde el punto de vista de las categorías gramaticales implicadas como desde los aspectos semánticos. Una vez establecida la taxonomía colocacional, en el siguiente punto se exponen las características específicas de cada categoría. Por último, se aborda la cuestión de la explotación de los diccionarios de uso –Diccionario del Español Actual (DEA)[1], el Diccionario de Uso del Español (DUE)[2] , Diccionario General de la Lengua Española de Vox y el Nuevo Diccionario de Voces de Uso Actual– como fuente de colocaciones y cómo las presentan los autores en los artículos con objeto de obtener material de partida que permitan detectar vínculos o relaciones semánticas entre los elementos que forman una colocación o un grupo de ellas, como el caso:

"dar se utiliza con sustantivos de golpe"

1.2. Enfoque estadístico.

Los primeros estudios sobre las colocaciones según los autores consultados se deben a Firth (1957), quien introduce el término **collocation** para referirse a la coaparición de palabras. Sin embargo, Gloria Corpas apunta trabajos previos de autores como Saussure, Bally o Porzig que ya tratan cuestiones estrechamente relacionadas con este fenómeno (Corpas, 1996). A partir de los trabajos de Firth surge la llamada escuela sistémica británica a la que pertenecen Halliday, Sinclair, Jones. Para tal corriente, las colocaciones se reducen a combinaciones frecuentes de palabras. Por utilizar el criterio de la frecuencia como rasgo principal de las colocaciones, en sus trabajos se propone el uso de técnicas estadísticas para la exploración de textos en busca de combinaciones de palabras cuya frecuencia de aparición pueda considerarse superior a la de otras combinaciones. Los cálculos se realizan a partir de las llamadas **líneas de concordancia**. Una línea de concordancia es un determinado fragmento en el que aparece una palabra dada en un texto o corpus textual. Por ejemplo, si se considera el texto:

> "**vio** los sagaces ojos que bajo el ala de ancho sombrero de terciopelo viejo resplandecían; **vio** la mano morena y acerada que empuñaba una vara verde, y el ancho pie que, al moverse, hacía sonajear el hierro de la espuela..."(Doña Perfecta, Benito Pérez Galdós).

---

[1] *Diccionario del Español Actual*, M. Seco y otros, Madrid, Agular, 1999. Dos volúmenes.
[2] *Diccionario de Uso del Español*, [M. Moliner], 2 vols., Gredos, Madrid, 1966.

Las líneas de concordancia o contextos de la palabra **vio** son:

vio los sagaces ojos que bajo

sombrero de terciopelo viejo resplandecían; **vio** la mano morena y acerada

Cada línea de concordancia corresponde con una aparición de la palabra en el texto –permite el conteo para cálculos de frecuencias. Para extraerlas es necesario fijar la palabra bajo estudio y cuántas palabras antes y después de la misma conformarán el contexto –en el ejemplo se fijó en 5. En general, se considera una cantidad fija, *n*, que recibe el nombre de **distancia colocacional** (**collocation span**). Sinclair y Jones estudiaron qué valor se le debe asignar y propusieron para *n* el valor de 4 (Koike, 2 001). Sin embargo, en este mismo trabajo Koike presenta un ejemplo en el que se muestra que la distancia colocacional no tiene por qué ser 4:

> Ambos juristas son expertos en la materia y han viajado a Madrid a fin de que la ***petición*** que en su momento pueda ***cursar*** el magistrado español sea acorde a la legislación británica (EPI, 27/10 al 2/1198:2).

En este caso la colocación *cursar petición* está separada por una cantidad de palabras mayor que 4. Otros autores han estudiado esta cuestión, Rogghe propone la distancia de 4 –salvo que sean adjetivos que será de 2–, Haskel, +/-3, Al Madi +/-1, pero que puede llegar a 5 ó 6 según los casos, Smadja propone 5 y Miall considera una distancia media de 9, aunque en ocasiones es necesario llegar hasta 15. (Corpas, 1 996).

Las técnicas descritas en las líneas anteriores forman parte de la disciplina conocida como Lingüística del Corpus. Además de la distancia colocacional, en esta disciplina también son objeto de estudio las medidas de lo estrechamente relacionadas que estén los dos elementos de la colocación, como por ejemplo: **frecuencia de aparición de *A* con *B***, **z-score** e **información mutua**. Aunque no son los únicos indicadores, sí son los más comunes: miden la frecuencia de aparición conjunta de palabras, pero no son capaces de reflejar aspectos más complejos presentes en las colocaciones. Existen trabajos que echan mano de técnicas estadísticas más avanzadas como el análisis factorial para intentar captar tales intimidades, pero requieren que se apliquen a una gran cantidad de textos para garantizar la fiabilidad de los resultados.

El ejemplo de la Tabla 1-1 (Koike, 2 001), recoge los datos sobre la coocurrencia de *medida* y *drástica* en un corpus e intenta averiguar si *medida drástica* es una colocación**.** Indica que *drástica* aparece con *medida*, pero *medida* con *drástica* no.

medida drástica	
	Frecuencia
medida	250
drástica	8
coocurrencia	4
drástica con medida	$\frac{4}{8} \times 100 = 50\%$
medida con drástica	$\frac{4}{250} \times 100 = 1'6\%$

Tabla 1-1. Ejemplo del cálculo de la frecuencia de coocurrencia.

A lo que Koike llama **frecuencia de aparición** de *A* con *B*, no tiene por qué presentar el mismo valor que la frecuencia de aparición de *B* con *A;* en realidad, es un indicador del porcentaje de veces que una palabra aparece con otra respecto al número total de veces que aparece. Si este valor es elevado, quiere decir que si aparece *A* es muy posible que también aparezca *B*. Para Koike es suficiente indicio de colocación que el valor sea superior al 20%.

Las medidas del **z-score** e **información mutua** constituyen valores más complejos y precisos utilizados para la extracción de colocaciones en Lingüística de Corpus. Este tema será tratado en profundidad en el capítulo 2, donde se analizarán estos y otros indicadores que permiten detectar colocaciones de forma automática en textos electrónicos.

1.3. Enfoque léxico-semántico.

Trabajos posteriores en la materia –Benson, Haussman, Mitchell, Coseriu, Cruse, Haensch o Mel'čuk– tratan las colocaciones como un fenómeno más complejo que la mera combinación usual de palabras. Apuestan porque las colocaciones ponen de manifiesto relaciones sintagmáticas o semánticas entre palabras.

Mitchell considera que tales vínculos se producen entre lexemas (Koike, 2001). Benson distingue entre 1) **colocaciones gramaticales** que se componen de una palabra dominante seguida de una preposición o estructura sintáctica –*consistir en*– y 2) **colocaciones léxicas** que se forman exclusivamente con nombres, adjetivos, verbos y adverbios –entre sus aportaciones a la teoría colocacional figura la primera clasificación establecida en la materia (Carballo, 1998), (Alonso, 1997) y (Koike, 2001).

Coseriu define las **solidaridades léxicas** que coinciden parcialmente con el concepto de colocación. Una solidaridad léxica es una relación orientada en la que existe un **lexema determinante** y un **lexema determinado** –en el sentido de que el determinado implica al determinante–, pero no a la inversa. Por ejemplo, *caballo alazán* es una solidaridad léxica, porque *alazán* implica a *caballo*, pero no el recíproco. Una **solidaridad léxica unilateral** es del

tipo *morder dientes, lamer lengua* –**?no suelen coocurrir. Sin embargo, *caballo alazán, ladrar perro...* que pertenecen al conjunto de las solidaridades léxicas multilaterales sí son colocaciones. Por otra parte, *radicalmente opuesto* es una colocación pero no es una solidaridad, puesto que no hay ninguna implicación en los lexemas que la forman, por ello no se puede afirmar que las colocaciones coincidan plenamente con las solidaridades léxicas, sino que algunas colocaciones coinciden con las solidaridades léxicas multilaterales (Corpas 1996, Koike 2001).

De especial importancia es la aportación de Haussman: llevó a cabo una taxonomía de las colocaciones y construyó una teoría sobre el estatus semántico de los elementos que intervienen en las colocaciones –definió **base** y **colocativo**. Para referirse a los dos lexemas que coocurren en una colocación, los lingüistas que tratan del tema en la actualidad usan el par de términos. Cada uno juega un papel distinto en la colocación, la base selecciona una acepción especial en el colocativo, que no puede ser descodificado sin la presencia de la base. Este rol diferenciado de los elementos que conforman la colocación se ilustra en la Tabla 1-2

Colocación	base	colocativo	Acepción del colocativo seleccionada
levantar veda	*veda*	*levantar*	**19. tr**. Hacer que cesen ciertas penas, prohibiciones o vejámenes impuestos por autoridad competente. *Levantar el entredicho, el destierro, el arresto, el embargo*
café fuerte	*café*	*fuerte*	**9.** adj. Dicho de un color o de un sabor: **intenso**.
lógica aplastante	*lógica*	*aplastante*	**2.** adj. Abrumador, terminante, definitivo.

Tabla 1-2 Semántica de los elementos de la colocación

Mel'čuk desarrolla sus trabajos desde la perspectiva de las **funciones léxicas** definidas para representar las posibilidades combinatorias de un determinado lexema que expresan un cierto sentido. Al igual que una función matemática, una función léxica asocia un conjunto de valores al elemento al que se aplica o **elemento llave** y que se corresponde con la base de Hausmann –el valor que toma la función es el colocativo o un conjunto de ellos. El vínculo establecido entre los elementos que forman una colocación determina la función que se debe escoger; se permiten utilizar funciones compuestas como combinación de varias funciones léxicas para expresar relaciones más complejas. Algunas funciones léxicas con los sentidos que expresan se recogen en la Tabla 1-3.

Función	Sentido	Ejemplo
Son	Emitir sonido típico	*Son(gato) = maullar*
Magn	Muy, Intenso	*Magn(enemigo) = acérrimo*
Mult	Conjunto de	*Mult(oveja) = rebaño*
$Fact_i$	Realizarse, Llevarse a cabo	$Fact_0$*(sospecha)=confirmarse, corroborarse*
Bon	Bueno	*Bon(clima) = benigno*
Ver	Tal como debe ser	*Ver(cuchillo) = afilado*
Sign	Porción de	*Sign(pan) = tostada*
$Func_i(C)$	Verbo semánticamente vacío que toma C como sujeto gramatical e i como su primer complemento	$Func_i$*(viento) = soplar*
$Oper_i(C)$	Verbo semánticamente vacío que toma C como sujeto gramatical e i como su primer complemento	$Oper_i$*(atención) = prestar*
Incep	Comenzar	$IncepFunc_0$*(epidemia)=declararse*

Tabla 1-3. Ejemplos de funciones léxicas.

Las funciones léxicas constituyen una herramienta útil a la hora de representar las relaciones comunes que conectan distintos lexemas presentes en las colocaciones; sin embargo, entre las críticas que reciben se mencionan aquí las siguientes:

1. Pueden producir combinaciones que no lo son, como en el caso de: *Sing(biblioteca) = libro.*
2. Existen colocaciones que no pueden ser descritas con las funciones léxicas, como en el caso de *discusión bizantina* (Koike, 2001).

1.4. Las colocaciones en español.

Los estudios sobre colocaciones en español se producen a partir de la década de los 90, si bien a finales de los 70 Seco introduce el concepto de **contorno de la definición lexicográfica** que hace referencia al término colocación (Corpas, 1996), (Penadés, 2001) y (Castillo, 1998).

Corpas las considera un tipo de **unidades fraseológicas**[3].

Bosque las supone una manifestación de **selección léxica**[4], en las que los colocativos que actúan como predicados seleccionan sus argumentos, no de forma individual, sino grupal

[3] Una unidad fraseológica es una construcción formada por al menos dos palabras ortográficas —puede ser una oración compuesta— con las siguientes características potenciales: uso habitual, presenta cierta fijación, idiomaticidad —su significado no es composicional—, presenta variantes. Todos estos aspectos no aparecen en la misma medida en los distintos tipos de unidad fraseológica: colocaciones, locuciones, y enunciados fraseológicos (paremias y fórmulas rutinarias), sino que se dan en distinto grado (Corpas, 1996).

[4] Las colocaciones constituyen una manifestación de selección léxica. Los predicados —verbales, adjetivales, adverbiales o preposicionales— seleccionan sus argumentos, y al hacerlo restringen el

–del mismo, o de varios campos semánticos– que conforman lo que se denomina **clase léxica**. De esta manera, lo realmente importante radica en detectar la clase que selecciona un determinado colocativo; entre los ejemplos que presenta en su trabajo aparecen:

supino *{ignorancia, incompetencia, inutilidad, necedad, desconocimiento, estupidez, ridiculez, imbecilidad, irresponsabilidad, egoísmo, cinismo}.*

terminantemente *{prohibir, negar, desmentir, excluir, rechazar, oponerse}*

En algunos casos establece la relación entre el colocativo y algún rasgo definitorio de la clase, como en el caso:

*"**universalmente** + verbos que denotan aceptación"*

Zuluaga se refiere a las colocaciones como un tipo de combinación de al menos dos lexemas, en las que puede aparecer algún elemento gramatical y de las que no se puede decir que sean combinaciones libres, ni tampoco unidades fraseológicas, pero que comparten propiedades de unas y otras.

Por último citamos a Alonso que estudia las colocaciones en el marco de la teoría Sentido-Texto de Mel'čuk.

Si bien se aprecian distintas interpretaciones del fenómeno colocacional entre los lingüistas españoles, aparece como un denominador común de sus trabajos el que descartan la interpretación exclusivamente estadística de las colocaciones por presentar los siguientes problemas:

1. No todas las combinaciones frecuentes cuentan como colocaciones.
2. Puede que no se considere colocación aquella que forman unas palabras de baja frecuencia.
3. El uso exclusivo de la frecuencia no puede manifestar fenómenos semánticos presentes en las mismas.

conjunto de entidades que pueden denotar en función de unos rasgos semánticos – a veces muy abiertos y en otras considerablemente restringidos. Frente a lo que parece sugerir una buena parte de la bibliografía, en muchísimos casos las unidades seleccionadas por los predicados no son piezas léxicas aisladas, sino clases léxicas de mayor o menor intensión. (Bosque, 2001).

1.5. Características.

En los siguientes epígrafes se ahonda en los rasgos específicos que resultan fundamentales a la hora de discriminar si una combinación de lexemas constituye o no una colocación. En general, existe una serie de características propias de las colocaciones, algunas de ellas diferenciadoras de otros tipos de combinaciones y otras que comparten al menos en parte debido a la transición gradual **combinación libre→colocación→locución** (Zuluaga, 2002), (Blasco, 2002), (Penadés, 2001) y (Corpas, 1996) –he aquí una de las causas por las que es difícil decidir si una determinada construcción es o no una colocación. Cabe destacar las discrepancias entre autores a la hora de catalogar como colocaciones determinadas construcciones; por ejemplo, pueden ser controvertidos los siguientes casos (Bosque, 2001):

ejecutar un castigo
empantanarse un proyecto
enfilar la calle principal
disminución progresiva...

Penadés advierte que el Diccionario fraseológico del español moderno (Varela, Kubarth, 1994) existen numerosas colocaciones incluidas como locuciones; entre otras:

dar corte
dar el cambiazo
dar el espectáculo
dar el golpe.

1.5.1. Flexibilidad

En contraposición con las estructuras por lo general fijas de las locuciones (Koike, 2001), la primera propiedad que se trata en este apartado se refiere a la **flexibilidad formal** de las colocaciones. Admiten[5]:

1. Cambios de categoría gramatical:

   comida frugal / comer frugalmente

2. Modificación adjetival: En las colocaciones es lícito incluir un adjetivo que modifique al sustantivo, la colocación:

   hacer un aterrizaje

   también se puede usar:

   hacer un aterrizaje forzoso

[5] Todos los ejemplos se obtuvieron de (Koike, 2001).

sin embargo, esto no es válido en las locuciones, por ejemplo, la locución:

quemarse las pestañas

no admite ser utilizada en la forma:

quemarse las pestañas largas

3. Transformación a pasiva. Podemos transformar a pasiva una colocación sustantivo + verbo:

trasplantar órganos

en pasiva:

el órgano fue transplantado

lo que no es posible en el caso de las locuciones:

escurrir el bulto

no se utiliza :

el bulto fue escurrido

4. Relativización:

seguir la línea / Este libro marca la línea que deben seguir sus partidarios.

echar el ojo (a algo) / El ojo que acabo de echar a ese vestido.

5. Nominalización: Es posible en las colocaciones como se observa en el ejemplo:

trasplantar un órgano / el trasplante de órganos

pero no en las locuciones:

escurrir el bulto

no se usa:

el bulto que fue escurrido

6. Pronominalización:

Asumió el **cargo** de alcalde, pero su repentina enfermedad le impidió **desempeñarlo** (desempeñar cargo)

De aquí que se afirme que la combinación realmente se da entre los lexemas[6] y no entre las lexías simples. Tal aseveración explica las colocaciones que se forman con una unidad léxica simple y una locución, como: *dar un golpe de Estado*. Por tanto, la exploración de un texto para la extracción de colocaciones no puede llevarse a cabo sobre palabras gráficas.

Por otra parte, los elementos que intervienen pueden sustituirse en general por otros relacionados, bien porque:

1. Pertenecen al mismo campo semántico:

desempeñar *{un cargo, una función o un papel}* (Corpas, 1996)

2. Son sinónimos:

{mantener, sostener} una ***conversación***

armar *{alboroto, algarabía, barullo}*

3. Son hiperónimos, hipónimos:

cariarse *{un diente, la dentadura}*

escamar *{pescado, besugo}*

En virtud de esta característica se pueden describir propiedades combinatorias mediante definiciones del tipo: "*dar combina con sustantivos de golpe*", o bien "*dar selecciona sustantivo de golpe*" que expresamos en forma de reglas:

"*dar + sustantivos de golpe*"

"*universalmente + verbos que denotan aceptación*"

1.5.2. Determinación.

Los elementos que forman una colocación reciben el nombre de colocados, uno de ellos goza de autonomía semántica y recibe el nombre de **base**, el significado del otro depende precisamente de ésta, está determinado por ella y se denomina **colocativo**. Como se mencionó con anterioridad, esta nomenclatura fue introducida por Hausmann en la terminología colocacional y en la actualidad su uso se ha generalizado entre todos los estudiosos del tema consultados. Respecto a las características señaladas de los dos colocados, se dice que la base es una palabra **autosemántica** –no necesita del colocativo para ser definida– y el colocativo es **sinsemántica**

[6] **.lexema**. 1. m. Ling. Unidad mínima con significado léxico que no presenta morfemas gramaticales, por ejemplo, sol; o que, poseyéndolos, prescinde de ellos por un proceso de segmentación; por ejemplo, terr en enterráis. (DRAE).

–su definición necesita ser completada por la base. Esta característica es fundamental en las siguientes definiciones de colocación que aportan Alonso (1994) y García Platero (2002):

> Una **colocación** o "semi-frasema" (FL) AB es una combinación de dos lexemas *A* y *B*, de tal forma que su significante es la suma regular de los significantes de los lexemas constituyentes y su significado incluye el significado del lexema *A* y un significado *C* que es:
>
> a) Bien ***C*** **<>** ***B*** y
>
> i. *C* es vacío: el lexema *B* es un auxiliar que se usa para sostener una configuración sintáctica
>
> *dar un paseo...*
>
> ii. *C* no es vacío, pero el lexema *B* expresa *C* solo en combinación con *A* o con otros pocos lexemas similares
>
> *odio mortal*
>
> *interés vivo...*
>
> b) Bien ***C*** **=** ***B*** y
>
> El lexema *B* se selecciona de modo restringido: en combinación con *A* no se puede reemplazar por otro sinónimo
>
> *café fuerte*...
>
> c) *C* incluye el sentido de *A*
>
> *pelo rubio*
>
> *nariz aguileña...*
>
> Todos los casos posibles se pueden explicar mediante alguna de las posibilidades:
>
> 1. La base selecciona una acepción especial en el colocativo; generalmente se usa en sentido figurado y se reconoce su significado gracias a la base con la que aparece. Los tres primeros casos en la definición de Alonso se ajustan a este tipo. Es el caso de colocaciones como:
>
> *banco de peces*
>
> *precio astronómico*
>
> *levantar sesión*
>
> *locamente enamorado*
>
> *dar envidia...*
>
> Se han subrayado los colocativos. Se incluyen las definiciones que proporciona el DRAE para la acepción en que se utilizan tales términos:

banco. 5. m. Conjunto de peces que van juntos en gran número.

astronómico. **2.** adj. coloq. Que se considera desmesuradamente grande. Sumas, distancias astronómicas.

sesión. levantar la ~.

1. Concluirla.

locamente fr. 2. adv. m. Excesivamente, sin prudencia ni moderación

dar. 23. tr. Causar, ocasionar, mover. *Dar gusto, gana*.

Se advierte en estas definiciones que los significados de *banco, astronómico*, *levantar*, *locamente* y *dar* están determinados por cada uno de los otros colocados; si no estuvieran presentes, no se sabría con qué acepción se utiliza el colocativo respectivo –tal comportamiento refleja la precisión semántica que sustenta la colocación.

2. No se puede definir el colocativo sin la base, pero en este caso lo que existe es una relación típica entre los dos elementos colocados. Los colocativos son términos que especifican propiedades, acciones o procesos específicos de las bases; por eso se dice que existe un vínculo típico entre ambos. Valen como ejemplos:

triángulo isósceles
querochar abeja
pelo lacio
zarpar barco
trinchar carne...

Así, *isósceles* es un tipo de triángulo y no existe ambigüedad respecto a este término. En tales ocasiones, por estar la base contenida en el colocativo, se consideran las colocaciones como casos de *solidaridades léxicas.*

isósceles. V. triángulo isósceles

querochar. 1. intr. Dicho de las abejas y de otros insectos: Poner la querocha.

Uno de los significados de *lacio* en el DRAE manifiesta que es una propiedad específica del *pelo.*

lacio, cia. 3. Dicho del cabello: Que cae sin formar ondas ni rizos.

Además de existir una relación típica entre los colocados, se observa que el significado de *lacio* queda determinado por la base.

El caso de *zarpar barco* es similar al anterior, puesto que la segunda acepción de este verbo requiere para su definición la presencia de la base de la colocación, y ésta a su vez selecciona una de las acepciones posibles, pero no en un sentido figurado.

> **zarpar**. **2**. intr. Dicho de un barco o de un conjunto de ellos: Salir del lugar en que estaban fondeados o atracados.

Lo mismo sucede con el último ejemplo:

> **trinchar**. **1.** Partir en trozos la comida para servirla.

Si los dos elementos que configuran una determinada combinación son vocablos autosemánticos estamos ante una combinación libre, en cambio, si una unidad es autosemántica y la otra sinsemántica se está ante una colocación. Los elementos que forman una locución son palabras extrasemánticas, porque el significado de la misma no proviene de los significados de sus constituyentes. (García Platero, 2002).

1.5.3. Frecuencia, Preferencia.

Las colocaciones son combinaciones usuales de palabras que los hablantes nativos de una lengua repiten tan habitualmente en situaciones comunicativas similares que las han convertido en estereotipos. Aunque teóricamente se podría escoger como colocación cualquier combinación posible, en muchas ocasiones ¿no son usuales o de uso preferente; así que no todas las posibilidades de elección −más o menos amplia− de los colocativos deberían considerarse como colocaciones, sino las que han sido sancionadas por la comunidad.

Así, *fervorosamente* y *fervientemente* se combinan de forma preferente, o casi exclusivamente con *rezar* (u *orar*), aunque los diccionarios suelen dar como sinónimos *vivamente, ardientemente,* o *vehementemente.* De este modo, *fervientemente* y *fervorosamente* se especializan en su empleo con *rezar*, en detrimento de otros adverbios de manera (García-Page, 2001).

En *lanzarse al ataque*, *lanzarse* ha sido preferido a *arrojarse,* y *ataque* ha sido preferido a *lucha*. (Zuluaga, 2002). En *guerra mundial*, *mundial* ha sido preferido a *internacional* (Zuluaga, 2002).

Tal perspectiva supone un problema a la hora de buscar generalizaciones sobre las propiedades combinatorias de algún lexema dado.

A pesar de que la elevada frecuencia de uso de una combinación respecto a otras posibles no sea un criterio determinante para asegurar que forme una colocación, constituye una

característica que se debe considerar. Por otra parte, es una consecuencia, en parte, del uso de clichés o combinaciones preferentes.

En los lenguajes de especialidad es normal el uso de combinaciones típicas de términos que derivan en colocaciones; por ejemplo

en informática:

disco duro
cargar el fichero
eliminar la carpeta
implementar un programa...

o en derecho:

{interponer, cursar, presentar} un recurso
{abrir, instruir, tramitar} un proceso...

En tales casos, los vínculos léxicos son más fácilmente identificables, porque el problema aparece delimitado; más difícil resultaría detectarlos en la lengua general.

1.5.4. Composicionalidad.

En cuanto al significado de una colocación se da un abanico de posibilidades que abarca desde la **composicionalidad** a la **seudocomposicionalidad** como se explica a continuación:

Composicionalidad total: sucede tanto en *triángulo isósceles* como en *querochar abeja*. El significado se obtiene de la suma de los significados de los dos colocados. La colocación se puede descodificar fácilmente y se dice que es totalmente transparente semánticamente. Nótese que tal característica aparece en las mismas colocaciones que se incluyen como ejemplos en el punto 2 del apartado Determinación.

Seudocomposicionalidad o composicionalidad parcial. En las colocaciones en las que la base selecciona un uso figurado del colocativo se necesita un conocimiento específico por parte del hablante para poder asignarles el significado; por tanto se dice que presentan cierta *opacidad semántica* (García Platero, 2002) y no son tan fácilmente descodificables.

Esta absoluta transparencia que atribuye al significado de las partes, no siempre se da con la misma intensidad; por ejemplo en *dinero negro*, si al hablante no nativo se le da ya hecha la combinación no podrá conocer exactamente su significado, si previamente no se le ha informado de que la base *dinero* en presencia de *negro* asigna al colocativo una acepción distinta de la que posee (‘ilegal’) (Castillo Carballo, 1998).

La seudocomposicionalidad se aprecia en todas las colocaciones incluidas en el punto 1 del apartado Determinación.

Las combinaciones libres son totalmente composicionales:

limpiar la guitarra
dar un caramelo...

Por el contrario, la locuciones no se pueden descodificar sin conocer el significado de toda la construcción, puesto que el significado de los elementos constituyentes no contribuye al significado global:

quemar las pestañas
cortar el bacalao
arrojar la toalla...

1.5.5. Gradualidad.

Hay que resaltar que las características expuestas de las colocaciones no son exclusivas de tal fenómeno léxico, sino que cada una va apareciendo (determinación, fijación) o desapareciendo (flexibilidad, composicionalidad, transparencia semántica) gradualmente desde las combinaciones libres a las locuciones; las colocaciones representan el grupo intermedio, en el que se manifiesta la transición progresiva de las primeras a las segundas como se recoge en el esquema de la Figura 1-1

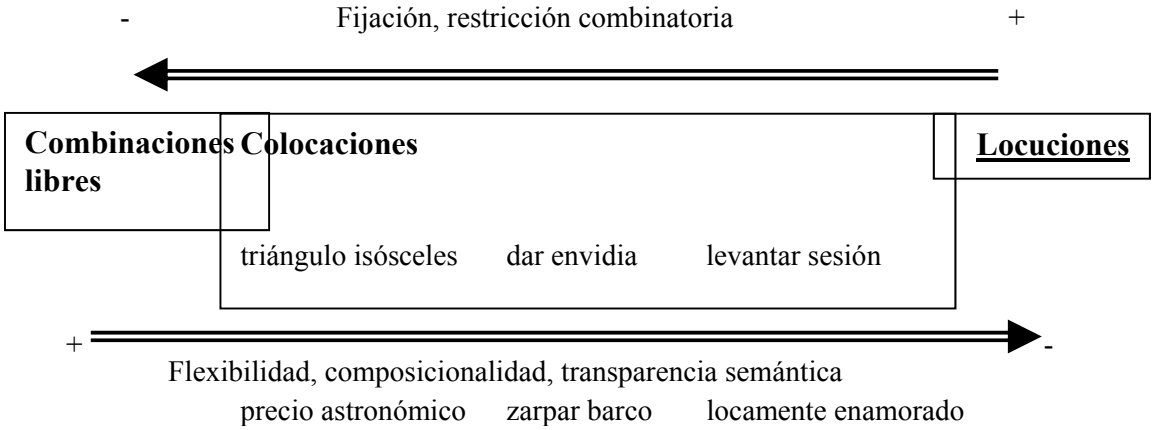


Figura 1-1 Transición progresiva entre combinaciones libres, colocaciones y locuciones.

Asimismo, la transición dentro del conjunto de las colocaciones es gradual y continua, es decir, en unas se dan algunas características con mayor intensidad que en otras. Por ejemplo, la flexibilidad en

triángulo isósceles

es menor que en

dar hambre

puesto que estos vínculos se podrían representar como:

"triángulo + tipo de triángulo" {isósceles, escaleno, rectángulo, etc.}

mientras que la relación en la segunda colocación corresponde a la estructura:

"dar + sustantivo de estado físico o de ánimo" {hambre, sueño, celos, etc.}

Debido a tal gradualidad, resulta difícil delimitar dónde terminan las combinaciones libres y dónde empiezan las colocaciones, o dónde terminan éstas y empiezan las locuciones; por tanto, no es extraño que se catalogue alguna combinación en un grupo erróneo, lo cual no quiere decir que las características de cada grupo no aparezcan bien claras y precisas (Penadés, 2001), (Bosque, 2001) y (Zuluaga, 2002).

1.6. Clasificación.

En cuanto a las variedades existentes se pueden clasificar las colocaciones según distintos criterios: la categoría gramatical de los colocados, el grado de restricción entre los colocados y el significado de la colocación.

1.6.1. Categoría gramatical de los colocados.

Varias clasificaciones se pueden citar en este apartado: la de Hausmann y la de Benson, o las clasificaciones de colocaciones en español propuestas por Corpas, Castillo Carballo, Penadés y K. Koike, con diferencias mínimas.

La primera taxonomía propuesta se debe a Hausmann, quien la estableció basándose en la función gramatical de los elementos exclusivamente, se presenta en la Tabla 1-4. La tipología de Benson de la Tabla 1-5 se define sobre las **colocaciones léxicas** que se forman con: *nombres, adjetivos, verbos y adverbios* a diferencia de las **colocaciones gramaticales** que siguen el esquema:

(nombre, adjetivo, adverbio) + preposición + estructura sintáctica

Se puede apreciar que para delimitar alguno de los grupos se apoya también en características semánticas de alguno o ambos colocados.

Tipo	Base
sustantivo + adjetivo	sustantivo
sustantivo + verbo	sustantivo
verbo + sustantivo $_{objeto}$	sustantivo
verbo + adverbio	verbo
adjetivo + adverbio	adjetivo
sustantivo + preposición + sustantivo	sustantivo

Tabla 1-4. Clasificación de Hausmann.

	Tipo	Base
Gramaticales		
Léxicas	verbo (‘creación’ o ‘activación’) + nombre / pronombre	sustantivo
	verbo (‘erradicación’, ‘anulación’) + nombre	sustantivo
	adjetivo + nombre	sustantivo
	nombre + de + nombre (uno representa a la unidad y otro el conjunto).	sustantivo
	adverbio + adjetivo	adjetivo
	verbo + adverbio	verbo

Tabla 1-5. Clasificación de Benson.

Las siguientes propuestas que se recogen se refieren al español. Nótese que en las tablas correspondientes se resaltan en negrita las diferencias; por ejemplo, en la clasificación de Corpas de la Tabla 1-6 aparece la clase:

$$verbo + sustantivo_{objeto}$$

que Carballo en la Tabla 1-7 la muestra más detallada:

$$verbo + sustantivo_{objeto}$$

$$verbo + preposición + sustantivo$$

también se marca la entrada:

$$adjetivo + adverbio$$

puesto que las dos autoras difieren en el orden

adjetivo + *adverbio* para Corpas

adverbio + *adjetivo* para Carballo

En lo esencial, ambas se pueden considerar la misma clasificación (Koike, 2001) –se aprecia que están basadas en las de Hausmann y Benson.

Tipo	**Base**	**Ejemplos**
sustantivo $_{sujeto}$ + verbo	sustantivo	*correr un rumor* *estallar una guerra* *zarpar un barco*
verbo + sustantivo$_{objeto}$	sustantivo	*desempeñar un cargo* *zanjar un desacuerdo* *conciliar el sueño* *dar comienzo*
adjetivo + sustantivo	sustantivo	*fuente fidedigna* *enemigo acérrimo* *éxito fulgurante* *oído fino*
sustantivo + sustantivo	sustantivo	*hombre clave*
sustantivo+preposición+sustantivo	sustantivo	*tableta de chocolate* *enjambre de abejas* *ronda de negociaciones*
verbo + adverbio	verbo	*caer pesadamente* *negar rotundamente* *fracasar estrepitosamente*
adjetivo + adverbio	adjetivo	*firmemente convencido* *estrechamente ligado* *rematadamente loco* *diametralmente opuesto*

Tabla 1-6. Clasificación de Corpas.

Tipo	Base	Ejemplos
sustantivo $_{sujeto}$ + verbo	sustantivo	*correr un rumor* *estallar una guerra* *zarpar un barco*
verbo + sustantivo $_{objeto}$	sustantivo	*desempeñar un cargo* *dar comienzo* *tener repercusión*
verbo + preposición + sustantivo	sustantivo	*poner en cuestión*
adjetivo + sustantivo sustantivo + sustantivo	sustantivo	*fuente fidedigna* *enemigo acérrimo* *éxito fulgurante* *oído fino* *hombre clave*
sustantivo$_1$+preposición+sustantivo$_2$	sustantivo	*tableta de chocolate* *enjambre de abejas* *ronda de negociaciones*
verbo + adverbio	verbo	*caer pesadamente* *negar rotundamente* *fracasar estrepitosamente*
adverbio + adjetivo	adjetivo	*firmemente convencido* *estrechamente ligado* *rematadamente loco* *diametralmente opuesto*

Tabla 1-7. Clasificación de Castillo Carballo.

En la Tabla 1-8, Penadés revisa las colocaciones en las que interviene un verbo; al igual que Koike de la Tabla 1-9 incluye las de:

verbo + adjetivo

	Subtipo	**Base**	**Ejemplos**
verbo + SN	verbo + nombre	sustantivo	*correr un rumor* *estallar una guerra* *zarpar un barco*
	verbo + determinante + nombre		*desempeñar un cargo* *dar comienzo*
verbo + SP	verbo+preposición + determinante+ nombre	sustantivo	*poner en cuestión*
verbo + adverbio		verbo	*caer pesadamente* *negar rotundamente* *fracasar estrepitosamente*
verbo + adjetivo		adjetivo	*resultar ileso* *salir malparado*
Estructura comparativa			*dormir como un lirón* *dar vueltas como un trompo*

Tabla 1-8. Clasificación de Penadés.

y propone considerar las que reproducen una estructura comparativa, que para Koike es una colocación compleja del tipo:

verbo + locución adverbial

La propuesta de Koike recoge las anteriores pero detalla algunos subgrupos; también abarca las colocaciones complejas no contempladas en otras tipologías, al menos específicamente. Como novedad, en esta clasificación se considera un único tipo

sustantivo + verbo

pero que se subdivide en:

$$sustantivo_{sujeto} + verbo,\ verbo + sustantivo_{CD}$$

y en:

verbo + preposición + sustantivo.

	Tipo	Subtipo	Base	Ejemplos
Simples	sustantivo+verbo	$sustantivo_{sujeto}$+ verbo	sustantivo	*correr un rumor* *estallar una guerra* *zarpar un barco*
		verbo + $sustantivo_{CD}$		*desempeñar un cargo* *dar comienzo*
		verbo+preposición+ sustantivo		*poner en cuestión*
	adjetivo+ sustantivo		sustantivo	*fuente fidedigna* *enemigo acérrimo* *oído fino*
	sustantivo+de+ sustantivo		sustantivo	*tableta de chocolate* *enjambre de abejas*
	verbo + adverbio		verbo	*caer pesadamente* *negar rotundamente*
	adverbio+ adjetivo / participio		adjetivo	*firmemente convencido* *rematadamente loco*
	verbo + adjetivo		adjetivo	*resultar ileso* *salir malparado*
Complejas	verbo+locución nominal			*dar un golpe de estado* *levantar castillos en el aire*
	locución verbal+SN			*poner en juego un recurso* *pasar a limpio un borrador*
	sustantivo+ locución adjetival			*dinero contante y sonante*
	verbo+locución adverbial			*reírse a mandíbula batiente* *llorar a moco tendido*
	adjetivo+locución adverbial			*bruto como un arado* *sano como una manzana*

Tabla 1-9. Clasificación de Koike

1.6.2. Grado de restricción de los colocados.

Según el grado de restricción Cowie diferencia entre colocaciones libres, restringidas, estables y categoría puente (Corpas, 1998). El grado de restricción va aumentando gradualmente en el orden citado. El significado en las **colocaciones libres** es totalmente composicional y pueden aparecer en numerosas colocaciones. En las **colocaciones restringidas** se está ante una acepción especial del colocativo que selecciona la base. El campo colocacional en las **estables** es mínimo, se limita a uno o dos colocados. Por último, el uso figurado del colocativo y el que quepa la sustitución sinonímica señalan las características de la **categoría puente**[7] (Tabla 1-10).

Libres	*provocar una pelea / guerra / discusión* *empezar una pelea / guerra / discusión* *ganar una pelea / guerra / discusión*
Restringidas	*correr peligro / suerte / aventuras*
Estables	*conciliar el sueño*
Categoría Puente	*levantar una calumnia / una mentira*[8]

Tabla 1-10. Clasificación de Cowie según el grado de restricción.

En la misma línea, Koike establece dos categorías: **colocaciones amplias** y **colocaciones estrechas** como aparecen en la Tabla 1-11. En las primeras, el colocativo tiene la capacidad de combinarse con numerosas bases debido a que su significado es general y polisémico en la mayoría de las ocasiones. En las segundas, el vínculo entre los dos lexemas es fuerte, las bases obligatoriamente tendrán que ser compatibles con el significado específico del colocativo[9]. Las colocaciones amplias abarcan las categorías libres y restringidas de Cowie y las estrechas recogen la estable y la categoría puente.

Colocaciones estrechas	Colocaciones amplias
escrutar votos	*recontar {votos, reses, presos}*
extirpar quiste	*quitar {quiste, muela}*
asestar puñetazo	*dar {puñetazo, consejo, paseo}*

Tabla 1-11. Colocaciones amplias y estrechas de Koike.

Una distinción similar propone García Platero al hablar de colocaciones libres y restringidas. En las primeras existe una gran cantidad de elementos que pueden combinarse con la base, mientras que en las restringidas como:

---

[7] Ejemplos recogidos del Manual de Fraseología
[8] Levantar una calumnia es una colocación, al contrario que levantar una mentira que no lo es.
[9] Ejemplos recogidos de (Koike, 2001).

fruncir el ceño

el colocativo es difícilmente sustituible.

1.6.3. Semántica.

El significado de la colocación también lleva a Koike a proponer dos categorías diferenciadas dentro de los tipos:

sustantivo + verbo

y

sustantivo + adjetivo

que son: las **funcionales** y las **léxicas**.

Colocaciones funcionales: el colocativo pierde su significado y pasa a tener un valor funcional. En el caso de las *sustantivo + verbo* de la Tabla 1-12, el verbo habilita al sustantivo para funcionar como un verbo complejo, aporta la información de número, persona y tiempo, mientras que el sustantivo aporta el significado. En general, existe un verbo simple relacionado morfológicamente con el sustantivo, aunque no siempre.

Funcional verbo+ sustantivo	Verbo relacionado
tener miedo	*temer*
hacer una aclaración	*aclarar*
dar un golpe	*golpear*
albergar esperanza	*esperar*

Tabla 1-12. Colocaciones funcionales: verbo + sustantivo.

En las *sustantivo+adjetivo*, el adjetivo intensifica cuantitativa o cualitativamente al sustantivo como se muestra en la Tabla 1-13.

Funcionales sustantivo+adjetivo	
error garrafal	cuantitativa
módico precio	cuantitativa
dolor atroz	cuantitativa
comida atroz	cualitativa
tiempo horroroso	cualitativa
chico fenomenal	cualitativa

Tabla 1-13. Colocaciones funcionales: adjetivo + sustantivo.

Colocaciones léxicas: Los elementos que conforman la colocación conservan su significado léxico plenamente, de tal forma que en tal caso se está ante colocaciones totalmente composicionales, como aparecen en la Tabla 1-14.

Verbo + sustantivo	Sustantivo + adjetivo
moler café	*pelo lacio*
cruzarse de brazos	*poder adquisitivo*
zarpar barco	*barrio céntrico*
amasar fortuna	*cuchillo afilado*
afrontar riesgo	*gobierno autoritario*

Tabla 1-14. Colocaciones léxicas.

En las colocaciones *sustantivo + verbo* hay que incluir las **aspectuales**; se llaman así porque su significado indica alguno de los siguientes aspectos: *incoativo* (comienzo), *durativo-reiterativo* (duración), o *terminativo-resultativo* (final). A diferencia de las funcionales, el verbo tiene en tal caso una cierta carga semántica; por ejemplo, en

asaltar duda

asaltar indica acometer de pronto algo, en este caso una *duda.*

Aspecto	Colocación
Incoativo	*asaltar duda*
durativo-reiterativo	*sostener conversación*
terminativo-resultativo	*cerrar sesión*

Tabla 1-15. Colocaciones aspectuales.

Sobre la base de estos grupos, y recurriendo a la definición de Alonso se consideran los siguientes casos:

Colocativo con valor plenamente funcional. Se está ante un verbo deslexicalizado[10] y un sustantivo generalmente deverbal, o bien ante un adjetivo que juega un papel meramente intensificador en alguno de los siguientes sentidos de: muy grande, magnífico, extraordinario, etc. Se puede considerar dentro de este grupo el caso ‘a). i’ de la definición de Alonso.

dar un golpe | *golpear*

DRAE: golpear. Dar un golpe o golpes repetidos.

[10] Un verbo deslexicalizado contribuye mínimamente al significado de la colocación, su función principal es aportar el tiempo y número a la construcción que conjuntamente funciona como un verbo cuyo significado proviene del significado del sustantivo.

tener miedo | *temer*

DRAE: temer. Tener a una persona o cosa por objeto de temor.

tener esperanzas | *esperar*

DRAE. esperar. Tener esperanza de conseguir lo que se desea.

hacer una aclaración | *aclarar*

DRAE. aclarar. 10. Hacer clara, perceptible, manifiesta o inteligible alguna cosa, ponerla en claro, explicarla.

soberana paliza

DEA. soberano –na. 2. Superior o extraordinario. Frec con intención ponderativa.

dolor atroz

DEA. atroz. 2. Muy grande o extraordinario.

inteligencia colosal

DEA. colosal. 2. Magnífico o extraordinario.

Vínculo típico. El colocativo expresa alguna propiedad, acción, o proceso característico de la base. Es el caso que aparece en el punto 'c)' de la definición de Alonso.

zarpar barco

DEA. zarpar. 1. Salir [un barco] del lugar en que estaba fondeado o atracado.

libar abeja

DEA. libar. 1. Chupar [un insecto el néctar, u otra sustancia azucarada, de las flores].

enhebrar aguja

DEA. enehebrar. 1. Pasar la hebra por el ojo [de la aguja (*cd*)].

dilapidar fortuna

DEA. dilapidar Derrochar o despilfarrar.

iglesia católica

DEA. católico. 1. b) De los católicos o de la Iglesia católica.

pelo lacio

DEA. lacio. 3. [Cabello] que cae liso y blando.

Selección de acepción en el colocativo. Abarca los casos 'a).ii' y 'b)' de la definición de Alonso.

albergar esperanza

DEA. albergar. 4. Tener de manera estable [un sentimiento].

administrar anestesia

DEA. administrar. 4. Dar o aplicar [algo, esp. medicamentos].

asaltar la duda

DEA. asaltar. 3. *(col)* Acometer repentinamente [a alguien una idea o un sentimiento].

perder el tren

DEA. perder. 7. No llegar a tiempo de poder utilizar [un medio de transporte (*cd)*].

tocar el piano

DEA. tocar. 6. b) Hacer sonar [un instrumento musical o un dispositivo de señales acústicas].

amasar una fortuna

DEA. amasar. 3. Acumular o reunir [bienes o dinero].

calar melón

DEA. calar. 2. Cortar un pedazo [de un melón o de otra fruta (*cd*)] para probarlos.

color crudo

DEA crudo. 8. [Color].blanco o amarillento.

alimentación equilibrada

DEA. equilibrada. 2. Que tiene equilibrio [4, 5 y 6].

equilibrio. 5. Justa proporción de los elementos [de algo].

cuestión vital

DEA. vital. 2. b) De suma importancia.

1.7. Características por categorías.

En los apartados siguientes se exponen las características propias de los distintos tipos de colocaciones que se recopilaron de la bibliografía analizada. Los epígrafes se corresponden con cada una de las categorías según la función gramatical de los colocados, si bien en cada uno se alude a las categorías establecidas según la semántica, en caso de ser necesario.

1.7.1. Sustantivo + Verbo.

La información sobre las colocaciones

sustantivo + verbo

es más numerosa que las restantes, debido a que son las más abundantes y también las más estudiadas del español.

Los verbos que forman parte de colocaciones se denominan colocacionales, los que no, se llaman no colocacionales; figuran en la Tabla 1-16 (Koike, 2001). Los colocacionales[11], a su vez se dividen en:

Funcionales: No mantienen su significado léxico. Por ejemplo, *hacer* o *sacar* en:

hacer una aclaración

sacar la conclusión...

Léxicos: Mantienen su significado léxico. Por ejemplo

afinar el piano...

Según con qué se combine un mismo verbo puede tener valor funcional o léxico. Por otra parte, tanto unos como otros pueden ser generales o específicos.

Generales: Los verbos generales tienen un número elevado de acepciones y forman numerosas colocaciones. El campo colocacional es amplio.

Específicos: Forman colocaciones estrechas que reflejan relaciones típicas entre el verbo y el sustantivo.

Colocacionales	**Funcionales**	Generales	*dar, tener, hacer, sentir, poner, echar*
		Específicos	[12] abrigar, albergar, fraguar, brindar, tejer, tramar, etc.
	Léxicos	Generales	*tocar, gastar*
		Específicos	*tañer, dilapidar*
No colocacionales	*querer, pensar, indicar, afirmar, preguntar, verbos de decir, asegurar, advertir, vender, comprar, agradecer, obligar, invitar, gustar, parecer, interesar, considerar, creer, hallar, poder, saber, querer, intentar, etc.*		

Tabla 1-16. Clasificación de los verbos.

En cuanto a los sustantivos, cabe la división entre concretos y abstractos de la Tabla 1-17 (Koike, 2001), ya que los primeros son los que intervienen en las colocaciones léxicas y los segundos los que seleccionan los sentidos figurados de los verbos.

[11] Koike, 2001

[12] Koike incluye aquí blandir, pero si es blandir espada, arma, etc. es léxico.

Concretos	instrumentos musicales	*guitarra, piano, cuerdas*	**Colocaciones léxicas**
	utensilios	*plancha, mordaza, tornillo*	
	medios de transporte	*tren, barco, guagua*	
	partes del cuerpo humano	*corazón, pelo, barba*	
	prendas de vestir	*sombrero, gafas, descosido*	
	alimentos	*café, pan, plato*	
	animales	*balidos, paloma, ganado*	
	otros		
Abstractos	de acción	*discusión, vigilancia, pago*	**Colocaciones funcionales.** Selección de acepción especial en el colocativo.
	de situación o estado de cosas	*crisis, norma, reputación*	
	de cualidad	*paciencia, habilidad, sensibilidad*	
	de evento	*boda, aniversario, congreso*	
	de sentimiento o sensación	*cariño, abatimiento, desesperación*	
	otros	*rumor, proyecto, indicio*	

Tabla 1-17. Clasificación de los sustantivos.

1.7.2. Verbo deslexicalizado + sustantivo$_{CD}$

Compuesta por verbos funcionales generales, donde el sustantivo generalmente es deverbal:

"*dar + sustantivo de golpe*":

dar golpe
dar bofetón
dar empujón
dar puñetazo
dar tropezón...

"*dar + sustantivo terminado en –azo excepto los aumentativos y algunas palabras primitivas como cazo* o *cedazo*"*:*

dar bocinazo
dar cambiazo
dar frenazo...

"*dar + sustantivo de voz o sonido*":

dar berrido...

dar:

dar albergue
dar alcance
dar aprobación
dar asentimiento
dar asilo
dar cobijo...

dar un:

dar un abrazo
dar un aviso
dar un barrido...
darse un(a):
darse un atracón (de)
darse un batacazo
darse una ducha...

darse:

darse autobombo
darse maquillaje
darse postín...

"*tener + sustantivo referido a cualidad; generalmente terminado en el sufijo –"dad"*:

tener piedad
tener curiosidad
tener habilidad
tener conocimiento
tener costumbre...

"*tener + sustantivo referido a sentimiento*":

tener amor
tener cariño
tener angustia
tener pasión...

"*tener + sustantivo referido a estado físico*":

tener jaqueca
tener sueño
tener tos
tener fatiga
tener apetito...

tener:

tener un altercado (con)
tener la ambición (de)
tener anhelo (de)
tener un apuro
tener atraso
tener confianza...

tener + su:

tener su inicio
tener su reflejo
tener su origen
tener su causa
tener su apogeo
tener su auge
tener su esplendor...

"*hacer + sustantivo abstracto terminado en -ción, -sión*; *que son derivados devebarles*":

hacer aclaración
hacer alusión
hacer reclamación
hacer adquisición...

hacer:

hacer acopio (de)
hacer alarde
hacer una afrenta
hacer una advertencia
hacer aspavientos
hacer boicot
hacer fortuna
hacer un gesto (de)
hacer pis
hacer un viaje...

"*sentir + sustantivo de estado anímico o físico*":

sentir alegría
sentir calor
sentir fatiga

sentir inquietud

sentir júbilo...

"*sentir + sustantivo de sentimiento*":

sentir cariño

sentir celos

sentir odio

sentir pudor...

"*echar + sustantivo que indican acto de decir*":

echar bronca

echar perorata

echar reprimenda...

"*echar +sustantivo relacionado con la vista*":

echar vistazo

echar mirada

echar ojeada...

"recibir, sufrir, llenarse, padecer, tener, obtener, encontrar, experimentar, pasar, seguir, cosechar, ganarse y algunos otros" + sustantivo abstracto forman colocaciones con valor pasivo:

recibir un achuchón

sufrir un atentado

tener una conmoción

padecer un desengaño

seguir un tratamiento...

"*causar, producir, provocar, etc. + sustantivo de estado emotivo"; forman colocaciones con valor causativo*:

acarrear un disgusto

infundir aliento

imponer condiciones

proporcionar alegría...

1.7.3. Verbo deslexicalizado + (sustantivo$_{CD}$) + preposición + sustantivo.

Verbo intransitivo + sintagma preposicional forman colocaciones con valor pasivo:

caer en el olvido

incurrir en desprecio

estar fuera de uso...

*llenar, sumir, llevar, reducir, etc***.** forman colocaciones con valor causativo. abismar (a alguien) en el dolor, dejar (a alguien) en la ruina, incurrir en el enojo.

1.7.4. Selección de una acepción especial en el verbo.

Verbo funcional específico + sustantivo$_{CD}$:

abrigar esperanzas

albergar propósitos

fraguar idea

brindar ayuda

tejer una trama

tramar un engaño...

En las definiciones se aprecian clases léxicas para las bases.

Verbo aspectual[13] *+ sustantivo*:

Sustantivo$_{SUJETO}$ de estado de ánimo + verbo incoativo:

asaltar duda

invadir melancolía

entrar fiebre...

Verbo de inicio de algo + sustantivo$_{CD}$:

abrir una crisis

cobrar cariño

emprender una marcha...

{caer, empeñarse, entrar, estallar, montar, ponerse, prorrumpir, romper} + en + sustantivo:

caer en el pecado

empeñarse en una discusión

estallar en llanto

romper en sollozos...

Las colocaciones de aspecto terminativo-resultativo se presentan en los tres esquemas sintácticos.

pasarse el enfado

[13] Los verbos aspectuales son aquellos que indican alguno de los aspectos que se recogen en la siguiente tabla; generalmente, cuando se usan en sentido figurado.

Aspecto	**Valor**	**Ejemplo**
incoativo	*comienzo*	*estallar una guerra*
terminativo-resultativo	*final*	*zanjar una cuestión*
durativo-reiterativo	*duración*	*encontrarse en un dilema*
intensificativo	*intensificación*	*corroer la envidia*

(Koike, 2001).

desvanecerse el dolor
abandonar una actitud
levantar la sesión
zanjar un asunto
dejarse de bromas
quedar en la impunidad
distraer (a alguien) de preocupación...

La mayoría de las colocaciones aspectuales durativas-reiterativas siguen el esquema:

verbo + preposición + sustantivo

y el verbo suele ser *andar* o *estar*.

andar con chismes
andar de caza
estar de broma
estar en circulación
seguir en coma...

Las de aspecto intensificativo también siguen principalmente el esquema:

verbo + preposición + sustantivo

También las hay en forma negativa. El sustantivo suele ser de sentimiento o de estado de ánimo.

anegarse en llantos
arder en deseos
consumirse de envidia
rebosar de alegría
angustiar temor
clavar la mirada (en)
no ahorrar esfuerzos
no escatimar medios...

1.7.5. Colocaciones léxicas.

Estas colocaciones están formadas por un verbo léxico y un sustantivo concreto entre los que hay una relación típica, como por ejemplo:

tocar la guitarra
afinar el piano
perder el tren

moler café

balar la oveja

apuntillar un toro

tejer una alfombra...

El significado es totalmente composicional. Son muy pocos los casos en que aparecen sustantivos abstractos, aunque también existen:

amasar fortuna

zanjar discusión...

son algunas de las presentadas en los ejemplos por Koike(2001); sin embargo, los esquemas

verbo léxico general + sustantivo concreto

y

verbo + sustantivo abstracto

se encuadran en las de selección de una acepción especial.

1.7.6. Sustantivo$_{sujeto}$ + verbo.

Los verbos suelen ser intransitivos o pronominales. Las construcciones pronominales impersonales se incluyen dentro de este tipo de colocaciones:

declararse una epidemia

desatarse una polémica...

Entre las colocaciones aspectuales existen numerosos ejemplos de éstos. Los verbos que intervienen en este tipo de colocaciones:

Denotan una acción característica de la persona o cosa designada por el sustantivo*:*

palpitar corazón

ensortijarse el pelo

zarpar el barco...

Indican fenómenos meteorológicos*:*

despuntar el día

caer la lluvia

soplar el viento

encapotarse el cielo...

Se refieren a sonidos emitidos por animales*:*

aullar el lobo

cacarear el gallo

relinchar el caballo...

Todos los casos se corresponden con colocaciones léxicas.

1.7.7. Sutantivo$_1$ + de + Sustantivo$_2$.

En estas colocaciones, el sustantivo$_2$ es la base y el sustantivo$_1$ es el colocativo que expresan alguno de los significados siguientes:

El sustantivo$_1$ indica grupo o colección del sustantivo$_2$ si no puede expresar otros significados:

banco de peces
ovillo de hilo
nube de polvo
bandada de aves...

El sustantivo$_1$ significa porción del sustantivo$_2$ en combinación con éste:

rebanada de pan
jarra de cerveza
copo de nieve
tableta de chocolate...

El sustantivo$_1$ indica impulso violento o repentino, manifestación brusca del sentimiento especificado mediante el sustantivo$_2$*:*

arrebato de ira
acceso de histeria
ataque de tos...

El sustantivo$_1$ en los dos primeros casos es cuantificador, y puede ser*:* **acotador** (*pedazo, trozo, tajada, etc.),* de **medida** (*kilo, litro, galón, etc), o de* **grupo** (*grupo, serie, manada, etc).*

1.7.8. Sustantivo + Adjetivo.

Los adjetivos también pueden ser colocacionales o no colocacionales, según intervengan o no en la formación de colocaciones. Los que pertenecen a la primera categoría son solo los calificativos, pero de éstos se pueden excluir los situacionales Tabla 1-18 puesto que los únicos que presentan restricciones combinatorias con sustantivos son los relacionales y los cualitativos (Koike, 2001).

Relacionales		*Nacionalidad, religión, instituciones, clase social, étc.*	*francés, republicano, carnívoro, musical, etc.*
Cualitativos	**Cualidades físicas**	*Tamaño, altura, anchura, volumen, color, temperatura, peso, velocidad, tiempo, gusto, tacto, vista, olfato, oído, sonido, etc.*	*alto, bajo, largo, corto, ancho, estrecho, gordo, delgado, rojo, blanco, caliente, frío, pesado, ligero, rápido, veloz, joven, viejo, efímero, insípido, amargo, claro, oscuro, fragante, ruidosos, silencioso.*
	Cualidades no físicas	*espirituales, vicios y virtudes, estado anímico, afecto, etc.*	*audaz, temeroso, orgulloso, perezoso, contento, aburrido, alegre, enemigo, indiferente*
Situacionales		*cuasi determinativos,*	*siguiente, próximo, pasado, referido*
		adverbiales modales	*posible, probable, presunto,*
		marcadores de la intensión	*único, propio, principal*
		circustanciales	*antiguo, próximo, actual*
		aspectuales	*frecuente, constante*

Tabla 1-18. Clasificación de los adjetivos calificativos, (Koike, 2001).

El colocativo en este caso es el adjetivo y al igual que en la construcción

sustantivo + verbo

o bien adquiere un valor funcional (adjetivos funcionales) o su significado léxico (adjetivos léxicos). Los adjetivos funcionales intensifican cuantitativa o cualitativamente el sustantivo; cuentan con dos posibilidades: que el sentido literal del adjetivo sea intensificador (adjetivos funcionales), o bien que el adjetivo se convierta en intensificador cuando se usa en sentido figurado (adjetivos funcionalizados). Un mismo adjetivo puede actuar como léxico o como funcional según el sustantivo con el que se combina. En la Tabla 1-19 se recogen ejemplos de ambos tipos de adjetivos colocacionales.

Funcional	Recto	*Dimensión física*	*grande* *profundo*	*pena grande* *profundo dolor*
		Cantidad	*mucho* *bastante* *poco*	
		Intensidad	*intenso* *fuerte* *débil*	*genio fuerte* *carácter débil*
	Figurado		*mortal* *terrible* *atroz*	*susto mortal* *prisa terrible* *actuación atroz*
Léxico	Recto		*afilado* *elocuente,* *mortal*	*cuchillo afilado* *orador elocuente* *veneno mortal*

Tabla 1-19. Clasificación de los adjetivos colocacionales, (Koike, 2001).

Las combinaciones

sustantivo abstracto + adjetivo funcional

como

pena grande

son colocaciones, el adjetivo se usa con valor intensificador; sin embargo,

sustantivo concreto + adjetivo funcional

como

casa grande

no pasa de ser una combinación libre. Las colocaciones funcionales en las que interviene un adjetivo funcional son amplias, mientras que en las que se usa un adjetivo léxico en sentido figurado son estrechas. En los ejemplos siguientes se presentan las definiciones de algunos adjetivos funcionalizados; como se aprecia, alguno de los significados son: *muy grande, extraordinario, muy bueno, etc*. –en general se está ante una intensificación positiva superlativa.

horroroso –sa

DRAE: Que causa horror. **2.** fam. Muy feo.

DEA 2. (*col*) Muy grande o extraordinario.

atroz

DRAE: 2. Enorme, grave.

3. fam. Muy grande o desmesurado.

DEA: 2. Muy grande o extraordinario.

fenomenal

DRAE: 3. Fam. Tremendo, muy grande.

4. fig. Estupendo, admirable.

DEA: 1. Muy bueno o extraordinario.

2. Muy grande o tremendo.

3. Muy bien.

Al contrario de lo que sucedía en las colocaciones

sustantivo + verbo

en este caso sí que se pueden tener sustantivos referidos a persona como en:

tonto redomado

lector empedernido

soldado bisoño...

En la definición de los adjetivos suele hacerse referencia a las bases con las que puede combinarse; en general, tales bases que no son únicas, sino que el vínculo se puede establecer con un conjunto de sustantivos pertenecientes al mismo campo semántico: *fino* se puede combinar con sustantivos que denotan sentidos, como *oído, vista u olfato*. (Corpas, 1996).

Por último se señala que existe un tipo de construcción:

sustantivo + sustantivo

que autores como Corpas o Castillo incluyen dentro de esta categoría, porque uno de los sustantivos actúa en realidad con función de adjetivo dentro de la misma, como en:

hombre clave

viaje relámpago...

sin embargo, para Koike es un tipo de compuesto.

1.7.9. Verbo + Adverbio.

El adverbio es el colocativo en estas combinaciones, en la mayoría de los casos es de modo –terminado en *mente*–; funciona como un intensificador de la acción denotada por el verbo como se ve en los siguientes ejemplos:

llover torrencialmente

desear ardientemente...

Algunas excepciones a esta característica son:

hablar claro

pisar firme...

La mayoría de las colocaciones de este grupo presentan una variante en la forma:

sustantivo + adjetivo

de manera que el sustantivo de éstas es deverbal y se corresponde con el verbo de aquéllas y el adverbio–terminado en *mente*– derivado del adjetivo como se aprecia en los ejemplos de la Tabla 1-20.

sustantivo + adjetivo	**verbo +adverbio**
lluvia torrencial	*llover torrencialmente*
comida opípara	*comer opíparamente*
cierre hermético	*cerrar herméticamente*

Tabla 1-20. Algunas colocaciones *sustantivo + adjetivo* y la correspondiente *verbo + adverbio*

Sin embargo, las colocaciones *sustantivo + adjetivo* que se citan a continuación no tienen correspondencia con las *verbo + adverbio*:

diferencia abismal

cuestión capital

sueldo ridículo

error garrafal

fe ciega…

ni tampoco:

diametralmente opuesto

clínicamente muerto

guardar celosamente…

presentan la versión *sustantivo + adjetivo*.

Por último incluimos en este apartado el caso: *verbo + gerundio* con valor adverbial como:

salir zumbando

salir pitando.

1.7.10. Adverbio + Adjetivo.

En este caso también el adverbio es el colocativo y actúa como intensificador, pero de un adjetivo –en muchas ocasiones es un participio con valor adjetival–, como en:

locamente enamorado

firmemente convencido

profundamente dormido

materialmente imposible...

aunque hay casos en que la base es un verdadero adjetivo como en:

locamente enamorado

mundialmente famoso...

Nótese que en el caso

adverbio + participio

suele existir la correspondiente versión:

verbo del que proviene el participio + adverbio

o bien:

sustantivo deverbal + adjetivo

como se observa en la Tabla 1-21.

Se pone de manifiesto la flexibilidad formal en las colocaciones, en las que se puede dar el cambio de categoría gramatical de los colocados.

sustantivo + adjetivo	**verbo + adverbio**	**adverbio + adjetivo**
relación estrecha		*relacionado estrechamente*
	dormir profundamente	*profundamente dormido*
	afectar visiblemente	*visiblemente afectado*

Tabla 1-21. Correspondencia entre las colocaciones adverbio + adjetivo y otras estructuras sintácticas

1.7.11. Verbo + Adjetivo.

Son muy pocas las colocaciones en este grupo, incluso hay autores que no contemplan tal caso en sus clasificaciones; ejemplos de ellas son:

{salir, resultar} ileso (de)

{salir, quedar} incólume (de)

{resultar, salir} redondo (el negocio)

andar {ajetreado, liado}

caer simpático, dejar frío...

1.7.12. Verbo + como + sustantivo.

Este tipo de colocaciones las introdujo Penadés en la tipología colocacional;.reproducen una estructura comparativa:

dar vueltas como un trompo

dormir como un tronco...

1.8. Las colocaciones en los diccionarios.

En español se dispone del DICE (Diccionario de Colocaciones del Español) de Alonso en el que se describen las restricciones combinatorias de un lexema mediante las funciones léxicas de Mel'čuk y los diccionarios Redes y Práctico de Bosque en el que se recogen las clases léxicas cuyos miembros coaparecen con un colocativo dado. Hay que mencionar también el Diccionario Fraseológico del Español Moderno (Varela, Kubarth, 1994), pero que solo recoge locuciones, complejos fraseológicos con casillas vacías y expresiones. Las dos primeras son locuciones, mientras que las expresiones coinciden con las fórmulas rutinarias de la clasificación de Corpas. En definitiva, tampoco recoge las colocaciones.

Además de los diccionarios específicos de colocaciones se puede explotar la información presente en los diccionarios tradicionales, especialmente en los de uso –Diccionario de Uso del Español (DUE, María Moliner) y el DEA (Diccionario del Español Actual, Seco)– sobre las restricciones combinatorias de las palabras.

1.8.1. Colocaciones en el DEA.

En las definiciones de este diccionario se especifica el **contorno**[14] de la misma, delimitado por corchetes y formado, en muchas ocasiones, por una lista de unidades léxicas que forman colocación con la unidad léxica que se define.

> **panorámico -ca I** *adj* **1** De(l) panorama. | FQuintana-Velarde *Política* 59: Tales son los rasgos básicos y generales .. Es preciso ahora obtener, además de la visión panorámica general que ya poseemos, algunos detalles concretos. **b)** [Imagen] que representa una vasta extensión de terreno. *Tb n f.* | * Es una fotografía panorámica del pueblo. **c)** [Autocar o vagón] que permite ver

[14] La definición está constituida por dos clases de elementos, el contenido –abarca el significado en sí– y el contorno –recoge el contexto.

cómodamente el paisaje en todas direcciones. | J. Oyarzun *Abc* 15.10.70, 11: Estas viejecitas de trajes floreados y sombreros de paja que a veces vemos circular por Madrid en grandes autobuses panorámicos .. esconden bajo su aparente fragilidad una gran entereza. **d)** (Cine) [Pantalla] de superficie cóncava y más ancha que la normal, destinada a películas en cinemascope y similares. | * Este cine tiene pantalla panorámica.

paranormal *adj* (*Psicol*) [Cosa, esp. fenómeno] que no tiene explicación científica. | *Ya* 13.2.75, 43: La autora estudia el tema agrupándolo así: creencias fantásticas y supersticiones; los muertos; .. facultades y fenómenos paranormales.

sentir[1] *(conjug* 60) A tr **1** Experimentar [una sensación o sensaciones]. *Tb abs*. | Arce *Testamento* 20: Comenzaba a sentir sed. Gambra *Filosofía* 135: Hay algunos que defienden el conocimiento en los vegetales, aduciendo a su favor ciertos hechos que parecen probar que los vegetales sienten. **b)** Experimentar [un estado afectivo o de ánimo]. | CNavarro *Perros* 92: Pensó en lo que sería la vida íntima de un ser como Fidel y sintió una lástima inmensa por su mujer. **c)** Ser afectado [por algo (*cd*)]. | *Ama casa* 1972 91: Los rododendros. Es un árbol hermoso .. Se cultiva en tierra de brezo y siente mucho el frío.

A la vista de los ejemplos se observa que en algunos casos se dan rasgos generales para especificar el contorno como "*una sensación o sensaciones*" en la acepción 1 de sentir, "*estado afectivo o de ánimo*", en el b). En el caso de paranormal se tiene un elemento general como "*cosa*" y uno específico como "*fenómeno*" y en panorámico solo se dan elementos específicos en el contorno como "*imagen*" en b), "*autocar o vagón*" en c) y "*pantalla*" en d). Obsérvese que en todos estos casos se puede decir que se está ante colocaciones formadas por el lexema de la entrada y los elementos del contorno; aunque esto no sucede siempre, sí son numerosas las oportunidades en las que se da. Nótese que en algunos casos se obtiene una restricción combinatoria entre la entrada y una clase léxica:

"*sentir + sustantivo de estado afectivo o de ánimo*"[15]

mientras que en los restantes aparecen colocaciones específicas:

imagen panorámica
autocar panorámico
vagón panorámico
pantalla panorámica
fenómeno paranormal

[15] Koike también resalta esta característica para indicar las posibilidades de combinación del verbo sentir: "el verbo sentir puede combinarse con sustantivos que expresan 'estado anímico o físico' y 'sentimiento'".

Si bien "*cosa paranormal*" no es una colocación, en el ejemplo aparece

facultad paranormal

que sí lo es.

En general, cuando el lexema que se está definiendo es el colocativo, las bases de las colocaciones se recogen en el contorno, aunque en el caso de que el colocativo sea un adjetivo también se pueden encontrar en las explicaciones en cursiva que utilizan expresiones del tipo:

"*Dicho normalmente de, Referido esp. a,* etc.".

Por otra parte, si en el contorno se habla de rasgos generales que definen una clase léxica, no quiere decir que todos los elementos que presenten ese rasgo formen colocación, puede haber excepciones. Además, como se aprecia en el apartado b) de panorámico solo hay un elemento en el contorno: "imagen"; sin embargo,

fotografía panorámica

también es colocación, si bien fotografía es un vocablo que se refiere a un tipo de imagen. En definitiva, a pesar de que en algunos casos si se consideran estrictamente los contornos, se puede incurrir en un exceso o defecto de producción de las bases que combinan con un colocativo, no supone un obstáculo para afirmar que la información contenida en el contorno es muy valiosa a la hora de caracterizar las restricciones combinatorias de los colocativos.

En algunos casos, también se dispone de los verbos con los que se combina un determinado sustantivo, como ejemplifica la Tabla 1-22; en muchos casos, el marcador *frec con el v* en la definición de algunos sustantivos en el DEA se corresponde con colocaciones.

Sustantivo	**Frec con el v**
gustirrinín	*dar*
leñazo	*dar*
golpe	*dar*
golletazo	***dar***

Tabla 1-22. Ejemplos del DEA.

1.8.2. Colocaciones en el DUE.

En las bases es donde se presentan las posibilidades combinatorias de los lexemas –en forma de enlaces frecuentes. Para los sustantivos se dan: los adjetivos y verbos; para los verbos, los adverbios y las preposiciones con que se construyen.

En el caso de los adjetivos, las colocaciones o las solidaridades léxicas pueden aparecer señaladas por "*se aplica a*", "*aplicado a*", "*dícese de*" o "*se dice (de)*". También se introducen en los ejemplos que aparecen en letras simples y entrecomilladas, en el apartado de palabras afines o bien en el de frases y modismos, en cursiva y al final del artículo (Corpas, 1992).

1.8.3. Diccionario VOX

Este diccionario se puede considerar como paradigma de la precisión semántica que aportan las colocaciones a la lengua, abundando su uso en las definiciones de los lemas que recoge. Algunos ejemplos en los que se manifiesta este fenómeno

animar: Infundir ánimo y energía

ilusionar: Despertar esperanzas, gralte. atractivas

truncar: Quitar [a uno] las ilusiones o esperanzas

confiar: Dar confianza o esperanza [a uno]

desahuciar: Quitar [a uno] las esperanzas de conseguir lo que desea

hechizar: Cautivar el ánimo, embelesar

Del mismo modo que en el DEA, se observan colocaciones en los sustantivos marcados como contornos de los verbos:

arriar: Bajar [una vela o bandera que estaba izada]

enarbolar: Levantar en alto [estandarte, bandera, etc.]

largar: Desplegar [la bandera, las velas, etc.]

vibrar: Agitar en el aire [la pica, la lanza, etc.]; arrojar con ímpetu y violencia [una cosa que vibre]

suspender: especialmente, figurado. Privar temporalmente [a uno del sueldo o empleo] que tenía

levantar: Vigorizar [el ánimo, la moral, etc.]; engrandecer, ensalzar; impulsar hacia cosas altas [el pensamiento, el corazón, etc.]

Aparecen colocaciones funcionales cuando se definen verbos relacionados con un sustantivo:

desesperanzar: Quitar la esperanza [a uno]

esperanzar: Dar esperanza [a uno]

esperanzar: Tener esperanza

apasionar: Causar, excitar alguna pasión [a uno]

apasionar: Llenarse de pasión

respetar: Tener respeto

En cuanto a los adjetivos, las colocaciones se observan en el contorno:

acéfalo: [sociedad, secta, etc.] Que no tiene jefe.

acéfalo: [feto] Sin cabeza o sin parte considerable de ella.

abisal: [zona del mar profundo] [zona del mar profundo] Que se extiende más allá del talud continental, y corresponde a profundidades mayores de 2000 m

frío: por extensión. [color] Que produce efectos sedantes como el azul, el verde, etc.

imberbe: [joven] Que no tiene barba.

interactivo: [programa] Que permite una interacción, a modo de diálogo, entre el ordenador y el usuario.

recurrente: [fenómeno] Que vuelve a su punto de partida.

visceral: [impresión, sentimiento, etc.] Intenso, profundo y arraigado: odio visceral.

1.9. Elementos ajenos a las colocaciones.

Por la utilidad que puede suponer a la hora de implementar un reconocedor automático de colocaciones, se incluye una recopilación de elementos que se deben descartar por no formar parte de las mismas.

El verbo *ser* no puede formar colocaciones sustantivo + verbo.

Los adjetivos demostrativos, indefinidos y relativos.

Verbos auxiliares.

Los verbos derivados de sustantivos que son bases de colocaciones funcionales: *dar un consejo / aconsejar.*

Los verbos *querer, creer, pensar, indicar, afirmar, preguntar, verbos de decir, asegurar, advertir, vender, comprar, agradecer, obligar, ayudar, invitar, gustar, parecer, interesar, considerar, creer, hallar, poder, saber, querer, intentar*, etc.

Los sustantivos propios en las sustantivo + verbo.

Los numerales o indefinidos.

Los adjetivos situacionales.

Los adverbios funcionales que solo intensifican o cuantifican la acción denotada por el verbo: *absolutamente, enteramente, completamente, intensamente, tremendamente, terriblemente, etc.*

Adverbios de tiempo, de cantidad, de afirmación, de negación, de duda o de lugar.

2. Extracción automática de colocaciones utilizando criterios estadísticos.

2.1. Medidas de asociación

Las técnicas de extracción automática de colocaciones se basan en el criterio de la frecuencia de aparición conjunta de palabras en corpus textuales como manifestación cuantificable de la característica del uso preferente frente a las combinaciones libres. Los primeros estudios se deben a Firth (1957), quien utiliza por primera vez el término **collocation** para referirse a la coaparición de palabras. A partir de los trabajos de Firth surge la llamada escuela sistémica británica a la que pertenecen Halliday, Sinclair, Jones; para tal corriente, las colocaciones se reducen a combinaciones frecuentes de palabras; por utilizar el criterio de la preferencia como rasgo principal de las colocaciones, en sus trabajos se propone el uso de técnicas estadísticas para la exploración de textos en busca de combinaciones de palabras cuya frecuencia de aparición pueda considerarse superior a la de otras combinaciones.

El análisis de las combinaciones en esta línea forma parte de la disciplina conocida como Lingüística de Corpus, en la que se consideran esenciales los indicadores de los vínculos motivados entre los dos elementos de la colocación. Por lo general, determinan una puntuación para cada par de palabras que puede usarse para establecer un ranking, o bien seleccionar aquellos casos que se confirman como colocaciones por medio de valores de corte. Se basan en la frecuencia de aparición conjunta de palabras y en general intentan captar anomalías respecto a lo que se espera que suceda si la aparición de la combinación en el corpus se debiera a la mera casualidad: las variables aleatorias que representaran la aparición en el corpus de los elementos que constituyeran las combinaciones libres serían independientes. La independencia estadística ocurre cuando no existe relación alguna entre ambos fenómenos, ninguno de ellos incide en la ocurrencia del otro. En otras palabras, las probabilidades de ocurrencia conjunta, *P(x, y)* e individual de base y colocativos, *P(x), P(y)* respectivamente, verifican la relación:

$$P(x, y) = P(x) * P(y)$$

Desde los primeros trabajos en la materia se han ido proponiendo distintas modalidades de indicadores que van desde el más sencillo y directo, la **frecuencia relativa** a los test estadísticos como **z-score**, **t-score**, **fórmula de Dunning**, **test de Poisson**, o la **información mutua** adaptada de la teoría de la información. Todos ellos se calculan a partir de valores de frecuencias de aparición conjunta de dos palabras en el corpus: $f(x, y)$ y de la frecuencia de aparición de forma individual de cada una de ellas: $f(x)$, $f(y)$. En base a estos valores se puede

obtener como estimación de la probabilidad la frecuencia relativa: $p(x) \cong \frac{f(x)}{N}$, siendo *N* la cantidad de veces que se repite el experimento, en este caso, la cantidad total de palabras en el corpus. En virtud de la ley de los grandes números, la frecuencia relativa de un suceso se aproxima a su verdadera probabilidad cuanto mayor es el número de repeticiones de un suceso. Es decir, cuanto mayor sea el número de repeticiones del experimento, mayor será la fiabilidad de la aproximación.

2.1.1. Frecuencia Relativa

También denominada frecuencia de aparición de *x* con *y* (Koike, 2 001); en realidad, es un indicador del porcentaje de veces que una palabra aparece con otra respecto al número total de veces que aparece.

$$frec(x|y) = \frac{frec(x,y)}{frec(x)} \times 100$$

Si este valor es elevado, quiere decir que si aparece *x* es muy posible que también aparezca *y*; no tiene por qué presentar el mismo valor que la frecuencia de aparición de *y* con *x*. Para Koike es suficiente indicio de colocación que el valor sea superior al 20%.

2.1.2. Información Mutua

La medida de **información mutua** puntúa la coocurrencia de dos palabras mediante la expresión (Church, Hanks, 1 990):

$$I(x,y) = \log_2 \frac{p(x,y)}{p(x) * p(y)}$$

Tal valor mide la cantidad de información (en bits) que la coocurrencia aporta sobre la información de las apariciones individuales de las dos palabras. De forma intuitiva, la información mutua mide en qué cantidad se reduce la incertidumbre de una variable *X* debido al conocimiento de la variable *Y*. En este caso se interpreta como la información que la aparición de una palabra aporta sobre la otra, cuando la información mutua es nula, estamos ante sucesos independientes. Las probabilidades se calculan directamente usando frecuencias relativas. Esta medida es inestable para cantidades pequeñas –se calcula solo si la frecuencia absoluta $f(x,y) > 5$.

La información mutua constituye un indicador más complejo y preciso que la frecuencia relativa utilizado para la extracción de colocaciones en Lingüística de Corpus.

2.1.3. z-score

El **z-score** mide si es significativa la frecuencia de aparición de *x* con *y* en el corpus, respecto a la frecuencia esperada, bajo la hipótesis de que las variables son independientes. Se basa en la aproximación normal de la variable aleatoria binomial que considera como éxito el que la palabra *x* ocurra con *y*:

$$z = \frac{f(x,y) - \bar{f}(x,y)}{\sqrt{f(x,y) * \big(1 - p(y)\big)}}$$

$$\bar{f}(x,y) = p(y) * f(x) * |D|$$

$$p(y) = \frac{f(y)}{N - f(x)}$$

Siendo *p(y)* la probabilidad de que ocurra *y*, en una posición distinta en el corpus, *f(x)* la frecuencia de *x*, *N* es el número de palabras en el corpus, $|D|$ el número de posibilidades en las que puede aparecer *y* alrededor de *x* –coincide con dos veces la distancia colocacional que se esté usando en las líneas de concordancias (Pearce, 2 002).

2.1.4. t-score

Este valor se utiliza ampliamente para la extracción de colocaciones y se inspira en la aplicación del *t-test de Student* para la media con varianzas desconocidas en distribuciones normales. Sin embargo, la distribución de las frecuencias de coocurrencia de palabras en el corpus no cumple con las condiciones de la teoría estadística en que se fundamenta el test original, por lo que su buen comportamiento le da rango de valor heurístico del z-score y no de un test de hipótesis (Evert, 2005).

$$t = \frac{f(x,y) - \frac{f(x) * f(y)}{N}}{\sqrt{f(x,y)}}$$

2.1.5. Test de Dunning

Se determina el ranking de colocaciones utilizando la razón de verosimilitud como test estadístico fiable con independencia del tamaño del corpus, ya que no se requiere la exigencia de normalidad en la distribución de la variable (Dunning, 1993). La razón de verosimilitudes se

calcula para la distribución binomial en la que bajo la hipótesis nula, la aparición de las dos palabras es independiente:

$$\log \lambda = \log L(f(x,y), f(x), p) + \log L(f(y) - f(x,y), N - f(x), p)$$
$$- \log L(f(x,y), f(x), p_1) - \log L(f(y) - f(x,y), N - f(x), p_2)$$

$$p = \frac{f(y)}{N}, p_1 = \frac{f(x,y)}{f(x)}, p_2 = \frac{f(y) - f(x,y)}{N - f(x)}()$$

$$\log L(k, n, p) = k \log p + (n - k) \log(1 - p)$$

$$p = \frac{f(y)}{N}, p_1 = \frac{f(x,y)}{f(x)}, p_2 = \frac{f(y) - f(x,y)}{N - f(x)} \text{ (Manning, Schütze, 1 999)}$$

2.1.6. Test de Poisson

Este test se construye bajo la suposición de que las coocurrencias siguen una distribución de Poisson, que corresponde a la extensión para grandes muestras de la distribución binomial. El valor se determina (Quasthoff, Wolff , 2002):

$$sig(x,y) = \frac{-\log p_k}{\log n} \approx \frac{\lambda - k \log \lambda + \log k!}{\log n}$$

$$\lambda = np_x p_y$$

$$p_k = (más\ de\ k\ ocurrencias\ de\ x, y) = \sum_{l=k}^{\infty} \frac{1}{l!} \lambda^l e^{-\lambda}$$

2.2. Base de datos de información combinatoria.

Los valores de las distintas medidas de asociación de una combinación de palabras quedan determinados por las frecuencias de aparición conjunta de los elementos que la componen así como de las apariciones individuales de cada uno de ellos. Estos datos se deben obtener a partir de un corpus textual que será considerado como una muestra del lenguaje, por lo que es importante garantizar que sea representativo. Asimismo, su volumen debe ser tal que permita su tratamiento mediante computadores en un tiempo razonable. Este trabajo se centra en las colocaciones léxicas del español, por lo que se recopilaron una gran cantidad de textos, de una amplia gama de géneros, albergando una colección de unos 11000 textos aproximadamente. Entre otros géneros este corpus reúne obras de literatura, tanto clásica como contemporánea,

española y universal, poesía y prosa, teatro, narrativa, ensayos, discursos y artículos periodísticos.

2.2.1. Características del Corpus

El corpus utilizado consiste en una colección de textos planos, sin ningún tipo de etiquetas que proporcionen información lingüística. Además de obras periodísticas se pueden encontrar obras literarias y obras no literarias. De las primeras, entre las que se han catalogado cuentos, relatos, poesía, novelas y otras encontramos 7758. En el segundo grupo se recopilaron 4108 tratados sobre arte, biografías, ciencias, cine, cristianismo, derecho, economía, educación, esoterismo, etnología, filosofía, geografía, historia, lingüística, literatura, política, psicología, religión, teatro y otros. Además, un grupo aparte lo conforman los periódicos. La cantidad de obras recogidas en cada uno de estos grupos se muestra en la Tabla 2-1 y la Tabla 2-2.

Tipo	
arte	21
biografías	85
ciencias	35
cine	220
cristianismo	44
derecho	513
economía	35
educación	40
esoterismo	22
etnología	21
filosofía	170
geografía	85
historia	265
lingüística	25
literatura	1068
política	346
psicología	70
religión	103
teatro	63
otros	877
Total	4108

Tabla 2-1 Distribución de las obras no literarias

Tipo	
novela	2053
teatro	1265
relato	2248
cuentos	632
otros	596
Total	7758

Tabla 2-2 Distribución de las obras literarias

En definitiva, una amplia muestra del español con un número total de palabras que está en torno a los 300 000 000, cuya distribución se resalta en el Gráfico 2-1 y Gráfico 2-2

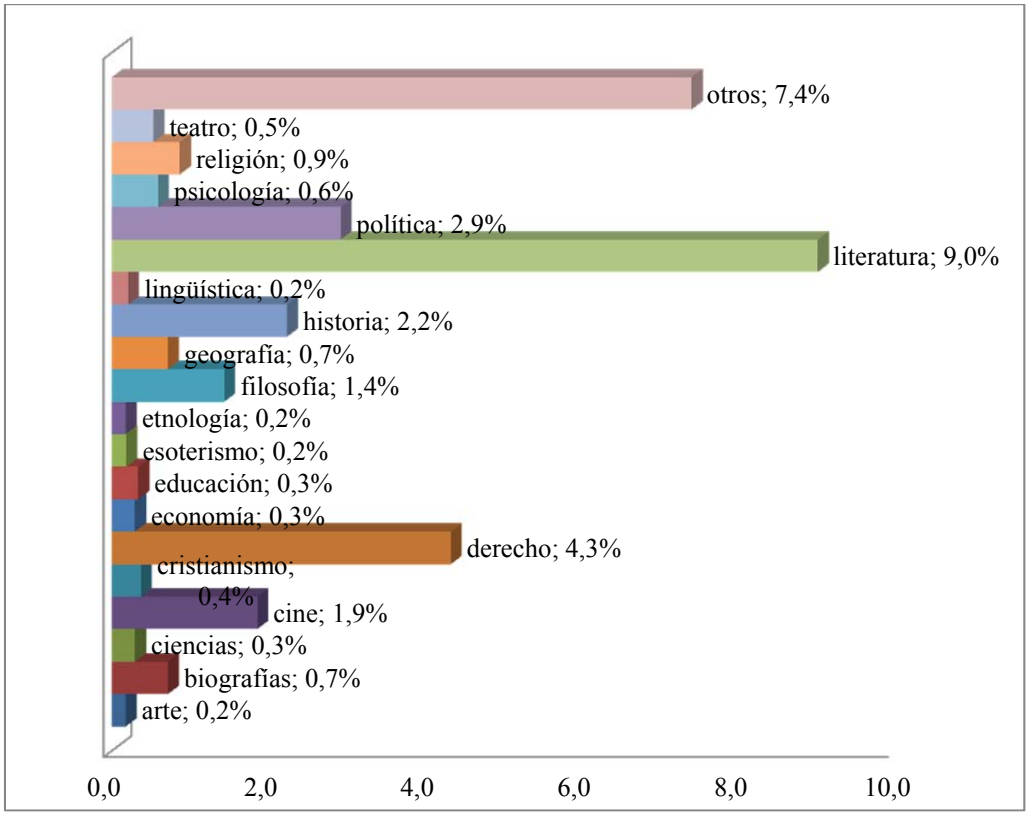


Gráfico 2-1 Distribución de las obras no literarias del corpus

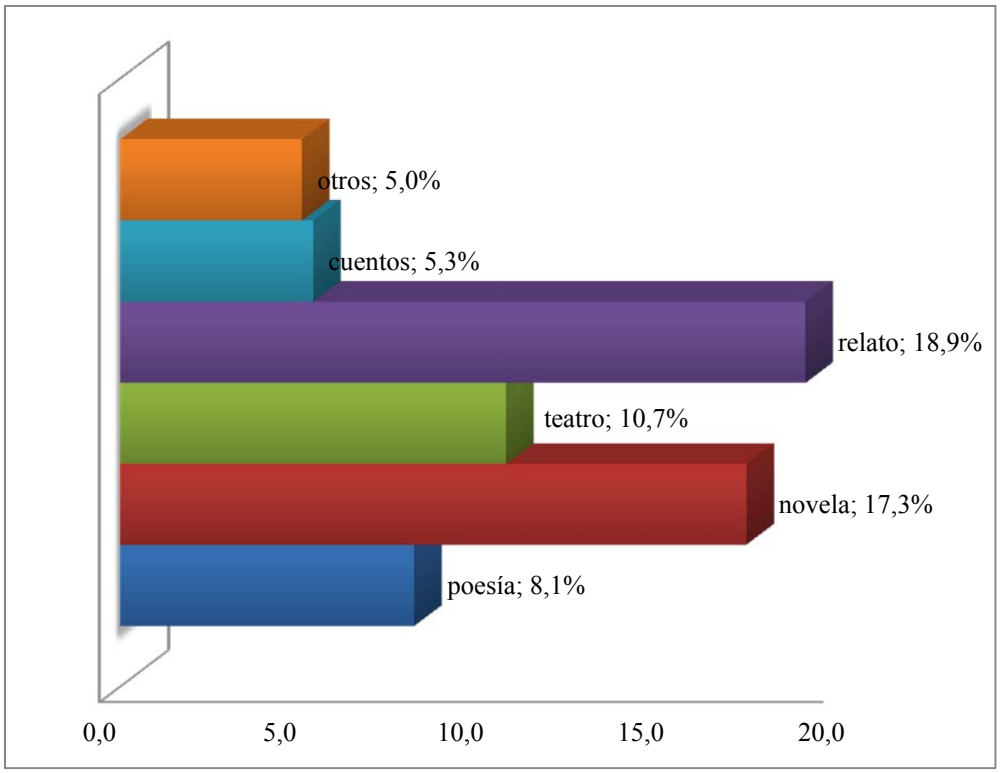


Gráfico 2-2 Distribución de las obras literarias del corpus

2.2.1.1. Tratamiento del corpus

El primer proceso llevado a cabo en esta investigación lo constituye la obtención de todas las combinaciones que se toman en consideración así como su frecuencia de aparición en el corpus. Los cálculos se realizan a partir de las llamadas *líneas de concordancia* extraídas de éste. Una línea de concordancia es un determinado fragmento de una longitud fija, en términos de cantidad de palabras que abarca, en el que aparece una palabra dada en un texto o corpus textual. La aparición conjunta de una determinada combinación en el ámbito de la línea de concordancia se considera una manifestación del uso conjunto de la misma o muestra de la combinación. En la Figura 2-1 se exponen líneas de concordancias obtenidas en el *corpus de referencia del español actual* (CREA) cuando se consulta sobre la palabra *arrojar*.

puesto muy cuesta arriba para su equipo, pero sin **arrojar** la toalla, porque "restan 90 minutos y no se
cuatro años. Pero al nuevo presidente le será difícil **arrojar** de sí la mala conciencia de tenerse de presta
el exilio no significa que el volcán deje de **arrojar** incertidumbre. Florentino Pérez vuelve a la carga, reclamando la
profesionales y características psicológicas de cada persona, así como **arrojar** un listado de arrojar un listado de los quince

Figura 2-1 Líneas de concordancia de la palabra arrojar en el CREA

Se salva en parte la falta de cualquier tipo de información lingüística en los textos gracias a las pautas marcadas por algunas características fundamentales de las colocaciones léxicas en español que se expusieron en el capítulo 1 y que se enumeran a continuación:

a) Debido a la flexibilidad formal de las colocaciones se recuentan combinaciones de formas canónicas en lugar de combinaciones de palabras gráficas. Según esto se consideran muestras de la combinación *progreso-civilización: el progreso de las civilizaciones, la civilización progresó o los progresos de una civilización,* o bien de *reprobar-condena: reprobaron la condena, reprobó la condena, las condenas fueron reprobadas* o de *guerra-civil: guerra civil, las guerras civiles, la guerra entre civiles*. Para ello se ha utilizado el *"Flexionador y Lematizador de palabras del español" del Grupo de Estructura de Datos y Lingüística Computacional de la Universidad de Las Palmas de Gran Canaria* (Santana et al, 2007, 1999, 1997).
b) Se considera un catálogo de *palabras vacías* como artículos, preposiciones, conjunciones, interjecciones, determinantes numerales, determinantes ordinales, adjetivos demostrativos, indefinidos y el relativo *cuanto,* los verbos *ser, estar y haber*, etc. descartándose del proceso todas las que se incluyeron en él. (Koike, 2001).
c) Se consideran exclusivamente las estructuras colocacionales (Koike, 2001) que se pueden encontrar en el español, desde la perspectiva del Lematizador se pueden agrupar en los casos:

sustantivo + verbo
sustantivo + adjetivo
sustantivo + de + sustantivo
verbo + adverbio
verbo + adjetivo
adverbio + adjetivo

por lo que en el recuento se contabilizan solo aquellas combinaciones que se corresponden con alguno de estos patrones en orden indistinto.

El programa simula la obtención de líneas de concordancias para cada una de las formas canónicas en el corpus, generando un diccionario de formas canónicas y otro de combinaciones de formas canónicas a partir de ellas. Cada línea de concordancia está dentro del ámbito de una frase y se toma un ancho de ventana 10. Es decir, se recopilan los valores de las frecuencias conjuntas para distancias entre palabras que van desde 1 hasta 10, tanto a la izquierda como a la derecha. Se obtienen las combinaciones de formas canónicas que se corresponden con alguna de las estructuras de las colocaciones en español y su frecuencia en el corpus. Si el tamaño de frase es menor que la amplitud fijada para la concordancia, no se complementa con palabras en la

siguiente frase. Cada elemento en la línea de concordancia actual que no sea una palabra vacía se procesa con el lematizador que proporciona la correspondiente información morfológica.

Para una mayor eficiencia en el procesamiento, cada palabra en el corpus solo se lee una vez; se almacena temporalmente la información morfológica que se genera gracias a una estructura de cinta circular, con tamaño igual a la distancia colocacional. En cada iteración se determina el elemento de dicha cola donde se deben registrar las formas canónicas con las que tiene correspondencia la palabra actual, en los restantes estarán almacenadas las formas canónicas que se obtuvieron cuando fueron procesadas las palabras previas en el corpus. Cuando se lee la palabra siguiente, pasará a ser palabra actual. Las palabras previas permanecen en la cinta siempre que se esté en el ámbito de la línea de concordancia. Para lograr esto, en cada iteración antes de insertar la información de la palabra actual extraída del corpus, se agregan al diccionario las combinaciones compuestas por las formas canónicas en el registro de la cinta que se va a actualizar y las formas canónicas en todos los restantes elementos en ella, siempre que correspondan a alguna de las estructuras colocacionales del español. En cada caso se actualizará la frecuencia si fue registrado previamente o creando la nueva entrada en caso contrario. Por último, es necesario especificar que si la palabra actual es vacía, no se lematiza y se deja la posición correspondiente en la cola vacía. De esta forma se logra obtener las parejas de formas canónicas con palabras previas y con palabras posteriores en la línea de concordancia. A continuación se ilustra el procesamiento realizado al corpus sobre un fragmento de Odisea 2001, de Arthur C. Clarke, novela que forma parte del corpus, en concreto, el primer párrafo:

> La sequía había durado ya diez millones de años, y el reinado de los terribles saurios tiempo a que había terminado. Aquí en el ecuador, en el continente que había de ser conocido un día como África, la batalla por la existencia había alcanzado un nuevo clímax de ferocidad, no avistándose aún al victorioso. En este terreno baldío y disecado solo podía medrar, o aun esperar sobrevivir, lo pequeño, lo raudo o lo feroz. (Odisea 2001, Arthur C. Clarke).

La tabla se utiliza para representar la cola circular en la que se irá insertando temporalmente la información morfológica de cada palabra que se va leyendo en el fichero. Se muestra dicha estructura después de haber leído las 5 primeras palabras, las casillas sin información se corresponden a palabras vacías (Figura 2-2).

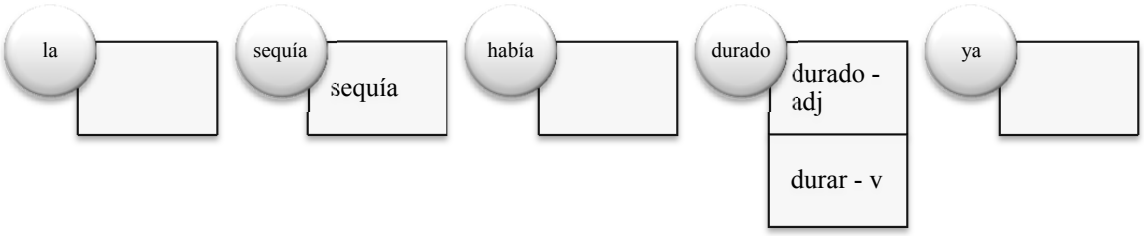


Figura 2-2 Inicialización de la cinta para la extracción de combinaciones

En la siguiente iteración la ventana se desplaza hacia la palabra *diez*, que será la siguiente que se lea en el corpus. La palabra actual es *diez* y se inserta en la cinta, no es necesario agregar información a los diccionarios de combinaciones, porque en la posición en la cola circular no hay elementos que correspondan previamente a una palabra vacía. Por el contrario, el diccionario de formas canónicas procesadas será actualizado con la información que produce la lematización de *diez* (Figura 2-3).

~~La~~ sequía ~~había~~ durado ~~ya~~ diez millones ~~de~~ años, ~~y el~~ **reinado** ~~de los~~ terribles saurios tiempo ~~a~~ ~~que~~ había terminado. ~~Aquí en el~~ ecuador, ~~en el~~ continente ~~que había de~~ ser conocido ~~un~~ día ~~como~~ Africa, ~~la~~ batalla ~~por la~~ existencia ~~había~~ alcanzado un nuevo clímax ~~de~~ ferocidad, ~~no~~ avistándose ~~aún al~~ victorioso. ~~En este~~ terreno baldío ~~y~~ disecado ~~soo~~ podía medrar, ~~o aun~~ esperar sobrevivir, ~~lo~~ pequeño, ~~lo~~ raudo ~~o lo~~ feroz.

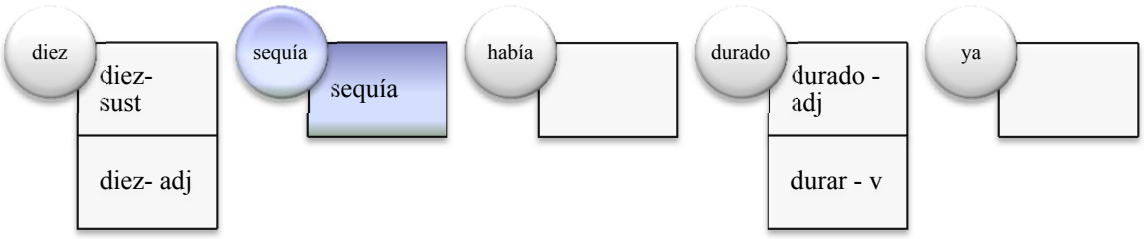


Figura 2-3 Actualización de la cinta para la extracción de combinaciones.

A continuación, se inserta la información morfológica relativa a *millones;* debe actualizarse las combinaciones de formas canónicas, es decir, la forma canónica *sequía / sustantivo* y sus combinaciones factibles con todas las formas canónicas en la cinta en ese instante: *sequía + durado, sequía + durar, sequía + diez, sequía + diez*

~~La~~ sequía ~~había~~ durado ~~ya~~ diez millones ~~de~~ años, ~~y el~~ **reinado** ~~de los~~ terribles saurios tiempo ~~a~~ ~~que~~ había terminado. ~~Aquí en el~~ ecuador, ~~en el~~ continente ~~que había de~~ ser conocido ~~un~~ día ~~como~~ Africa, ~~la~~ batalla ~~por la~~ existencia ~~había~~ alcanzado un nuevo clímax ~~de~~ ferocidad, ~~no~~ avistándose ~~aún al~~ victorioso. ~~En este~~ terreno baldío ~~y~~ disecado ~~solo~~ podía medrar, ~~o aun~~ esperar sobrevivir, ~~lo~~ pequeño, ~~lo~~ raudo ~~o lo~~ feroz.

Una vez actualizado el diccionario de combinaciones, se elimina sequía de la cinta y se agrega en su lugar millones, quedando con la configuración en la Figura 2-4.

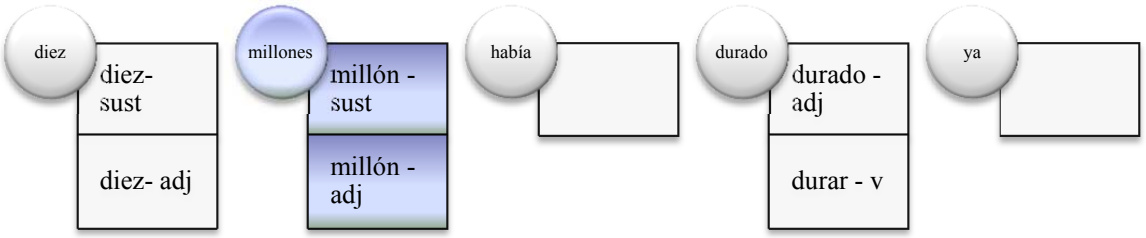


Figura 2-4 Actualización de una posición no vacía de la cinta de combinaciones.

Para mantener ambos diccionarios, el de formas canónicas y el de combinaciones de éstas se utiliza tablas hash cuyo tipo base permite registrar tanto el elemento como la frecuencia de aparición del mismo. De esta forma la ejecución permite generar la información de formas canónicas correspondientes a palabras no vacías y sus frecuencias en el corpus.

El lematizador del GEDLC permite identificar las formas canónicas cuya flexión o derivación produce la palabra por la que se consulta. Sin sumar la ampliación inherente a los prefijos y a los pronombres enclíticos, el universo está compuesto por alrededor de cuatro millones y medio de formas flexionadas y derivadas obtenidas a partir de las más de ciento veinticuatro mil formas canónicas con que opera el sistema. En los verbos, trata la conjugación simple y compuesta, los pronombres enclíticos, la flexión del participio como adjetivo verbal (género, número, grado superlativo y adverbialización) y el diminutivo del gerundio. Con las formas no verbales, considera: género y número en los sustantivos, adjetivos, pronombres y artículos; heteronimia por cambio de sexo en los sustantivos; grado superlativo en los adjetivos y adverbios; adverbialización y adverbialización del superlativo en los adjetivos; derivación apreciativa en los sustantivos, adjetivos y adverbios; variantes gráficas en todas las categorías gramaticales; formas invariantes tales como preposiciones, conjunciones, exclamaciones, palabras de otros idiomas y locuciones o frases. Se contempla la prefijación. (Santana et al, 2007). Por motivos de rendimiento, debido a la extensión del corpus, se ha utilizado la versión local, si bien está disponible como servicio web. En la Figura 2-6 se muestra la información que se obtiene al lematizar una palabra, se ha utilizado para generar el ejemplo la aplicación web disponible para consultas.

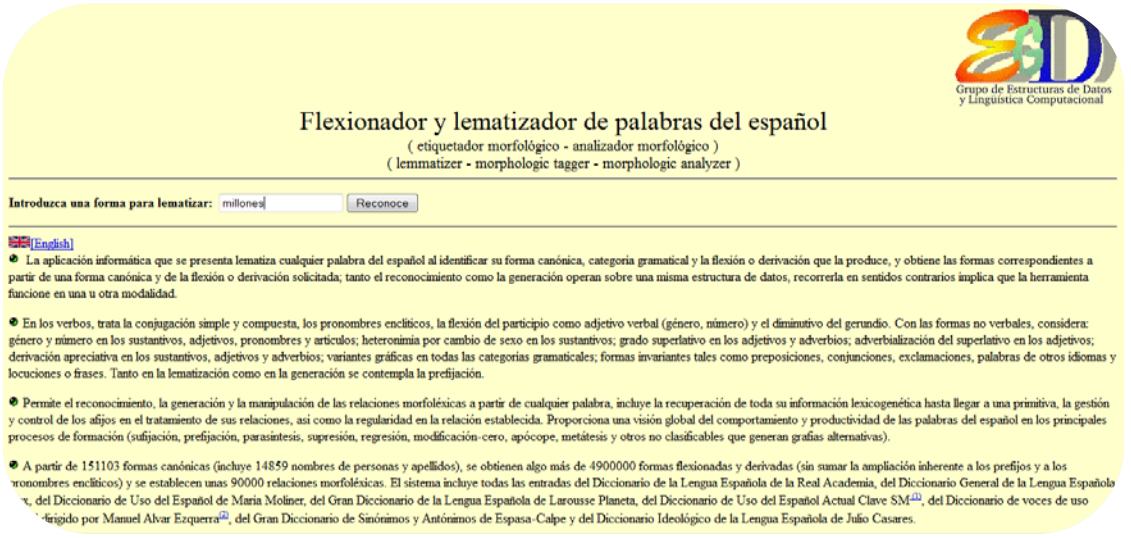


Figura 2-5 Lematizador del GEDLC de la ULPGC

http://www.gedlc.ulpgc.es/investigacion/scogeme02/lematiza.htm

Figura 2-6 Respuesta a una consulta realizada al lematizador

Se hace necesario hacer notar que al utilizar exclusivamente información morfológica es inviable determinar qué forma canónica es la que produce flexión o palabra derivada que se haya leído del corpus. Por este motivo para cada palabra se contemplan todas las formas canónicas posibles. Desde el punto de vista de las colocaciones, en una gran parte de los casos se producirán combinaciones de formas canónicas con distintas categorías correspondientes a la misma colocación, como manifestación de la característica de la flexibilidad formal.

La ejecución del programa sobre el corpus de trabajo produce un total de 14475136 combinaciones distintas con frecuencia de aparición mayores o iguales que 3 en alguna de las

distancias. Las combinaciones con frecuencia 1 ó 2 se desprecian debido a la explosión de combinaciones irrelevantes que producen.

2.3. La Base de Datos

Con el objetivo de manipular la cantidad ingente de información generada a partir del tratamiento del corpus de la forma más eficiente posible se almacenaron los resultados en una BDD (Base de Datos) en la que inicialmente se recogen formas canónicas y su frecuencia de uso así como combinaciones de formas canónicas y su frecuencia de uso. Se muestra en el Gráfico 2-3 la distribución de las combinaciones registradas en la BDD según las categorías gramaticales de los dos elementos de la combinación, como grupos más numerosos se catalogan 6 551 979 en el tipo *sustantivo + verbo* y 3 743476 en el tipo *sustantivo + adjetivo*.

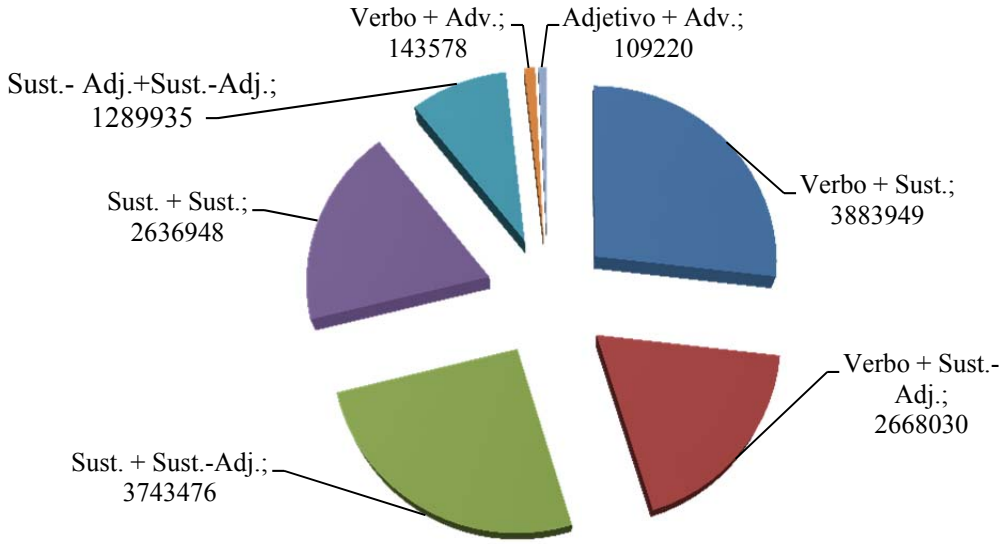


Gráfico 2-3 Clasificación de las combinaciones por categorías gramaticales de los elementos

A partir de la información inicial de frecuencias extraídas mediante el tratamiento del corpus, y utilizando el sistema gestor se amplía el conjunto de datos con los valores de los indicadores: **frecuencia relativa**, **z-score**, **t-score** y **test de Dunning** evaluados sobre todos los datos del corpus. Se incorpora también el catálogo de colocaciones recopiladas en los artículos que versan sobre colocaciones léxicas en español consultados para el desarrollo del trabajo (Alonso, 1994-1995), (Alonso, 2002), (Bargalló, 1997-1998), (Blasco, 2002), (Bosque, 2001), (Castillo, 1997-1998), (Castillo, 1998), (Castillo, 2001), (Corpas, 1996), (Corpas, 2001), (Corpas, 2003), (García 2002), (García-Page, 2001), (Koike, 2001), (Moreno, 1998), (Penadés, 2001), (Varela, 1994), (Zuluaga, 2002) y en el DCECR. Tanto unas como otras se usan para contrastar

la capacidad de las distintas técnicas basadas en la frecuencia a la hora de detectar colocaciones. El primer grupo se denominará a partir de ahora *combinaciones recopiladas*, donde se encontraron 2 356 combinaciones, y en el DCECR se registran 55 052.

En la Figura 2-7 se muestra el grafo relacional que refleja cómo está organizada la información.

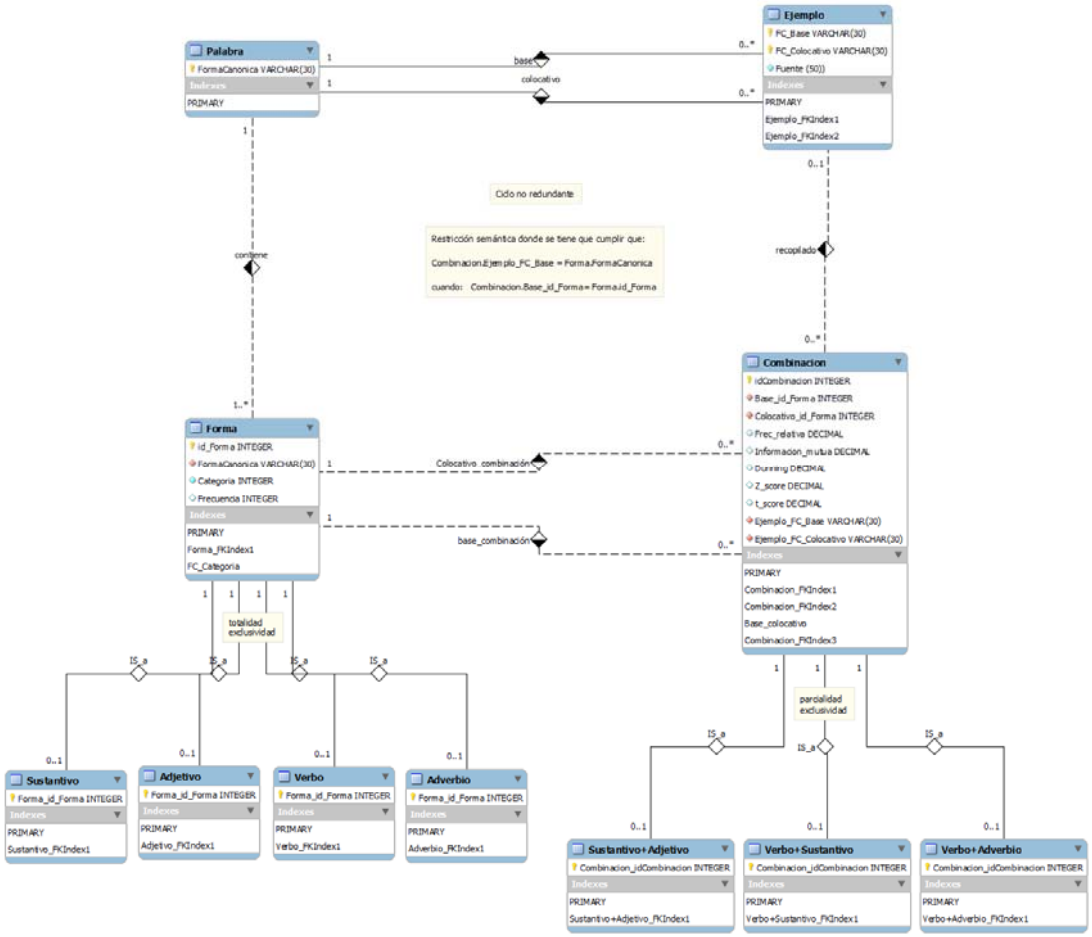


Figura 2-7 Grafo relacional de la Base de Datos de Combinaciones

A continuación se describen cada una de las entidades y sus atributos.

Entidad
Ejemplo
Descripción
Entidad que contiene la información de las combinaciones recopiladas y del DCECR.
Atributos
FC_Base: Forma Canónica que es base de la colocación. FC_Colocativo Forma Canónica que es colocativo de la colocación. Fuente: Puede ser Recopilada o DCECR.

Entidad
Palabra
Descripción
Entidad que contiene la información de la palabra que se toma para la forma canónica, a diferencia de ésta que abarca también la categoría, la frecuencia entre otros.
Atributos
FormaCanonica: Palabra que corresponde a una forma canónica.

Entidad
Forma
Descripción
Entidad que contiene la información de una Forma Canónica.
Atributos
id_Forma: identificador de la Forma Canónica FormaCanonica: Forma Canónica Categoría: Categoría de la Forma Canónica Frecuencia: Frecuencia de aparición de la forma canónica en el corpus.

Entidad
pg
Descripción
Entidad que contiene para cada grupo del Diccionario Ideológico Vox su probabilidad a priori. Pueden ser grupos de *verbos, sustantivos, adjetivos o adverbios*.
Atributos
idGrupo: identificador del grupo pg: probabilidad a priori del grupo idGrupo

Entidad
Combinación
Descripción
Entidad que contiene la información de una combinación que se ha extraído del corpus. Puede ser de alguno de los tipos: *verbo + sustantivo, verbo + adverbio, sustantivo + adjetivo*.
Atributos
id_Combinación: identificador de la combinación. Base_id_Forma: Indentificador de la base Colocativo_id_Forma: Identificador del colocativo Frec_absoluta: Frecuencia absoluta de la combinación Frec_relativa_base: Valor de la frecuencia relativa de la combinación respecto a la base Frec_relativa_colocactivo: Valor de la frecuencia relativa de la combinación respecto al colocativo Informacion_mutua: Valor de la información mutua de la combinación Dunning: Valor del test de Dunning de la combinación z_score: Valor del z-score de la combinación t_score: Valor del t-score de la combinación Ejemplo_FC_Base: Ejemplo_FC_Colocativo:

Entidad
IdeologicoVox
Descripción
Entidad que contiene la clasificación semántica de palabras del Diccionario Ideológico Vox (DIV)
Atributos
NumeroCabecera: Identificador de la Cabecera a la que pertenece el Grupo NumeroGrupo: Identificador del Grupo al que pertenece la palabra NumeroPalabra: Identificador de la palabra Palabra: Palabra en la clasificación Categoria: Categoría de la palabra

Entidad
Ideologico_VOX_Cabeceras
Descripción
Entidad que contiene la información de las cabeceras del DIV.
Atributos
Numero: Identificador de la cabecera Cabecera: Palabra que representa a la cabecera. Cuadro: Cuadro al que pertenece la cabecera, representa un nivel superior en la clasificación. CardSust: Cardinal del conjunto de sustantivos en la cabecera. CardVerbos: Cardinal del conjunto de verbos en la cabecera. CardAdj: Cardinal del conjunto de adjetivos en la cabecera. CardAdv: Cardinal del conjunto de adverbios en la cabecera

Entidad
Cuartiles
Descripción
Entidad que contiene los cuartiles y rango intercuartílico de las frecuencias relativas de cada forma del corpus para poder calcular estadísticos de valores atípicos. Pueden ser cuartiles de combinaciones *verbo + sustantivo, verbo + adverbio o sustantivo + adjetivo*.
Atributos
id_Forma: identificador de la Forma Canónica Q1: Cuartil 1 de las frecuencias relativas de las combinaciones en las que aparece la forma canónica. Q2: Cuartil 2 o mediana de las frecuencias relativas de las combinaciones en las que aparece la forma canónica. Q3: Cuartil 3 de las frecuencias relativas de las combinaciones en las que aparece la forma canónica. RI: Rango intercuartílico de las frecuencias relativas de las combinaciones en las que aparece la forma canónica.

2.4. Análisis de los resultados

Se utilizan los valores obtenidos de los indicadores para las combinaciones recopiladas y las del DCECR que aparecen en el corpus explorado para analizar la eficacia de las medidas de colocabilidad en la literatura de referencia, junto con su comparación con los datos extraídos de

la totalidad del corpus. Hay que destacar que a pesar de la extensión de éste, de las 2 356 combinaciones recopiladas se encuentran muestras de 1 832 de ellas y en cuanto a las 55 052 que provienen del DCECR se obtiene información de 28 403. Por tanto, si bien los datos registrados se consideran una amplia muestra de las capacidades de combinatoria de las palabras del español, el catálogo se considera totalmente abierto.

En los primeros epígrafes de este apartado se analizan los resultados obtenidos en general, para posteriormente realizar el estudio según la estructura de la combinación corresponda a *sustantivo* + *verbo*, *sustantivo* + *adjetivo* o *verbo* + *adverbio*.

2.4.1. Frecuencia Relativa e Información Mutua

Inicialmente se valora la posibilidad de establecer como puntos de corte para delimitar la frontera entre las colocaciones y las combinaciones libres una puntuación de 0,2 en el caso de la frecuencia relativa (Koike, 2001) y de 3 en el caso de la información mutua (Kenneth, 1990). En la Tabla 2-3 se presenta un resumen de los valores de ambos indicadores en la muestra tratada. A partir de ella se aprecia que el valor de corte en el caso de la frecuencia relativa es excesivamente restrictivo. Por otra parte, los valores máximos no necesariamente lo alcanzan las colocaciones. En este sentido, se detecta que valores de frecuencia relativa 1 corresponden con combinaciones cuya frecuencia absoluta es baja (Tabla 2-6), siendo raras las combinaciones que aparecen en el corpus con una frecuencia absoluta superior a 5 dentro de este grupo.

Fuente	Casos con Frec. Rel. ≥ 0,2	Casos con Inf. Mutua ≥ 3	Rango de la Frec. Rel.	Rango de la Inf. Mutua
Recopiladas	7%	92,79%	[0,00018196 0,8666667]	[-0,72 15,45]
DCECR.	2%	83%	[9,3194E-05 0,8833333]	[-1,25 16,50]
Corpus	0%	25,74%	[1,536E-06 2,7]	[-3,66 26,57]

Tabla 2-3. Resumen de valores de los indicadores obtenidos en el corpus

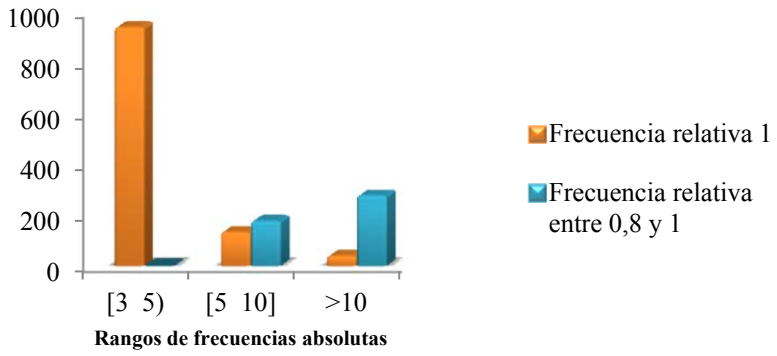


Gráfico 2-4 Distribución de las combinaciones según frecuencias totales

Los rangos encontrados para los valores de la información mutua indican que cuando se utiliza esta medida de asociación se hace necesario establecer además un límite superior. Una inspección exhaustiva de las combinaciones que alcanzan dichos valores máximos revela que en bastantes de ellas, ambas formas canónicas pertenecen al mismo campo semántico: *cambaleo-gangarilla, anabolismo-catabolismo, obrepción-subrepción* son ejemplos de ello, como se observa en las definiciones del DRAE. Este análisis manual pone de manifiesto que el límite inferior de la información mutua también debe ser modificado, ya que en caso contrario se seleccionarían como colocaciones una gran cantidad de combinaciones que no lo son.

cambaleo

m. Compañía antigua de la legua, compuesta ordinariamente de cinco hombres y una mujer que cantaba.

gangarilla

f. Compañía antigua de cómicos o representantes, compuesta de tres o cuatro hombres y un muchacho que hacía de dama.

anabolismo

m. *Biol*. Conjunto de procesos metabólicos de síntesis de moléculas complejas a partir de otras más sencillas.

catabolismo

m. *Biol*. Conjunto de procesos metabólicos de degradación de sustancias para obtener otras más simples.

obrepción

f. **Der**. Falsa narración de un hecho, que se hace al superior para sacar o conseguir de él un rescripto, empleo o dignidad, de modo que oculta el impedimento que haya para su logro.

subrepción

f. Acción oculta y a escondidas.

2. f. *Der*. Ocultación de un hecho para obtener lo que de otro modo no se conseguiría.

Cabe destacar que los valores de frecuencias relativas mayores que 1 se producen en los casos en el que en la misma línea de concordancia se repiten formas canónicas, se tienen solo 55 combinaciones de este tipo en el corpus, por lo que se opta por desecharlas del estudio. Uno de los ejemplos que mejor explica estos casos es la frase:

"*Matarile rile rile, matarile rile rón*".

La lematización produce los resultados:

Forma Canónica	*matarile*	*rilar*	*rilar*	*matarile*	*rilar*	*rón*
Categoría	Sustantivo	Verbo	Verbo	Sustantivo	Verbo	No reconocida

Tabla 2-4 Frecuencias relativas mayores que 1

Los valores de frecuencia que se recogen para las combinaciones *verbo + sustantivo* registradas en el texto del ejemplo se muestran en la Tabla 2-5.

	Frecuencia	**Frec. Relativa**
matarile + rilar	6	
A = Matarile	2	6/2 > 1
B = Rilar	3	6/3 > 1

Tabla 2-5 Frecuencias calculadas en el ejemplo

Se muestra en la Tabla 2-6 el ranking de las 20 combinaciones con mejores puntuaciones de la frecuencia relativa, en la Tabla 2-7 el de las 20 mejores cuando se usa la información mutua. Estos resultados dejan patente que las combinaciones libres se entremezclan en zonas de las que deberían quedar excluidas para hacer fiable la extracción automática por medio de estos indicadores. A la luz de estos resultados se ve la necesidad de fijar la atención también en los valores de la frecuencia absoluta de la combinación. Esto se debe a que gran parte de las combinaciones con este dato bajo y frecuencia relativa o información mutua superando los umbrales establecidos no son colocaciones. A partir del ensayo con distintas posibilidades se fija la exigencia de al menos 10 muestras de la combinación para considerar fiable la puntuación alcanzada por una combinación. En el Gráfico 2-5 se contrasta el valor mínimo de muestras requerido para considerar fiables los indicadores sobre las colocaciones recopiladas y las del DCECR.

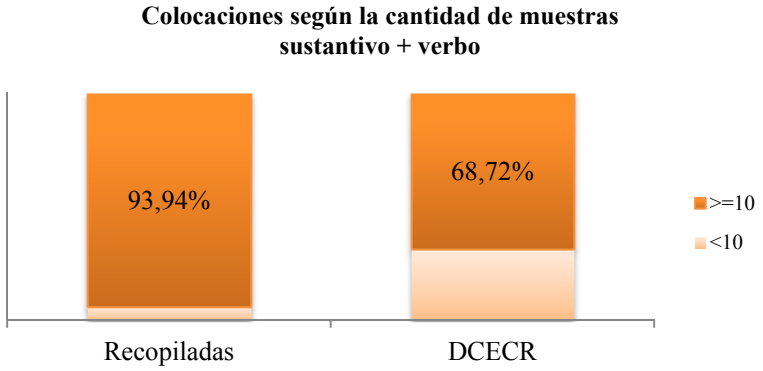


Gráfico 2-5 Colocaciones de los conjuntos de prueba que cumplen con la restricción del mínimo número de muestras

Palabra *x*	**Palabra *y***	**Frec. Relat. *x***	**Frec. Relat. *y***
recostadero	*tapiz*	1	0,0008219178
acristalamiento	*doblar*	1	6,739301E-05
antuviador	*disoluta*	1	0,003115265
antimoniuro	*india*	1	0,0001105899
guadianés	*palado*	1	0,00175644
desmelenadura	*telaraña*	1	0,001273885
ajedrista	*juan*	1	1,958608E-05
candonguero	*confirmar*	1	0,0001316713
caratulero	*morar*	1	7,478127E-05
bereque	*decir*	1	1,536072E-06
antuviador	*poner*	1	5,858757E-06
microfilmadora	*atar*	1	0,000108452
caciquesco	*hazañar*	1	0,0002938871
gomosidad	*húmedo*	1	0,0001965795
bereque	*curro*	1	0,00123406
caratulero	*infinito*	1	9,875568E-05
desalumbradamente	*decir*	1	1,536072E-06
galonista	*novatada*	1	0,06818182
impugnativo	*desahogar*	1	0,0008186655
breguero	*valeroso*	1	0,0003867474
desalinizadora	*planta*	1	0,0001832988

Tabla 2-6 Combinaciones en el corpus con valores máximos de frecuencia relativa

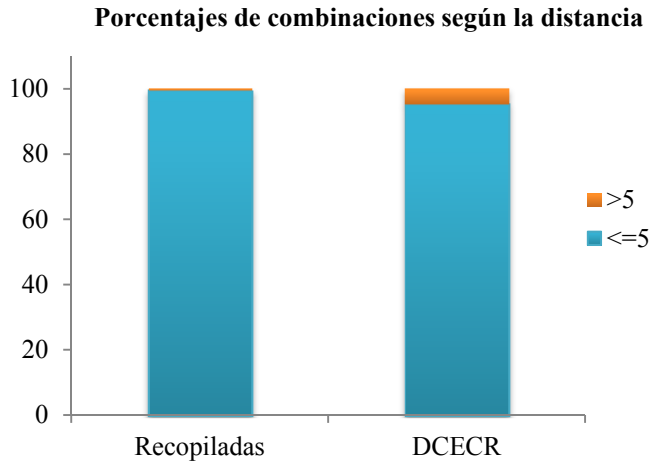


Gráfico 2-6 Distancia entre los colocados en los conjuntos de prueba

Palabra *x*	**Palabra *y***	**Inf. Mutua**
candonguero	*candonguear*	26,5754247590989
carroñoso	*lapidífico*	25,8384591649327
amentáceo	*balanífero*	25,8384591649327
formiato	*propilo*	25,8384591649327
antuviador	*piltraca*	25,8384591649327
lairén	*pedrojiménez*	25,5754247590989
hiperlipidemia	*lipemia*	25,5165310700453
espigadilla	*zollipar*	25,3530323762343
amarguero	*morrilla*	25,3530323762343
helenización	*panhelenismo*	25,1603872598201
cochinita	*pibil*	25,1603872598201
adarguero	*algareador*	25,1603872598201
mandilejo	*oracionero*	24,768069909176
cespitoso	*flexuoso*	24,7009556796546
abstergente	*restringente*	24,6160667596262
abstergente	*cefalálgico*	24,6160667596262
epigénesis	*preformismo*	24,5754247590989
norato	*redel*	24,5754247590989
rupicapra	*rupicapra*	24,5754247590989
ruado	*ruado*	24,5456775383339

Tabla 2-7 Combinaciones en el corpus con valores máximos de Información mutua

Asimismo, la revisión de los casos en que la combinación solo se da a distancias grandes revela que éstos no aportan información sobre la combinatoria de los lexemas, por lo que se desechan todas aquellas muestras de la combinación a distancias mayores que 5 (Gráfico 2-6)

En síntesis, se puede concluir que la metodología de trabajo exige fijar un número mínimo de muestras y una distancia máxima entre los elementos de la colocación. Empíricamente, se determina que si no se llega al menos a 0,0001 en la frecuencia relativa o de 1,5 en la información mutua no se deben tomar en consideración. Por otra parte, parecía suficiente indicio de colocación el disponer de al menos 10 muestras de la combinación y frecuencia relativa que supere el valor de 0,05 o la información mutua mayor que 6. Fruto de los hallazgos realizados surge la catalogación de las combinaciones extraídas en: combinaciones libres, casos dudosos y colocaciones aplicando los criterios expuestos anteriormente y que se resumen en la Tabla 2-8.

	Frec. absoluta		**Frec. Relativa**		**Inf. mutua**
Combinaciones libres			$\leq 0{,}0001$	O	$\leq 1{,}5$
Casos dudosos	< 10	O	$> 0{,}0001$ Y $< 0{,}05$	Y	$> 1{,}5$ Y < 6
Colocaciones	≥ 10	Y	$\geq 0{,}05$	O	≥ 6

Tabla 2-8. Catalogación de las combinaciones del corpus.

El grupo de combinaciones que ni se aceptan como colocaciones ni como combinaciones libres será identificado como combinaciones sin catalogar. Entre ellas se pueden encontrar combinaciones libres,

intimidar-ciclo

cobrar-madera

y colocaciones,

enternecer-caricia

acunar-sueño

estirar-cabello

Si se aplican estas reglas sobre las colocaciones con estructura *sustantivo + verbo* se observa la gran cantidad de combinaciones rechazadas como colocaciones (Gráfico 2-7) y que el conjunto de combinaciones que se identifican por ambos indicadores tiene un alto grado de coincidencia (Gráfico 2-8).

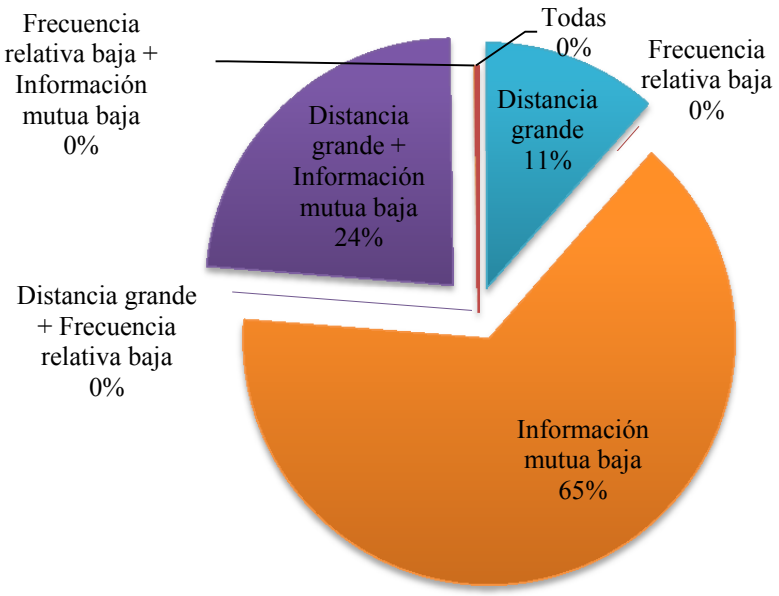


Gráfico 2-7 Combinaciones sustantivo + verbo que no cumplen ninguna de las reglas establecidas

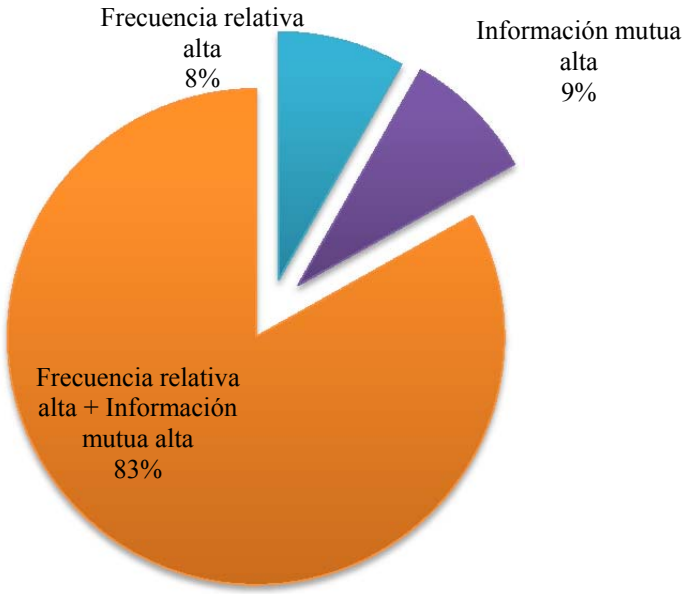


Gráfico 2-8 Combinaciones sustantivo + verbo aceptadas según las reglas establecidas.

2.4.2. Test estadísticos

Se analizan los datos obtenidos para el z-score, t-score y test de Dunning, los rangos en que varían tanto el z-score como el t-score no presentan gran diferencia entre las colocaciones recopiladas o las del DCECR. Este resultado parece útil para automatizar la catalogación de combinaciones como colocaciones si pertenecen a dicho intervalo respetando un cierto margen (Tabla 2-9).

	z-score	t-score	Dunning
Recopiladas	[-55,41 695,99]	[-1,57 127,55]	[-2293,62 1904484,55]
D. CC .RR.	[-72,64 695,99]	[-2,38 197,37]	[357,84 19664670]
CORPUS	[-172,32 5539,22]	[-28,66 304,99]	[-36533,65 19664695,48]

Tabla 2-9 Rangos de variación de los estadísticos z-score, t-score y test de Dunning

Son pocas las combinaciones cuyo valor del z-score superan el umbral marcado por los conjuntos de ensayo, en total 567. En este grupo se recogen anomalías como *matarile-rilar* o bien palabras del mismo campo semántico como *fulano-mengano, iza-rabiza, antineutrón-antiprotón* que también presentan como característica la baja frecuencia de los elementos que las componen. En este rango del z-score, al ceñirse a los casos en que las formas canónicas constituyentes tienen frecuencias altas, aparece una alta densidad de colocaciones (Tabla 2-10).

z-score	**FC *x***	**Frec *x***	**FC *y***	**Frec *y***
876,01	*administrativo*	8842	*contencioso*	1505
1189,47	*ceja*	10904	*enarcar*	1620
1160,71	*ceño*	11392	*fruncido*	3040
2388,16	*ceño*	11392	*fruncir*	14661
700,07	*distrito*	12837	*federal*	10248
796,14	*barro*	15357	*recreativo*	1639
719,03	*archivo*	15515	*protocolo*	3227
795,52	*columna*	19662	*vertebral*	965
1673,44	*estado*	40175	*unido*	42708
1226,58	*estado*	40175	*unir*	79751
711,36	*presidenta*	47988	*república*	42959
711,36	*presidente*	47988	*república*	42959
1407,71	*hombro*	61358	*encoger*	20376
770,38	*puerta*	203887	*abrir*	207390

Tabla 2-10 Combinaciones cuyos elementos aparecen con elevada frecuencia y z-score alto.

La representación de la frecuencia de la combinación frente al valor alcanzado por el z-score muestra que valores altos del indicador se dan en todos los espectros de frecuencia, tanto si ésta es alta como si no (Gráfico 2-9). Otro aspecto a destacar en el comportamiento del z-score es que los valores negativos corresponden a colocaciones funcionales (Tabla 2-11).

Recopiladas			DCECR.		
z-score	**FC A**	**FC B**	**z-score**	**FC A**	**FC B**
-27,29	*embargo*	*levantar*	-36,75	*profundo*	*saber*
-27,63	*tristeza*	*tener*	-38,17	*clase*	*dar*
-28,05	*locura*	*tener*	-38,13	*clases*	*dar*
-28,40	*reflejo*	*tener*	-38,754	*sitio*	*dar*
-28,43	*función*	*tener*	-39,15	*aire*	*dar*
-29,73	*angustia*	*tener*	-39,21	*amor*	*dar*
-30,38	*solución*	*hacer*	-40,21	*fuerza*	*dar*
-30,81	*término*	*dar*	-40,25	*imagen*	*dar*
-32,49	*sueño*	*tener*	-40,77	*vivo*	*dar*
-32,78	*idea*	*llevar*	-41,42	*colores*	*dar*
-32,89	*belleza*	*tener*	-41,59	*color*	*dar*
-35,00	*pasión*	*tener*	-42,92	*voz*	*dar*
-35,68	*terror*	*tener*	-43,00	*espacio*	*dar*
-35,76	*fin*	*dar*	-43,29	*atenciones*	*dar*
-37,36	*metro*	*ir*	-43,29	*atención*	*dar*
-41,35	*problema*	*señor*	-49,53	*pie*	*dar*
-41,35	*problema*	*señor*	-51,97	*guerra*	*dar*
-42,97	*relación*	*hacer*	-53,65	*forma*	*dar*
-46,09	*bajo*	*hombre*	-58,67	*cuerpo*	*dar*
-52,17	*causa*	*tener*	-64,27	*alto*	*dar*
-55,41	*amor*	*tener*	-68,23	*ver*	*dar*
			-72,62	*tiempo*	*dar*

Tabla 2-11 Combinaciones con z-score negativo

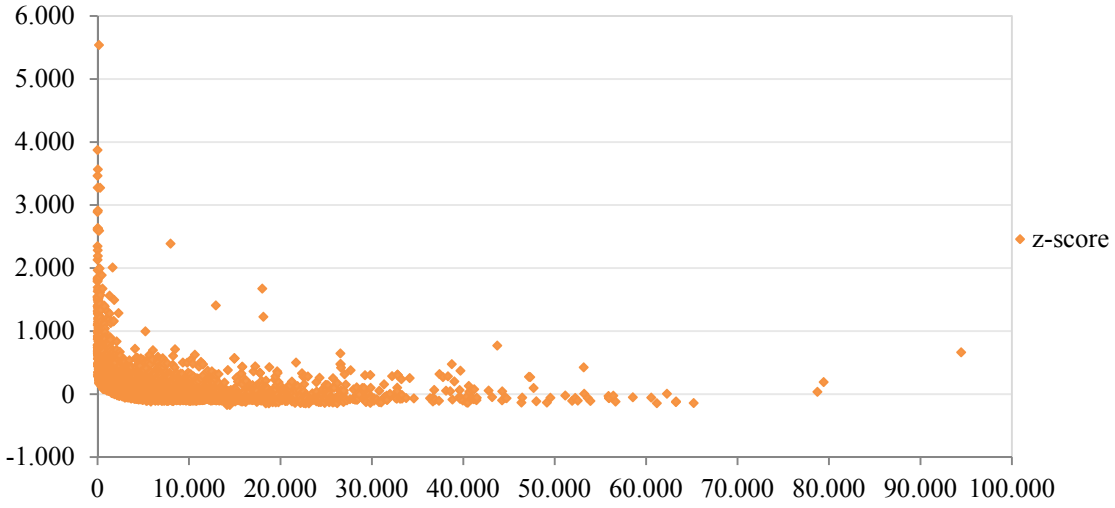


Gráfico 2-9 Valores del z-score según la frecuencia de la combinación

En contraste con los resultados observados para el *z-score* ocurre que los valores negativos del *t-score* se pueden considerar prácticamente inexistentes entre las colocaciones recopiladas y las del DCECR: 2 y 93 respectivamente. En el total del corpus se observa que estos casos se obtienen en combinaciones de frecuencias mínimas. También destaca la dependencia directa entre el valor del *t-score* y la cantidad de veces que aparece la combinación, a mayor frecuencia mayor *t-score*. Por tanto, no se recomienda establecer un ranking en el que se entremezclen palabras poco usadas con aquellas de frecuencia de aparición alta. En la Tabla 2-12 se resalta, a modo de ejemplo, una combinación libre que queda igual o mejor catalogada que colocaciones cuya frecuencia de aparición es igual o menor a la de ella. Este error está inducido por utilizar un único ranking entre todas las combinaciones en un texto o corpus dado, basado en este indicador.

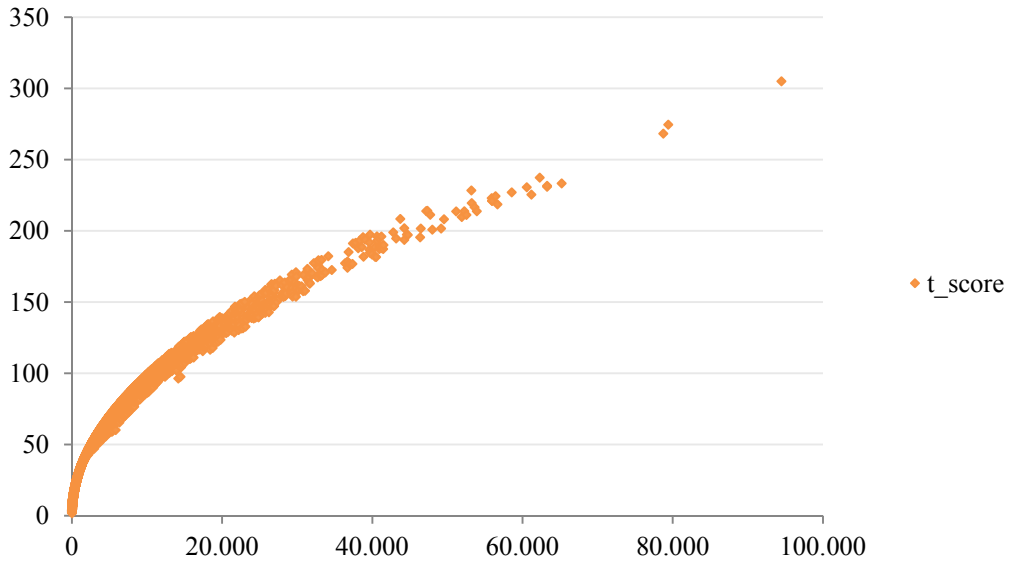


Gráfico 2-10 Valores máximos de t-score entre las combinaciones agrupadas por frecuencia de aparición.

Pal *x*	Pal *y*	Frec (*x,y*)	*t-score*
dar	*discípulo*	681	22,8
fuerte	*emoción*	426	20,29
dar	*emoción*	681	23,86
controlar	*emoción*	165	12,73
dominar	*emoción*	390	19,6
prestar	*colaboración*	143	11,86

Tabla 2-12 Influencia de la frecuencia de la combinación en el t-score

Respecto al *test de Dunnig*, se atenúa la influencia de la cantidad de muestras de una combinación en la puntuación que se obtiene. Este resultado es de esperar, ya que está diseñado para hacerlo más estable ante las variaciones en la cantidad de muestras. En la Tabla 2-13 se observan las colocaciones recopiladas para el verbo *ir,* destacando el amplio rango de variación de las frecuencias computadas para cada una de ellas, y la baja variabilidad del *test de Dunning*.

FC *x*	FC y	Frec(x, y)	*t-Dunning*
coche	*ir*	3088	11394028,69
metro	*ir*	669	11432604,93
automóvil	*ir*	250	11441127,50
automóvil	*ir*	250	11441127,50
taxi	*ir*	249	11441428,53
autobús	*ir*	223	11441910,29
detrimento	*ir*	60	11444571,56
descrédito	*ir*	7	11445482,55
desdoro	*ir*	7	11445487,50

Tabla 2-13 Test de Dunning evaluado en colocaciones recopiladas para el verbo ir

Para finalizar, se aprecia inviable la evaluación del *test de Poisson* en un corpus tan extenso. Se recuerda que la expresión incluye valores de *k!,* siendo *k* la frecuencia de la combinación.

2.4.3. Análisis por tipos de combinación según la categoría gramatical

Un segundo análisis de los resultados obtenidos se llevó a cabo según el tipo de combinación. Los grupos se han establecido según la categoría gramatical de las formas canónicas de los elementos que intervienen en las mismas, teniéndose en cuenta los casos:

sustantivo + verbo

sustantivo + adjetivo

verbo + adverbio.

2.4.3.1. Combinaciones sustantivo + verbo

De ellas se reconocen aproximadamente un 55% de las colocaciones recopiladas que se encuentran en el corpus y un 23 % de las que se encuentran en el DCECR. De las colocaciones recopiladas ninguna infringe el requisito de la frecuencia relativa, sin embargo 11 solo aparecen a una distancia de más de 5 palabras, 18 no superan el umbral de la información mutua y 153 no cumplen el requisito de la frecuencia absoluta mínima. En cuanto a las combinaciones del DCECR., 4 se descartarían por no llegar a la frecuencia relativa mínima, 1065 por no superar el valor impuesto de 1,5 a la información mutua, 1332 casos solo aparecen a distancia alta y 9995 no cumplen el criterio de contar con al menos 10 muestras de las mismas en el corpus (Tabla 2-14).

	Distancia > 5	**Frec. Relativa < 0,05**	**Inf. Mutua < 1,5**	**Frec. Total ≤ 10**
Recopiladas	11	-	18	153
DCECR.	1 332	4	1 065	9 995

Tabla 2-14 Colocaciones que infringen las restricciones determinadas en los conjuntos de prueba

Se comprueban los rangos de variación de *z-score*, *t-score* y *test de Dunning* evaluados para cada uno de los grupos de ensayo; en todos los casos, de uno u otro modo, se observa la influencia de la frecuencia de uso de los elementos de la combinación con la puntuación obtenida: valores mínimos de *z-score* para verbos como *hacer, tener, formar, decir,* etc. o valores máximos de *t-score* cuando ambos elementos de la combinación tienen una frecuencia alta de aparición (Tabla 2-15), o el caso del *test de Dunning*, donde ocurre que los puestos más altos en el ranking corresponden a las llamadas colocaciones funcionales producidas por verbos de elevada frecuencia de uso:

hacer, dar, tener, ir, poner, haber, ... + sustantivo

	z-score	***t-score***	***Dunning***
Recopiladas	[-55,41 695,99]	[-0,66 127,55]	[83,53 1904484]
D. CC .RR.	[-72,64 695,99]	[-2,38 197,37]	[1712,65 19664670]
Corpus	[-172,32 5539,22]	[-20,31 268,26]	[83,53 19044484,55]

Tabla 2-15 rangos de variación de los estadísticos en los conjuntos de prueba y en el corpus

Se puede considerar que los criterios establecidos en el caso de las frecuencias relativas y la información mutua para descartar combinaciones son efectivos; sin embargo, la cobertura del método en cuanto a la catalogación de las colocaciones resulta insuficiente. Se concluye además que cuando se usa cualquiera de los indicadores estudiados no parece muy útil la comparación mediante rankings, o valores de corte sobre el total del corpus, ya que su comportamiento queda determinado por la cantidad de muestras de la combinación y por la frecuencia de aparición de los elementos que la componen, es decir, el uso que se le da a una palabra. No se puede comparar el valor de un indicador basado en la frecuencia entre los verbos *dar* (1 118 012), *desempeñar* (18 903), o *traspirar* (100).

***z-score*, valores máximos**		***z-score***	
Frecuencia Combinación >=10		**valores mínimos**	
aranzada majolar	*nobel premiar*	*capitán poder*	*antiguo poder*
ruga rugar	*monóxido carbonar*	*vuelto poder*	*amor hacer*
puerta cerrar	*pistilo estambrar*	*negro hacer*	*ante tener*
cloruro clorurar	*rodilla hincar*	*bienes seguir*	*llama poder*
vito vitar	*dactilar hollar*	*bien seguir*	*siglo tener*
cigarro fumar	*viento soplar*	*alrededores decir*	*hombre formar*
i her	*café tazar*	*alrededor decir*	*año parecer*
mamante piantar	*salce pitar*	*alto querer*	*propio dar*
holístico agenciar	*diamela engraciar*	*artículo saber*	*negro poder*
hidrato carbonar	*atenciones prestar*	*mesa decir*	*junto saber*
sant sacramentar	*atención prestar*	*gobierno ver*	*voz ir*
vistazo echar	*iris arcar*	*español hacer*	*colores poder*
		vida llegar	*capitana poder*

Tabla 2-16 combinaciones del corpus con valores extremos del z-score

2.4.4. Combinaciones sustantivo + adjetivo

En este grupo se incluyó el total de combinaciones del tipo *sustantivo – adjetivo* o *sustantivo – sustantivo (que puede usarse como adjetivo en el corpus)*. El resumen de los datos registrados revela ligeras diferencias con las correspondientes a la estructura *sustantivo + verbo.* Se destaca que la distancia colocacional no debe restringirse a 5, las frecuencias relativas alcanzan valores mayores cuando se evalúan respecto al adjetivo en lugar del sustantivo (Gráfico 2-11). Entre las coincidencias en el comportamiento se observa la necesidad de exigir un número mínimo de

muestras, el que las combinaciones con frecuencia relativa 1 se deben descartar, el valor del *t -score* está directamente relacionado con la cantidad de veces que aparece la combinación en el corpus, y que el rango de variación de la información mutua es similar entre las colocaciones recopiladas y las del DCECR.

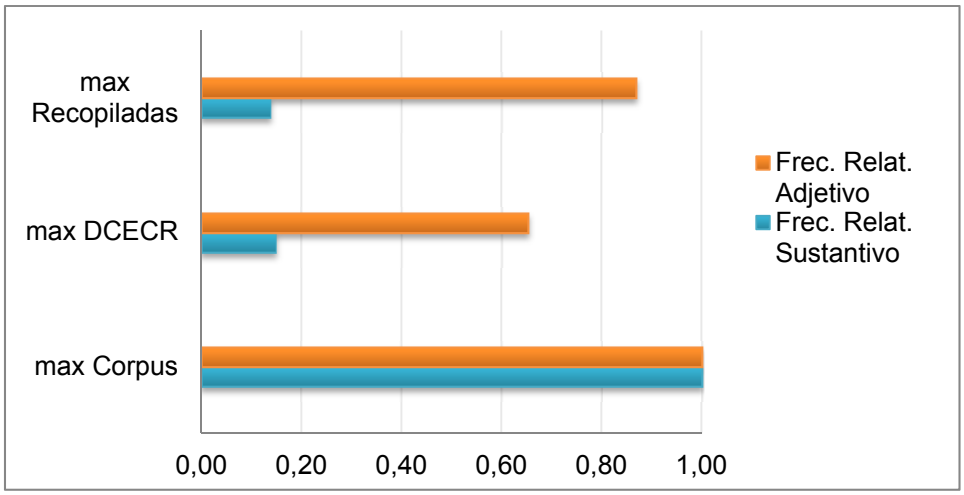


Gráfico 2-11 Frecuencias relativas en las combinaciones Sustantivo + Adjetivo

	Corpus	**Recopiladas**	**DCECR.**
Frec. Total ≥ 10	41,34%	87,56%	57,58%
Frec. Relat. Sust.	[0,000014 1]	[0,000012 0,13826]	[0,0000051 0,14894]
Frec. Relat. Adj.	[0,000005 1]	[0,00015 0,86667]	[0,000014 0,65165]
Inf. Mutua	[-3,41759 25,83846]	[-0,71624 15,45569]	[-0,73025 16,50326]
z-score	[-104,74 2634,92]	[-46,1 607,86]	[-17,32 335,7]
t-score	[-16,78 191,71]	[1,57 100,58]	[-1,61 40,41]
Dunning	[-19846,94 971272,33]	[-2293,62 6740597,91]	[357,85 5320378,84]

Tabla 2-17 Valores obtenidos por las combinaciones Sustantivo + Adjetivo.

En el caso del t-score se entremezclan numerosas combinaciones libres que obtienen buenas puntuaciones por ser palabras de uso frecuente en lengua española:

madre padre, hijo parte, don señor, aire bueno

2.4.5. Combinaciones verbo + adverbio

No se encuentra ningún caso del conjunto de colocaciones recopiladas en este grupo, por lo que el contraste se realiza entre las 2 148 extraídas del DCECR., y las 143 578 encontradas en el corpus. Los rangos de variación de los indicadores presentan comportamientos similares al caso *sustantivo + adjetivo*. Se observa que las frecuencias relativas son mayores cuando se evalúan respecto al adverbio en lugar del verbo, se debe exigir un número de muestras mínimo y se aconseja utilizar los rangos obtenidos en el conjunto de pruebas añadiéndoles un cierto margen. Destaca el caso del test de Dunning, donde se aprecia la influencia de la frecuencia de aparición del verbo en la puntuación obtenida: las combinaciones mejor posicionadas corresponden al verbo *decir*, las siguientes al verbo *hacer*, etc. Este resultado nuevamente revela cierta dependencia de los valores calculados con la cantidad de muestras de la forma canónica en el corpus.

	Corpus	**D. CC.RR.**
Frec. Total ≥ 10	37,78%	55,87%
Frec. Relat. Adv.	[5,907373E-05 1]	[0,00018 0,78588]
Frec. Relat. Verbo	[1,536072E-06 0,3571429]	[2,987902E-06 0,08926]
Inf. Mutua	[-2,00833 21,83846]	[0,39011 13,18894]
z-score	[-48,86878 749,99601]	[-17,71749 265,77742]
t-score	[-5,23625 96,45938]	[0,47386 42,63691]
Dunning	[228,8713 19664695,48157]	[8348,81265 19664515,42659]

Tabla 2-18 Valores obtenidos por las combinaciones verbo + adverbio

Frec. Rel	*mirar fijamente, oponer diametralmente, cerrar herméticamente, mirar fijamente, enamorar perdidamente, mirar interrogativamente, poder lícitamente, vivir maritalmente, abrir desmesuradamente, vivir disolutamente, consumir improductivamente, recibir hospitalariamente, girar preceptivamente, realizar preceptivamente, disponer preceptivamente, asignar preceptivamente, llorar desconsoladamente, pecar venialmente, llorar inconsolablemente*
Inf. Mutua	*enjaezar vistosamente, legalizar solidariamente, enjaezar ricamente, amueblar vistosamente, guarnir ricamente, emanar patrióticamente, asignar preceptivamente, absolver sacramentalmente, departir amigablemente, encuadernar lujosamente, enjaezar lujosamente, infringir civilmente, lisar llanamente, amueblar suntuosamente, reclinar muellemente*
z-score	*mirar fijamente, lisar llanamente, guarnir ricamente, oponer diametralmente, enamorar perdidamente, enjaezar ricamente, amueblar lujosamente, cerrar herméticamente, iluminar brillantemente, participar activamente, ataviar ricamente, emanar patrióticamente, ligar íntimamente, sudar copiosamente*
t-score	*mirar fijamente, poder fácilmente, poder solamente, poder realmente, ver claramente, decir realmente, decir solamente, tener solamente, tener realmente, hacer solamente, hacer realmente, decir propiamente, decir exactamente, saber exactamente, hacer exactamente, decir finalmente, saber perfectamente,*
Dunning	*decir macarrónicamente, decir rasamente, decir chuscamente, decir aficionadamente, …*

Tabla 2-19 Combinaciones Verbo + Adverbio mejor puntuadas por cada indicador

2.5. Influencia del tamaño del corpus

A la vista de la dependencia que muestran los resultados con respecto al uso que se hace de las palabras en el corpus, es necesario comprobar en qué medida la estrategia queda determinada por el tamaño del corpus. Para ello, se evalúan los datos utilizando un corpus sensiblemente menos extenso que compila una colección de novelas de D. Benito Pérez Galdós, con un total de 2 299 920 palabras.

Los rangos de variación de todos los indicadores se modifican sustancialmente, comprobándose que los valores umbral se deben alterar en cuanto se cambia el tamaño del corpus. Se manifiesta de nuevo la influencia de la cantidad de muestras que es posible recolectar de una combinación sobre los valores que marcan el límite entre lo que es o no colocación. Sería necesario revisar los puntos de corte, incluido el del número mínimo de muestras que se considere fiable. Una forma de resolverlo consiste en la revisión de los datos que se obtengan y establecerlo de forma empírica, al igual que se hizo en el epígrafe anterior. Otro método es recurrir a un conjunto de entrenamiento, en tal caso las colocaciones recopiladas o el DCECR,

con el inconveniente de la dependencia de disponer de ellas y de que el corpus que se procese albergue suficientes muestras de las mismas (Gráfico 2-12).

	Frec. Relativa	***Inf. Mutua***	***z-score***	***t-score***	***Dunning***
Recopiladas	[0,0049 0,7846]	[0,28 13,13]	[-57,07 13,96]	[-46,72 5,93]	[-26,71 46634]
D. CC .RR.	[0,0028 1]	[0,5 13,26]	[-61,11 10,98]	[-39,79 14,54]	
Corpus	[0,00092 1]	[-0,99 18,19]	[-136 76,48]	[-116,19 14,54]	

Tabla 2-20 Rangos de variación de los indicadores en el corpus de Galdós

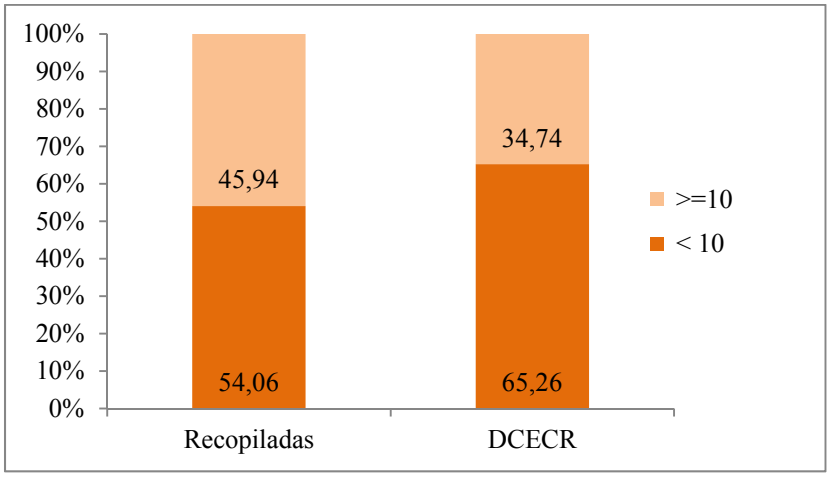


Gráfico 2-12 Muestras de las colocaciones en los grupos de ensayo en el corpus de Galdós

2.6. Las colocaciones como combinaciones atípicas

A la vista de los resultados obtenidos se hace obvia la necesidad de modificar la estrategia basada únicamente en umbrales de corte para delimitar la frontera entre colocaciones y combinaciones libres. El objetivo último es obtener una herramienta que permita automatizar en lo posible la extracción de las colocaciones presentes en un corpus con independencia de su tamaño. Para ello, se tendrá en cuenta la característica de preferencia presente en el concepto de colocación, según la cual parece lógico que, fijado un colocado, el análisis de sus capacidades combinatorias se realice entre las muestras obtenidas de él, y no sobre el total de combinaciones extraídas del corpus. De esta forma se evita la comparación entre palabras distintas desde el punto de vista de sus frecuencias de uso, cuyas diferencias aumentan a medida que el corpus sea más extenso.

Bajo estas hipótesis, se pretende construir un indicador que detecte aquellas combinaciones cuyo uso destaque con respecto a lo que es habitual, en contraposición a la idea subyacente en los estadísticos analizados en las secciones previas, en las que la decisión se toma en términos de rechazar la independencia en el uso de los colocados. En este sentido, se considera que las frecuencias relativas serán fundamentales, ya que aportan información del uso de una determinada combinación con respecto a lo que ha sido utilizada una determinada

palabra en el corpus sobre el que se trabaje. Así las cosas, la hipótesis se centra en qué valores atípicos de ésta, en relación a lo que es común en su rango de variación, proporcionarán indicios del fenómeno de la preferencia. El uso de este concepto permite contrastar los datos en términos relativos al uso que se da en el corpus a una palabra concreta, con mayor independencia respecto a cuánto haya sido usada en el corpus de referencia.

Siguiendo esta metodología se asume que cada palabra tiene su propia muestra, y se considera qué casos son atípicos dentro de ella, frente a considerar el corpus como una única muestra en la que se detecta dónde se rechaza la independencia de los colocados. Esto, además, permite alejarse del problema de establecer valores de corte dependientes del tamaño del corpus o de establecer rankings en los que se comparan datos que dependen directamente de lo frecuente que es el uso de las palabras implicadas en la combinación.

Este epígrafe se centra a partir de ahora en la experimentación con técnicas que tradicionalmente se usan para la identificación de valores atípicos en muestras. En estadística se utiliza el término *outlier* para referirse a los datos de una muestra que parecen ser inconsistentes con el resto del conjunto, es decir, son valores que parecen demasiado grandes o demasiado pequeños en comparación con las demás observaciones.

> Un outlier es una observación que se desvía tanto de las restantes como para sospechar que fue generada por un mecanismo diferente. (Aggarwal, 2013).

La hipótesis que se plantea consiste en considerar las colocaciones como anomalías en el mecanismo de generación de combinaciones, se presupone que el fenómeno de generación origina combinaciones libres y que la característica de la preferencia determina la diferencia en la producción de las colocaciones.

Se analizan 3 estrategias para la detección de estos elementos singulares ampliamente utilizadas en otros ámbitos: método basado en la *desigualdad de Chebyshev*, método de los *cuartiles* y por último el método basado en la *MEDA*. En el primero se determinan los outliers a partir de la media muestral de las frecuencias mientras que en los otros dos se hace a partir de los valores de los cuartiles de la muestra. Cabe destacar que todos ellos proporcionan un criterio objetivo para establecer un punto de corte respecto a las combinaciones que se deben considerar, frente a las técnicas propuestas anteriormente, en la que la frontera se establece de forma aproximada y es totalmente dependiente del tamaño del corpus.

2.6.1. Método basado en la desigualdad de Chebyshev

La desigualdad de Chebyshev garantiza que los datos de una muestra verifican:

$$fr(i: |x_i - \bar{x}| > ks) < \frac{1}{k^2}$$

Siendo $\bar{x}$ la media muestral y s la desviación típica de la muestra. Aplicado a distribuciones normales, significa que se cumplen las relaciones de la Tabla 2-21

% de datos que verifican	
68%	$fr(i: \|x_i - \bar{x}\| < s)$
95%	$fr(i: \|x_i - \bar{x}\| < 2s)$
99%	$fr(i: \|x_i - \bar{x}\| < 3s)$

Tabla 2-21 Aplicación de la desigualdad de Chebyshev

Sobre la base de estas propiedades se justifica considerar atípicas las observaciones que verifican: $|Z| > 3$, siendo:

$$Z = \frac{x_i - \bar{x}}{s}$$

De este modo se estarían limitando las anomalía al 0,13% de los casos (Leys *et al.*, 2013), algunos investigadores adoptan un criterio menos restrictivos, considerando una desviación de la media de 2 o 2,5 veces la desviación de la media.

En este trabajo: fijada una palabra, los valores x_i corresponden a las frecuencias relativas con las restante con las que combina y $\bar{x}$, s a la media muestral y desviación típica de dichos datos respectivamente, por lo que el estadístico que se usa:

$$ZChebyshev(X_i, Palabra) = \left| \frac{FrecRel(X_i, Palabra) - \overline{FrecRel(Palabra)}}{s_{FrecRel(Palabra)}} \right|$$

Las combinaciones $(X_i, Palabra)$ tales que $ZChebyshev(X_i, Palabra) > 3$ serán aceptados como colocaciones.

La Tabla 2-22 y la Tabla 2-23 resumen los resultados de aplicar esta estrategia sobre las combinaciones en el corpus completo y el corpus de Galdós, mostrando que el requisito que se exigió a las combinaciones para considerarlas colocaciones fue excesivamente restrictivo. No obstante, se aprecia que se filtran notablemente la gran cantidad de combinaciones libres que se introducen erróneamente en los catálogos generados por los restantes métodos. Se muestran,

como ejemplos, combinaciones que superan la restricción de Chebyshev en los 3 conjuntos de ensayo. Se presentan tanto las de mejor como las de peor puntuación (Tabla 2-24).

	Rango	**Combinaciones con *ZChebyshev* ≥ 3**
Recopiladas	[0 22,289966]	933
DCECR	[0 78,607647]	2735
Corpus	[0 78,607647]	455738

Tabla 2-22 Resultados de ZChebyshev en el corpus completo

	Rango	**Combinaciones con *ZChebyshev* ≥ 3**
Recopiladas	[0 4,35341]	45
DCECR	[0 11,650217]	83
Corpus	[0 16,537422]	2922

Tabla 2-23 Resultados de ZChebyshev en el corpus de Galdós

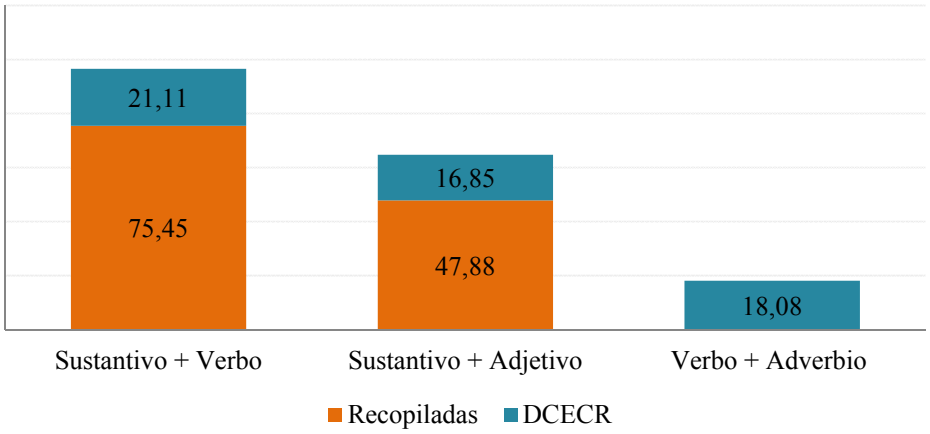


Gráfico 2-13 Combinaciones de los conjuntos de ensayo que superan el criterio de ZChebyshev (Z-Chebyshev≥3)

	Recopiladas		DCECR		Corpus	
***ZChebyshev* ≥ 3 Max**	*dar*	*golpe*	*dar*	*vuelta*	*dar*	*vuelta*
	hacer	*pregunta*	*dar*	*paso*	*don*	*doña*
	poner	*marcha*	*dar*	*gracia*	*estado*	*unido*
	tener	*intención*	*tomar*	*decision*	*tener*	*razón*
	tomar	*decision*	*dar*	*salto*	*hacer*	*falta*
	dar	*salto*	*dar*	*orden*	*tener*	*miedo*
	latir	*corazón*	*soplar*	*viento*	*decir*	*verdad*
***ZChebyshev* ≥ 3 Min**	*alisar*	*cabello*	*sumir*	*cuidado*	*cavado*	*fosa*
	alisar	*pelo*	*reponer*	*fuerza*	*alternativo*	*formar*
	salir	*crisis*	*camino*	*accidentado*	*llamada*	*recibir*
	viajar	*coche*	*perder*	*control*	*dar*	*evento*
	dar	*taconazo*	*poner*	*objection*	*día*	*evento*
	haber	*relación*	*dar*	*ubicación*	*torre*	*elevado*
	poner	*crisis*	*reinar*	*tranquilidad*	*enemigo*	*padre*

Tabla 2-24 Ejemplos de combinaciones que superan el criterio de Chebyshev

La Tabla 2-25 y el Gráfico 2-13 recogen el resumen de los datos según la estructura de la colocación, se muestran los rangos alcanzados y los porcentajes de combinaciones que serían reconocidas según este criterio. Se aprecia escasa cobertura entre las combinaciones del DCECR. Sin embargo, los listados que se obtienen se consideran mejores desde el punto de vista de la precisión, logrando filtrar la gran cantidad de ruido en el proceso automatizado.

	ZChebyshev
Verbo + Sustantivo	[0 78,6076]
Sustantivo + Adjetivo	[0 68,8782]
Verbo + Adverbio	[0 41,06655]

Tabla 2-25 Rango de ZChebyshev según la estructura de la combinación

2.6.2. Métodos robustos

Una fuente de error en los métodos de identificación de outliers la constituye precisamente el uso de los datos anómalos en el cálculo de la media aritmética y de la desviación típica usados para evaluar los estadísticos. El valor de la media y la desviación típica muestral están fuertemente influenciados por la presencia de outliers en la muestra, además la desigualdad de Chebyshev se cumple para distribuciones normales. El siguiente ejemplo ilustra estos problemas. Sea la muestra en (1), en la que el valor 1000 claramente es atípico, no sería

detectado mediante el indicador *ZChebyshev* como tal debido a la distorsión que él mismo provoca en los estimadores muestrales que se usan en su cálculo. Este ejemplo (Leys *et al.*, 2013) ilustra los problemas que se originan al considerar este indicador, aunque su uso es habitual.

$$\begin{gathered}\{1,3,3,6,8,10,10,1000\}\\ \bar{x}=130{,}13\\ s=328.8\\ \bar{x}+3s=1116{,}52\\ \bar{x}-3s=-856.27\end{gathered}\qquad(1)$$

La mediana, sin embargo, es un estimador para la medida de tendencia central que es invariante a la presencia de outliers, lo que sugiere que se utilicen indicadores basados en este estadístico para resolver el problema que estamos tratando. En este epígrafe se presenta la adaptación de dos métodos de detección de outliers a la extracción automática de colocaciones, ambos basados en los cuartiles de la muestra.

2.6.3. Método basado en el diagrama de cajas y bigotes

En el análisis exploratorio de datos estadísticos se utilizan cálculos sencillos para resumir los datos. El **diagrama de cajas y bigotes de Tukey**, o **Box-Plot** es una herramienta para la visualización de 5 valores que aportan información sobre la tendencia, la simetría y la variación de la distribución del conjunto. Además permite localizar la existencia de outliers. En un diagrama de cajas y bigotes se representan $Q_1, Q_2 \text{y } Q_3$, cuartiles 1, 2 y 3 de la distribución y se utilizan para delimitar la caja. Los bigotes se marcan en los puntos $Q_1 - 1{,}5 \cdot RI \text{ y } Q_3 + 1{,}5 \cdot RI$. Estos últimos fijan la zona a partir de la que se considerarán *valores atípicos moderados* en la distribución. Finalmente, destacan los *valores atípicos extremos*, que rebasan las fronteras del intervalo $Q_1 - 3 \cdot RI$ y $Q_3 + 3 \cdot RI$. Los valores en esta zona se marcan con un símbolo que los identifica como tales, un ejemplo puede verse en la Figura 2-8. En el interior de la caja se representa el valor de la mediana, si éste no aparece centrado la distribución de la variable es asimétrica.

En la Figura 2-9 se muestra el diagrama de cajas de las frecuencias relativas del verbo *palpitar* con los sustantivos en su contexto en el corpus, la mayor parte de combinaciones se concentran en la zona de frecuencia relativa inferior a 0,002 apreciándose un conjunto de valores atípicos extremos. El perfil de la representación en un diagrama de cajas y bigotes del resumen de la variable "*frecuencia relativa fijada una palabra*" (*FrecRel(Palabra)*) por lo general será similar al que se muestra en la Figura 2-9. Se observa que corresponde a una variable totalmente asimétrica, y en la que aparecen valores atípicos extremos, por lo que se justifica la propuesta de utilizar indicadores más robustos frente a outliers que *ZChebyshev*.

media	desv. típica	Q1	Q2	Q3	n
0.0112348	0.02690609	0.003634361	0.005947137	0.009581498	167

Tabla 2-26 Estadísticos resumen de las frecuencias relativas cuando se fija el verbo palpitar.

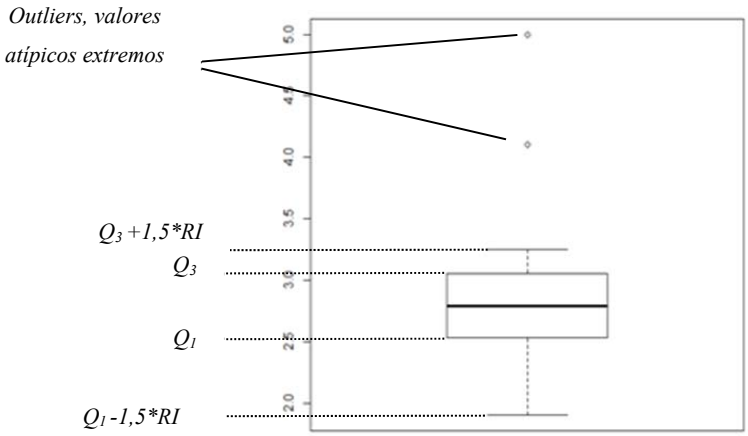


Figura 2-8 Explicación del diagrama de cajas de Tukey

Se define a continuación otro indicador basado en estas consideraciones, que identificaremos como *Box-plot(palabra, x_i)*. Este método también se encuentra en la literatura con el nombre *método de los cuartiles*. Para calcularlo, se debe fijar de antemano sobre qué palabra se calculará, y habrá que determinar el primer y tercer cuartil (*Q1, Q3*) de las frecuencias relativas respecto a la palabra de interés así como el rango intercuartílico (*RI*).

$$Box - plot(palabra, x_i) = \frac{FrecRelativa(x_i, palabra) - Q3}{RI}$$

$FrecRelativa(x_i, palabra)$ es la frecuencia relativa respecto a *palabra* de x_i con *palabra*

En la Tabla 2-27, Tabla 2-28 y la Tabla 2-29 se presentan las combinaciones *sustantivo + verbo*, *sustantivo + adjetivo* y *verbo + adverbio* que alcanzan los valores máximos que se obtienen al calcular este indicador fijando el colocativo en cada uno de esos 3 casos respectivamente. La cobertura que se alcanza se recoge en el Gráfico 2-14, donde se puede apreciar cómo suben los resultados frente al método de Chebyshev.

palabra	x_i	*FrecRelativa(palabra, x_i)*	*Box-plot (palabra, x_i)*
encoger	*hombro*	0,6348646	993,846208698718
trizar	*hecho*	0,3452381	688,999867633008
enguantar	*mano*	0,7609428	673,999826155181
abrir	*puerta*	0,210825	671,584602763095
enguantar	*manos*	0,7485971	662,999835490524
cerrar	*puerta*	0,2387841	576,282633476278
fruncir	*ceño*	0,5443012	530,800016942488
menear	*cabeza*	0,617914	481,083363372164
atrancar	*puerta*	0,6423659	462,999946849891
decir	*señora*	0,04031934	456,767419796828
decir	*señor*	0,04031934	456,767419796828
pronunciar	*palabra*	0,3318395	434,450978686954
vezar	*mucho*	0,1426878	431,531244894402
cerrar	*ojo*	0,1766763	426,108704142493
latir	*corazón*	0,3690867	412,812470856838
unir	*estado*	0,2274454	393,239122510575
mirar	*ojo*	0,07979666	382,259992494583
decir	*hombre*	0,03339114	378,098820930561
pablar	*san*	0,3634326	374,421032936214
roncar	*voz*	0,4186728	362,363633128052

Tabla 2-27 Combinaciones Sustantivo + Verbo con máximo valor de Box-plot

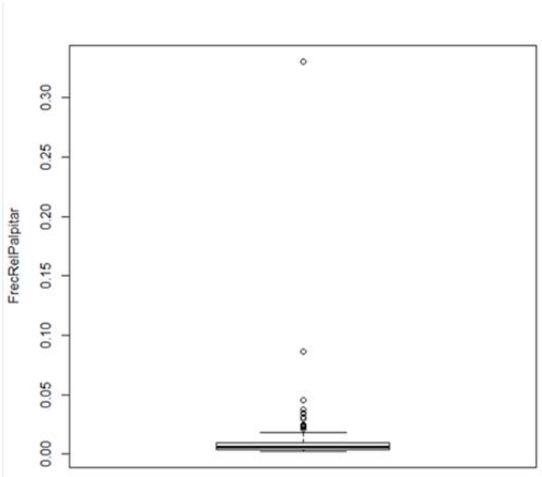


Figura 2-9 Diagrama de cajas para la distribución de las frecuencias relativas del verbo palpitar

palabra	x_i	***FrecRelativa(palabra, x_i)***	***Box-plot(palabra, x_i)***
pecado	*venial*	0,6516464	185,499984013151
prueba	*fehaciente*	0,278481	149,999992789701
fuerza	*centrífugo*	0,5914894	137,000007441268
nota	*discordante*	0,1366623	100,000013820363
camino	*trillado*	0,3438228	96,3333324454725
testigo	*ocular*	0,3409549	90,6363666121319
mirada	*penetrante*	0,1941656	87,214282613629
precio	*irrisorio*	0,094	86,999982304874
necesidad	*imperioso*	0,2272065	85,3333331557094
voz	*estridente*	0,2045455	84,7142806504476
ojo	*penetrante*	0,1874216	84,1428532289289
edad	*avanzado*	0,1275232	82,1249971597993
edad	*avanzada*	0,1275232	82,1249971597993
movimiento	*brusco*	0,1216772	79,6315753950049
documento	*fehaciente*	0,1500904	78,9999953648077
pregunta	*formular*	0,2516515	78,6999982346979
discusión	*acalorado*	0,1558185	76,9999971669168
libertad	*condicional*	0,1933497	76,0000066170476
luz	*cegador*	0,3292156	74,7142820524585
luces	*cegador*	0,3292156	74,7142820524585

Tabla 2-28 Combinaciones Sustantivo + Adjetivo con máximo valor de Box-plot

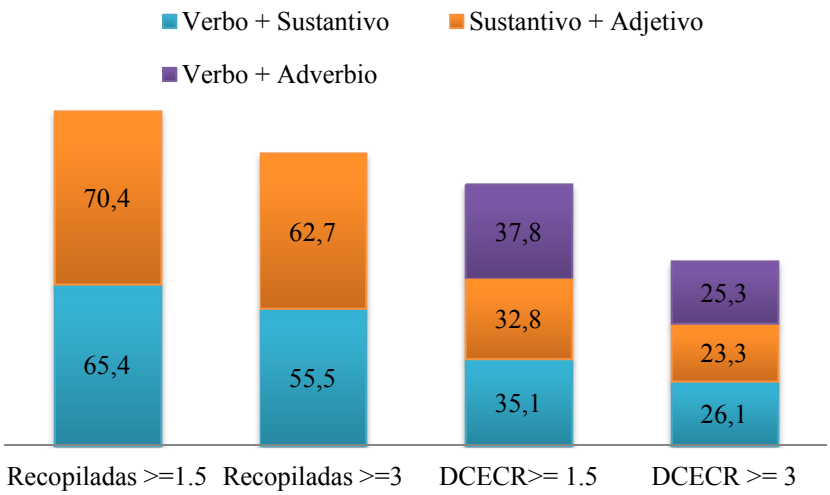


Gráfico 2-14 Cobertura en el caso de aplicar Box-Plot

palabra	x_i	***FrecRelativa(palabra, x_i)***	***Box-plot(palabra, x_i)***
herméticamente	*cerrar*	0,7858796	153,805255159383
comercialmente	*estrenar*	0,08759124	144,595113843274
olímpicamente	*ignorar*	0,3606557	139,279349376819
alfabéticamente	*ordenar*	0,508772	118,27913930758
celosamente	*guardar*	0,3695652	106,073907808064
rematadamente	*locar*	0,2102564	105,617925682556
activamente	*participar*	0,2290419	102,565217422258
frontalmente	*chocar*	0,2113821	100,519609091312
herméticamente	*sellar*	0,05555556	93,2356047672532
gratamente	*sorprender*	0,3082192	82,7287324380578
encarecidamente	*rogar*	0,2931243	81,9901037563911
ardientemente	*desear*	0,4692623	78,9741632248684
descaradamente	*mentir*	0,08625731	73,420234354581
terminantemente	*prohibir*	0,25459	65,3472818468441
rotundamente	*negar*	0,3542977	63,152643313143
acaloradamente	*discutir*	0,2864321	62,5256543244676
fervientemente	*desear*	0,3698225	61,957777978053
atentamente	*escuchar*	0,2001596	60,3643561695563
copiosamente	*sudar*	0,1366384	60,3583868324945
cordialmente	*detestar*	0,03237743	59,4035125720244

Tabla 2-29 Combinaciones sustantivo + verbo con máximo valor de Box-plot

2.6.4. Método de Hampel

El estadístico *MAD* (2), basado en la *MEDA* de los datos, es robusto frente a observaciones atípicas, por lo que métodos más efectivos para la identificación de outliers se basan en él, si bien los cálculos son más complejos (Leys y otros, 2013). La *MEDA* (3) es la mediana de las desviaciones a la mediana de los datos en valores absolutos (3) y *b* (4) en caso de distribuciones normales debe ser: 1,4826.

$$(2)\ \boldsymbol{MAD = b * MEDA}$$

$$(3)\ \boldsymbol{MEDA = mediana\{|x_1 - mediana(X)|, |x_2 - mediana(X)|, \cdots, |x_n - mediana(X)|\}}$$

$$(4)\ \boldsymbol{b = \frac{1}{Q_3}}$$

$X = \{x_1, x_2, \cdots, x_n\}$ es la muestra de la variable.

MAD es un estadístico más adecuado para el problema por las siguientes propiedades:

- El punto de ruptura de *MAD* es de 0,5, superior al del rango intercuartílico que es de 0,25, o 0 en el caso de la media. Este valor indica la proporción máxima de observaciones que pueden ser infinito (estar contaminadas) sin que el estimador se resienta, es decir, que de un valor falso.
- *MAD* no tiene restricciones respecto al tamaño muestral.
- Permite establecer un criterio objetivo fácilmente automatizable para discriminar entre las combinaciones que se consideran outliers y las que no. Por extensión proporciona un mecanismo para identificar automáticamente los candidatos a colocaciones.

El método de Hampel acepta como outlier los valores x_i que cumplen la condición $|MAD(x_i)| \geq 4{,}5$ **(5)**.

$$(\mathbf{5})\ \mathbf{MAD(x_i)} = \frac{\mathbf{|x_i - mediana(X))|}}{\mathbf{MEDA}}$$

La adaptación del método a este trabajo lleva a considerar:

$$MADF(X_i, Palabra) = \frac{|FrecRel(X_i, Palabra) - mediana(Palabra)|}{MEDAF(Palabra)}$$

Siendo:

$$mediana(Palabra) = \text{mediana}\left\{\left|FrecRel\left(x_j, Palabra\right) - mediana(Palabra)\right|_{j=1\cdots n}\right\}$$

y

$$MEDAF(Palabra) = mediana\left\{\left|FrecRel(x_j, Palabra) - mediana(Palabra)\right|_{j=1\cdots n}\right\}$$

Las Tabla 2-31 y Tabla 2-32 muestran los rangos obtenidos en ambos corpus cuando se intentan detectar los outliers a partir de las desviaciones respecto a la mediana de las frecuencias relativas. Hay que tener en cuenta que se pueden realizar los cálculos en los casos en que la *MEDA* sea distinta de cero y que es necesario calcular los valores del indicador para cada uno de los colocados, ya que si el uso respecto a la base no es preferente si lo puede ser respecto al colocativo y viceversa. En este caso, la cobertura aumenta significativamente, si bien en los ejemplos extraídos en torno al punto de corte teórico se aprecian falsas colocaciones (Tabla 2-30).

	Recopiladas		**DCECR.**		**Corpus**	
MAD≥4,5 Max	*hacer*	*alarde*	*dar*	*vuelta*	*encoger*	*hombro*
	echar	*vistazo*	*echar*	*vistazo*	*dar*	*vuelta*
	dar	*palmada*	*dar*	*gracia*	*don*	*doña*
	hacer	*pregunta*	*dar*	*paso*	*abrir*	*puerta*
	ladrar	*perro*	*soplar*	*viento*	*decir*	*señor*
	latir	*corazón*	*dar*	*prisa*	*don*	*quijo*
	poder	*ejecutivo*	*dar*	*salto*	*menear*	*cabeza*
MAD≥4,5 Min	*andar*	*barco*	*dorar*	*fama*	*juir*	*cerrado*
	incurrir	*ira*	*arranque*	*patriotismo*	*servido*	*regentar*
	cifra	*grande*	*fórmula*	*predicción*	*botón*	*suicidio*
	memoria	*frágil*	*sofocar*	*levantamiento*	*adquirir*	*círculo*
	fuente	*fidedigno*	*influenciar*	*dominante*	*sensación*	*huir*
	acarrear	*molestia*	*distorsionar*	*versión*	*maldito*	*amenazar*
	abrir	*crisis*	*estrechar*	*acuerdo*	*pluma*	*relación*

Tabla 2-30 Ejemplos de combinaciones que superan el criterio MAD≥4,5

	Rango	**Combinaciones con *MAD* ≥ 4,5**
Recopiladas	[0 2087,0032]	1532
DCECR	[0 4304,111]	15039
Corpus	[0 4309,9998]	4318191

Tabla 2-31 Resultados obtenidos en el corpus para MAD

	Rango	Combinaciones con *MAD* ≥ 4.5
Recopiladas	[0 68,00004]	933
DCECR	[0 344,9999]	2735
Corpus	[0 344,9999]	455738

Tabla 2-32 Resultados obtenidos en el corpus de Galdós para MAD

Los ejemplos extraídos reflejan cierta tendencia a que los verbos ocupen las posiciones más altas del ranking, y entre éstos, los funcionales como *dar, echar o hacer*.

Los rangos alcanzados en cada uno de los grupos, según la estructura de la combinación, se muestran en la Tabla 2-33. Se aprecian intervalos similares en los grupos *sustantivo + adjetivo* y *verbo + dverbio*, y destaca la amplitud mucho mayor en el caso *verbo + sustantivo*. Respecto a la cobertura en los conjuntos de ensayo, en todas las estructuras sube drásticamente respecto a los resultados obtenidos con el criterio de *ZChebyshev*.

	MAD
Verbo + Sustantivo	[0 4309,9999]
Sustantivo + Adjetivo	[0 2531,476]
Verbo + Adverbio	[0 2332,5601]

Tabla 2-33 Rango de Mad según la estructura de la combinación.

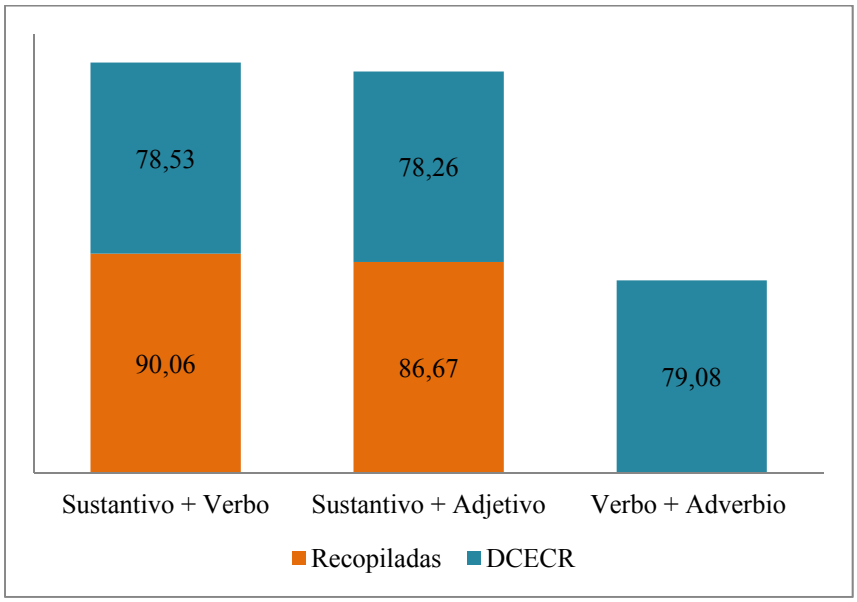
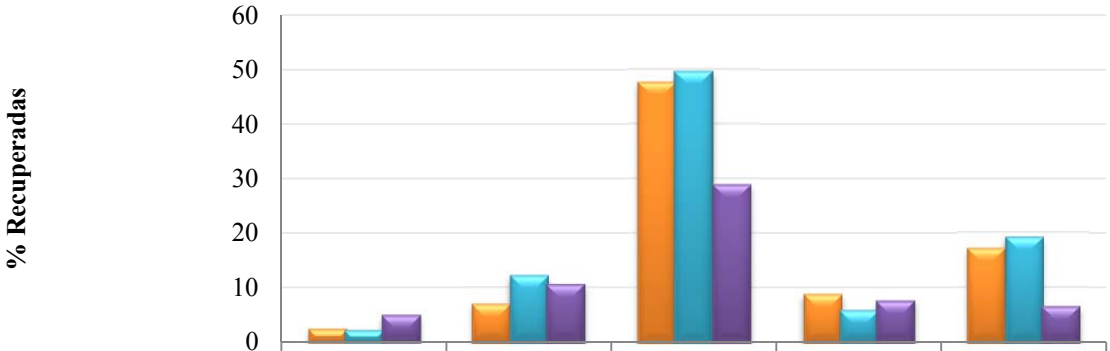


Gráfico 2-15 Combinaciones de los conjuntos de ensayo que superan el criterio de MAD ≥4,5

Para finalizar, se contrasta la cobertura de los distintos criterios barajados sobre los conjuntos de ensayo, teniendo en cuenta también la estructura de la colocación. Los resultados se resumen en

el Gráfico 2-16; en él se aprecia la gran diferencia de cobertura entre *MAD* y el resto. Se observa que en orden de más restrictivos a menos se tiene: frecuencia relativa (≥0,05), información mutua (≥6), *Zchebyshev* (≥3), *Box-Plot* y *MAD* (≥4,5) respectivamente. Sin embargo, ya se ha advertido que la frontera teórica marcada por *MAD* introduce ruido en los resultados.



	Frec. Relativa ≥0,05	Inf. Mutua≥6	MAD >4.5	Zchebyshev ≥3	Box-Plot≥3
Sustantivo + Verbo	2,24948659	6,870332459	47,60672671	8,62319746	17,0239228
Sustantivo + Adjetivo	1,966144696	12,10707363	49,55359021	5,683497238	19,0904178
Verbo + Adverbio	4,731611071	10,47555826	28,81930999	7,42564216	6,41699397

Gráfico 2-16 Comparativa de los criterios de extracción automática

2.7. Estrategia para la extracción de información colocacional generada por un corpus extenso

Fruto del análisis de los resultados registrados en la BD de colocaciones surgen las líneas que marcan su explotación en el desarrollo de un módulo de software que dé soporte a la presentación de las características combinatorias de las palabras en español: sugieren consideraciones a tener en cuenta a la hora de explotar corpus textuales extensos cuando se utilizan indicadores basados en las frecuencias de uso para medir las relaciones de asociación entre palabras. Los resultados de los epígrafes previos llevan a la necesidad de considerar la composición del corpus como una muestra en la que el uso que se da en él a las palabras influye en la posición en cualquier ranking que se intente establecer, y no solo el fenómeno fraseológico. Una muestra muy extensa del lenguaje agranda las diferencias de comportamiento según factores tales como el tipo de elementos que conformen la combinación, la estructura de la colocación o si intervienen palabras funcionales, de gran frecuencia de uso. Por otra parte, el límite entre lo que es colocación cuando los indicadores se basan en la independencia se debe variar en función del tamaño del corpus; sin embargo, los indicadores que pretenden captar el uso preferente permiten establecer un criterio objetivo de forma independiente a este factor. Sobre la base de estas consideraciones se propone seguir una estrategia en la que dada una palabra se muestren sus capacidades combinatorias, de forma que se presente al usuario un

resumen de los datos de frecuencias y se filtren aquellas que se comportan como colocaciones en su rango de variación de frecuencias relativas; es decir, considerando la muestra de su comportamiento en el corpus y no todo el corpus como una muestra de su comportamiento. Siguiendo esta estrategia, se consulta sobre una determinada palabra y se representa el ranking según el o los criterios seleccionados para cada una de las estructuras colocacionales en las que podría formar parte. En la Figura 2-10 se expone un esquema de este proceso que se describe a continuación:

i. Lematizar la palabra para obtener las formas canónicas que la representan en la B. DD.
ii. Dada una de las formas canónicas, se extraen de la BDD las listas ordenadas de combinaciones según alguno de las medidas de asociación: *frecuencia relativa, información mutua, z-score, ZChebyshev, Box-Plot* o *MAD*. Este proceso se lleva a cabo para cada una de las posibles estructuras: *verbo + sustantivo, sustantivo + adjetivo o verbo + adverbio* en las que pueda intervenir la forma canónica.

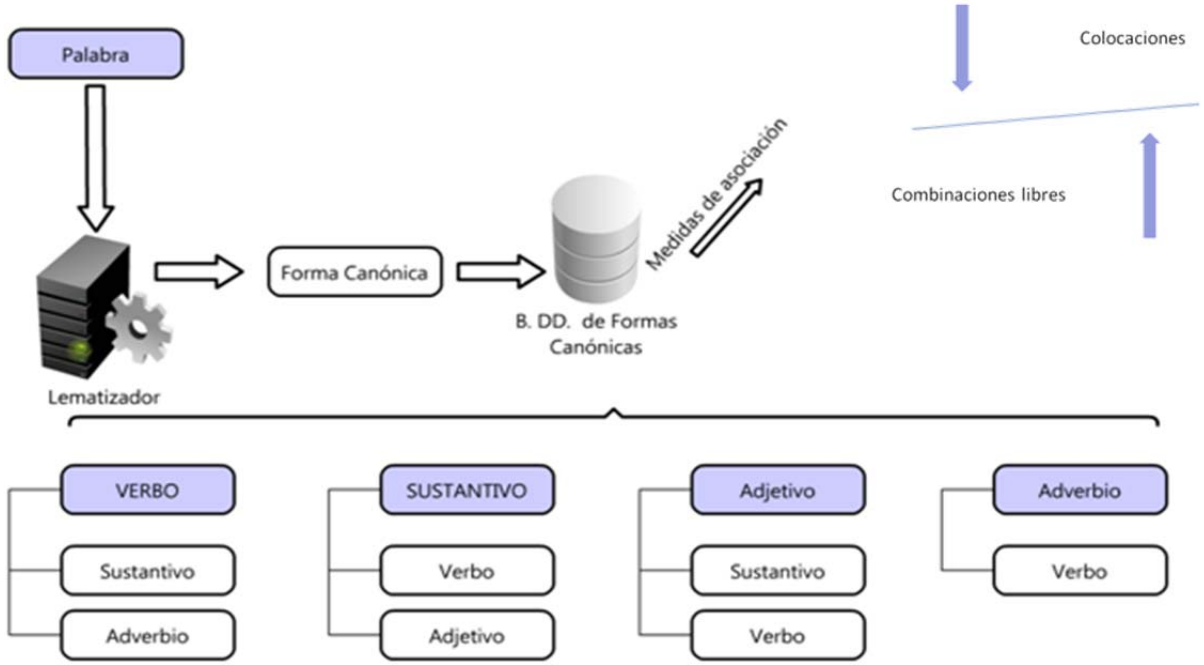


Figura 2-10 Extracción de colocaciones en corpus textuales extenso.

Un ejemplo de uso de esta metodología cuando la palabra de referencia es el verbo *palpitar* y se aplica el indicador *Box-Plot* evaluado sobre los sustantivos con los que aparece en el corpus, arroja los resultados mostrados en el Gráfico 2-17. Las 20 combinaciones que se obtienen los valores máximos de *Box-Plot* se listan en la Tabla 2-34.

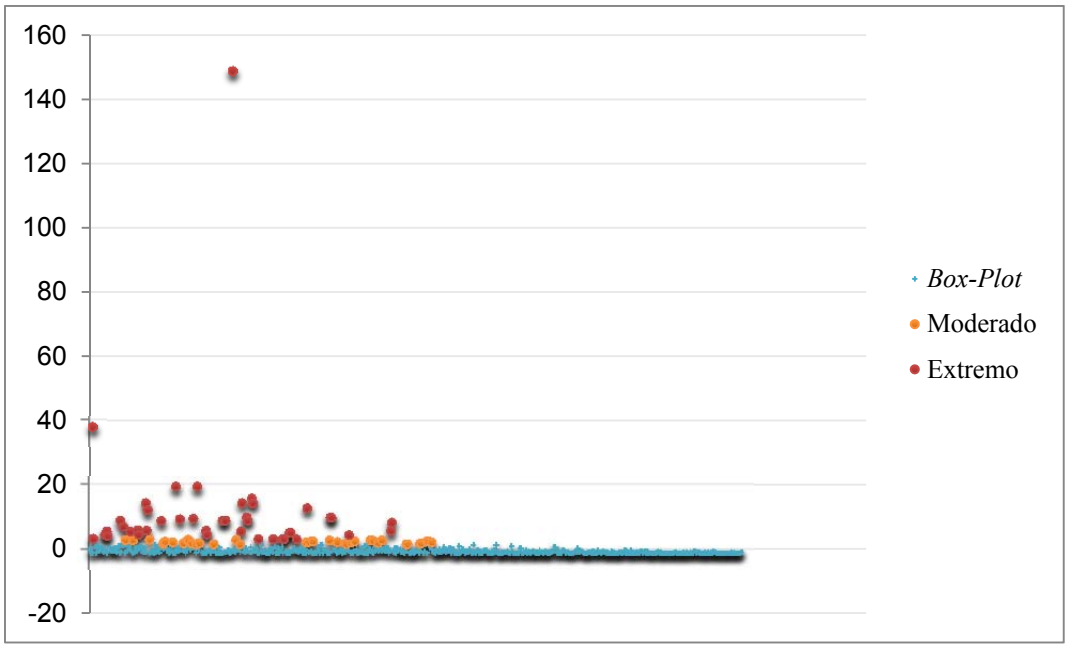


Gráfico 2-17 *Box-Plot* para las combinaciones palpitar + Sustantivo en el corpus

	x_i	*FrecRelativa(x_i, palpitar)*	*Box-Plot(x_i, palpitar)*
palpitar	*corazón*	0,330837	148,899993456854
	pecho	0,08634361	37,8999958880714
	vida	0,04537445	19,2999986998272
	vido	0,04537445	19,2999986998272
	ojo	0,03744493	15,6999990803656
	baja	0,03436123	14,2999983827119
	bajo	0,03436123	14,2999983827119
	amor	0,03414097	14,1999996617437
	fuerza	0,03083701	12,6999993974809
	sangre	0,02951542	12,099999460904
	mano	0,0246696	9,89999969345519
	seno	0,0246696	9,89999969345519
	manos	0,02444934	9,79999928120528
	cabeza	0,02378855	9,49999973573724
	vez	0,02334802	9,29999975687826
	luces	0,02246696	8,89999895351946
	luz	0,02246696	8,89999895351946

Tabla 2-34 Valores máximos de *Box-Plot* para el verbo palpitar + Sustantivo

2.7.1. Resultados obtenidos sobre el corpus completo

En las tablas Tabla 2-35 a la Tabla 2-58 se muestran ejemplos de los listados obtenidos siguiendo esta estrategia cuando se consultan formas canónicas (FC) que aparecen en distintos rangos de frecuencia en el corpus. Se han elegido los verbos *dar* y *propinar* (Tabla 2-35, a Tabla 2-40), los sustantivos *tierra* y *contrato* (Tabla 2-41 a Tabla 2-46), los adjetivos *público* y *omiso* (Tabla 2-47 a Tabla 2-52), y los adverbios *libremente* y *frugalmente* (Tabla 2-53). En todos los casos se utiliza *ZChebyshev*, *Mad* y *Box-Plot*. Se muestran los 10 mejores, los 10 más cercanos al umbral utilizado para discriminar y 10 de los que obtienen valor mínimo, siempre que sea posible. Hay que señalar que para el adverbio *frugalmente* se muestran todos los casos de combinaciones que aparecen en el corpus.

Max *ZChebyshev*		**Punto de corte *ZChebyshev***		**Min *ZChebyshev***	
vuelto	78,60	*bodego*	3,00	*pipado*	0
vuelta	74,39	*alusión*	3,00	*pretura*	0
golpe	52,28	*pe*	3,00	*elogiado*	0
pasa	51,76	*tarifa*	3,00	*pecunia*	0
paso	51,76	*ubicación*	3,00	*sene*	0
gracia	48,06	*desconcertado*	3,00	*anclote*	0
muestra	46,03	*marchado*	3,00	*molinosismo*	0
salto	45,83	*brote*	3,00	*canguelo*	0
orden	43,30	*inoportuno*	3,00	*remado*	0
órdenes	43,30	*evento*	3,00	*robla*	0

Tabla 2-35 Resultados ZChebyshev, verbo dar + Sustantivo

Max MAD		**Punto de corte Mad**		**Min Mad**	
vuelta	4304,11	*cáliz*	4,51	*jengibre*	0
vuelto	3969,09	*presunción*	4,51	*yunque*	0
palmado	1963,00	*aprobado*	4,51	*palabrita*	0
palmada	1962,49	*editorial*	4,51	*naire*	0
gracia	1674,36	*lentitud*	4,51	*mosquetería*	0
pasa	1471,76	*manual*	4,51	*molusco*	0
paso	1471,76	*montañoso*	4,50	*manillar*	0
priso	1382,75	*suplicación*	4,50	*judicatura*	0
prisa	1378,25	*candelaria*	4,50	*landa*	0
golpe	1329,11	*tensó*	4,50	*latamente*	0

Tabla 2-36 Resultados MAD, verbo dar + Sustantivo

Max *Box-Plot*		Punto de corte *Box-Plot*		Min *Box-Plot*	
vuelto	265,42	*intelectual*	3,03	*sociable*	-1,03
vuelta	259,01	*docena*	3,03	*socioeconómico*	-1,03
hombre	207,23	*comentarios*	3,03	*subjefe*	-1,03
señor	179,59	*comentario*	3,03	*substituto*	-1,03
señora	179,59	*acero*	3,02	*tecnicismo*	-1,03
diosa	178,23	*altar*	3,01	*teleología*	-1,03
dios	178,23	*cuartel*	3,01	*teologal*	-1,03
mujer	173,11	*tradición*	3,01	*sufijo*	-1,03
bienes	169,78	*traducción*	3,00	*susurrante*	-1,03
bien	169,78	*contacto*	3,00	*tetilla*	-1,03

Tabla 2-37 Resultados *Box-Plot*, verbo dar + Sustantivo

Min *ZChebyshev*		Min *ZChebyshev*	
golpe	13,69	*fin*	0
paliza	5,588	*muchacha*	0
bueno	5,075	*muchacho*	0
patada	4,936	*sonoro*	0
buena	4,424	*trasero*	0
puntapié	4,377	*joven*	0
puñetazo	4,191	*mordisco*	0
puñete	4,191	*sonora*	0
fuerte	3,632	*bofetón*	0
vez	3,25	*trasera*	0

Tabla 2-38 Resultados ZChebyshev, verbo propinar + Sustantivo

Max MAD		Punto de corte MAD		Min MAD	
golpe	301,00	*amor*	5,00	*vero*	0
paliza	127,00	*momento*	5,00	*vera*	0
paliza	126,99	*mandíbula*	5,00	*obra*	0
bueno	116,00	*hermana*	5,00	*problemas*	0
patada	113,00	*hermano*	5,00	*descubierta*	0
buena	102,00	*gracia*	5,00	*actitud*	0
puntapié	101,00	*oro*	5,00	*soldado*	0
puñete	97,000	*bebida*	5,00	*complacencia*	0
puñetazo	97,000	*sonrisa*	5,00	*complacencia*	0
fuerte	85,000	*caso*	5,00	*problema*	0

Tabla 2-39 Resultados MAD, verbo propinar +Sustantivo

Max *Box-Plot*		Punto de corte *Box-Plot*		Min *Box-Plot*	
golpe	42,14	*dólar*	6,14	*lejos*	-1
paliza	17,29	*generoso*	4,86	*manera*	-1
bueno	15,71	*padre*	4,57	*manero*	-1
patada	15,29	*codo*	3,71	*nalga*	-1
buena	13,71	*madre*	3,57	*esclava*	-1
puntapié	13,57	*tiempo*	3,43	*esclavo*	-1
puñetazo	13,00	*señor*	3,29	*estilo*	-1
puñete	13,00	*señora*	3,29	*vicioso*	-1
fuerte	11,29	*ante*	3,00	*felpa*	-1
vez	10,14	*antes*	3,00	*gratificado*	-1

Tabla 2-40 Resultados *Box-Plot*, verbo propinar

Se observa que los valores máximos de *ZChebyshev* y *MAD* arrojan buenos resultados en las regiones de máximos, produciendo además listados similares. Los valores con puntuaciones más altas de *Box-Plot* son peores para el caso del verbo *dar*, en el que las combinaciones que se detectan como anomalías corresponden a palabras de uso frecuente, más que de uso preferente con este verbo. Asimismo, los valores mínimos de los tres se producen entre combinaciones libres. La diferencia principal surge en las inmediaciones del umbral que

marca la zona de lo que se propondría como colocación. Tanto en *MAD* como *Box-Plot* corresponde a una región difusa donde se mezclan combinaciones libres: *propinar hermana* con otras con cierta fijación como *propinar amor* o *propinar sonrisa*. *ZChebyshev* establece un punto de corte más drástico, lo que se traduce en que la zona de aceptación, si se ciñe a palabras de uso más restringido, prácticamente coincide con los máximos seleccionados. Un comportamiento similar está presente en los restantes ejemplos barajados (Tablas Tabla 2-41 a Tabla 2-49).

Max *ZChebyshev*		**Punto de corte *ZChebyshev***		**Min *ZChebyshev***	
firme	33,83	*corona*	3,00	*nulidad*	0
cielo	31,00	*monte*	3,00	*mutación*	0
firmar	30,71	*venida*	3,00	*perturbación*	0
adentrar	30,60	*bordo*	3,00	*domeñar*	0
hacer	30,08	*demasiar*	3,00	*lotófago*	0
mar	29,74	*contornos*	3,00	*amainar*	0
cultivar	28,06	*contorno*	3,00	*ufanía*	0
tener	28,06	*muerte*	3,00	*impeditivo*	0
poder	27,49	*amarrar*	3,00	*desjugar*	0
decir	24,82	*carga*	3,00	*remado*	0

Tabla 2-41 Resultados ZChebyshev, sustantivo tierra + Adjetivo

Max MAD		**Punto de corte MAD**		**Min MAD**	
hacer	826,72	*fluvial*	4,66	*carrete*	0
cielo	784,81	*sinuoso*	4,66	*carbonero*	0
adentrar	780,66	*enemistad*	4,66	*caminata*	0
tener	771,54	*carbono*	4,66	*chachapoyo*	0
poder	756,09	*confinado*	4,62	*consciencia*	0
decir	683,36	*usura*	4,59	*cárdeno*	0
firme	682,84	*acorde*	4,59	*cincelado*	0
mar	680,78	*recitar*	4,54	*cigarrillo*	0
ver	654,00	*tenso*	4,50	*deliberación*	0
cielo	570,50	*medieval*	4,5	*desacato*	0

Tabla 2-42 Resultados MAD, sustantivo tierra + Adjetivo

Max *Box-Plot*		Punto de corte *Box-Plot*		Min *Box-Plot*	
hacer	182,07	*tronco*	3,05	*desasir*	-0,03
tener	169,93	*social*	3,04	*yate*	-0,03
poder	166,53	*recibido*	3,04	*languidecer*	-0,03
decir	150,53	*morado*	3,04	*empedernir*	-0,03
ver	144,07	*morada*	3,04	*amordazar*	-0,03
cielo	125,70	*rumbo*	3,03	*erar*	-0,03
dar	123,11	*cruel*	3,02	*reportar*	-0,03
hombre	109,40	*cortés*	3,02	*violar*	-0,03
ir	101,36	*dilatar*	3,02	*tortuoso*	-0,03
mar	95,42	*acostumbrar*	3,02	*traicionero*	-0,03

Tabla 2-43 Resultados *Box-Plot*, sustantivo tierra + Adjetivo

Max *ZChebyshev*		Punto de corte *ZChebyshev*		Min *ZChebyshev*	
poder	15,93	*firmado*	3,54	*fundamento*	0
hacer	15,79	*ver*	3,44	*idea*	0
firmar	14,88	*contrato*	3,41	*ideo*	0
tener	13,55	*contrato*	3,40	*número*	0
celebrar	11,16	*hecho*	3,38	*recordar*	0
decir	9,83	*cancelar*	3,25	*reinar*	0
matrimonial	7,97	*gobierno*	3,23	*solar*	0
firmar	7,10	*haber*	3,21	*ir*	0
social	6,48	*contrato*	3,16	*mercantil*	0
dar	6,27	*nuevo*	3,06	*contrato*	0

Tabla 2-44 Resultados ZChebyshev, sustantivo contrato + Adjetivo

Max MAD		Punto de corte MAD		Min MAD	
poder	192,33	*fecho*	3,22	*condicional*	0
hacer	190,66	*contrato*	4,62	*bastir*	0
firmar	179,83	*promesa*	4,62	*destruir*	0
matrimonial	174,00	*millo*	4,59	*demasiar*	0
tener	164,16	*millón*	4,57	*encargar*	0
celebrar	135,83	*nombre*	4,57	*dotar*	0
firmar	134,62	*pueblo*	4,52	*extinguir*	0
decir	119,99	*mercantil*	4,52	*manifestar*	0
celebrar	116,28	*arrendatario*	4,50	*salariar*	0
rescindir	86,999	*usar*	4,50	*cincuenta*	0

Tabla 2-45 Resultados MAD, sustantivo contrato + Adjetivo

Max *Box-Plot*		Punto de corte *Box-Plot*		Min *Box-Plot*	
poder	49,39	*vido*	3,17	*auxiliador*	-1,04
hacer	48,96	*parecer*	3,17	*contracto*	-1,04
firmar	46,13	*vida*	3,17	*contracto*	-1,04
tener	42,04	*padre*	3,17	*comodato*	-1,04
celebrar	34,65	*conocer*	3,09	*ayuno*	-1,04
decir	30,52	*pago*	3,09	*acordado*	-1,04
social	20,17	*volver*	3,09	*contacto*	-1,04
dar	19,52	*pueblo*	3,04	*abuelo*	-1,04
partir	17,87	*conseguir*	3,04	*ayunas*	-1,04
deber	17,78	*nombre*	3,04	*condesa*	-1,04

Tabla 2-46 Resultados *Box-Plot*, sustantivo contrato + Adjetivo

Max *ZChebyshev*		**Punto de corte *ZChebyshev***		**Min *ZChebyshev***	
hacer	29,05	*judicial*	3,03	*animal*	0
poder	27,70	*económico*	3,03	*terma*	0
privado	25,57	*acogida*	3,03	*distraído*	0
funcionario	24,70	*gasto*	3,02	*intrigante*	0
privar	24,27	*caso*	3,02	*imprescriptible*	0
instrucciones	23,51	*universidad*	3,02	*precaución*	0
instrucción	23,51	*dirigir*	3,01	*latas*	0
ministerio	20,97	*ilustrar*	3,01	*gana*	0
administración	20,54	*tener*	3,01	*arriesgar*	0
tener	20,47	*militar*	3,01	*trimestralmente*	0

Tabla 2-47 Resultados ZChebyshev, adjetivo público

Max MAD		**Punto de corte, MAD**		**Min MAD**	
hacer	636,21	*mentalidad*	4,50	*suficiencia*	0
poder	606,85	*adaptar*	4,50	*inválido*	0
subastar	591,00	*estampa*	4,50	*público*	0
privar	481,14	*suelo*	4,50	*quieto*	0
privado	469,14	*ventana*	4,50	*público*	0
tener	449,99	*transformación*	4,50	*conminar*	0
privada	427,00	*gubernamental*	4,5	*etiquetar*	0
instrucción	416,75	*bofetada*	4,5	*mortificar*	0
instrucciones	416,75	*sobrevenir*	4,5	*semillar*	0
funcionario	370,00	*agencia*	4,5	*público*	0

Tabla 2-48 Resultados MAD, adjetivo público + Adjetivo

Max *Box-Plot*		Punto de corte, *Box-Plot*		Min *Box-Plot*	
bien	54,41	*género*	3,09	*prosa*	-0,93
bienes	54,41	*tarde*	3,07	*rienda*	-0,93
leyes	51,77	*consideración*	3,07	*romería*	-0,93
ley	51,77	*decreto*	3,06	*acabo*	-0,93
obra	51,03	*garantía*	3,04	*basura*	-0,93
servicio	48,64	*voluntad*	3,03	*basuras*	-0,93
instrucción	47,41	*gestión*	3,03	*calificación*	-0,93
instrucciones	47,41	*comunidad*	3,03	*cuestor*	-0,93
opinión	46,43	*presencia*	3,01	*aristocracia*	-0,93
orden	46,13	*sentimiento*	3,01	*cuánto*	-0,93

Tabla 2-49 Resultados *Box-Plot*, adjetivo público + Adjetivo

Se destaca el comportamiento de formas canónicas fácilmente decodificables por tener una única acepción como el adjetivo *omiso:* en la zona de transición cuando se utiliza *MAD*, las combinaciones que aparecen tienen una relación estrecha que induce a catalogarlas como colocaciones: *remitente omiso, insulto omiso, parte omiso, protesta omiso*. Se observan también los efectos excesivamente restrictivos de *ZChebyshev* en palabras con estas características. En este caso, solo las formas canónicas que provienen de *hacer caso omiso,* superan la restricción impuesta por este indicador (Tabla 2-50). También destaca el buen comportamiento de *Box-Plot* que parece mejorar los resultados, introduciendo entre los casos anómalos palabras con menor frecuencia de uso.

Max *ZChebyshev*		
omiso	*hacer*	10,04
omiso	*casar*	10,01
omiso	*caso*	9,889
omiso	*hacendar*	2,980
omiso	*omiso*	2,814
omiso	*hacendar*	1,871
omiso	*casar*	1,855
omiso	*hacer*	0,563
omiso	*vanguardia*	0,492
omiso	*convención*	0,479

Tabla 2-50 Resultados ZChebyshev, adjetivo omiso + Adjetivo

Max MAD		**Punto de corte, MAD**		**Min MAD**	
hacer	1495,00	*remitente*	6,00	*culpar*	0
casar	1491,00	*señalar*	6,00	*insultar*	0
caso	1472,00	*soler*	6,00	*observar*	0
hacendar	457,000	*insulto*	6,00	*padre*	0
caso	91,0625	*llamar*	6,00	*procurar*	0
poder	76,0000	*sentir*	6,00	*recomendación*	0
casar	73,6000	*parte*	5,00	*regla*	0
decir	69,0000	*partes*	5,00	*sugerencia*	0
tener	58,0000	*ser*	5,00	*teléfono*	0
hecho	51,0000	*protesta*	5,00	*texto*	0

Tabla 2-51 Resultados MAD, adjetivo omiso + Adjetivo

Max Box-Plot		***Min Box-Plot***	
caso	367,25	*órdenes*	2,25
hombre	9,50	*mano*	2,25
pregunta	8,50	*manos*	2,00
palabra	7,00	*advertencia*	2,00
mujer	6,25	*presencia*	1,75
comentario	5,25	*interrupción*	1,50
comentarios	5,25	*bien*	1,50
dolor	4,50	*bienes*	1,50
dolores	4,50	*voz*	1,00
ley	3,50	*cámara*	0,75
leyes	3,5	*insulto*	0,75
protesta	3,5		

Tabla 2-52 Resultados *Box-Plot*, omiso + Sustantivo

Nótese como el comportamiento se reproduce también en los adverbios seleccionados para ilustrar los resultados, destacando el efecto restrictivo de *ZChebyshev* que excluye todas las combinaciones halladas del conjunto de colocaciones.

Max *ZChebyshev*		Punto de corte, *ZChebyshev*		Min *ZChebyshev*	
hacer	29,05	*relacionado*	4,24	*cierto*	0,01
poder	27,70	*vincular*	4,07	*entrelazar*	0,01
privado	25,57	*poder*	4,03	*limitado*	0,01
funcionario	24,70	*relacionado*	3,90	*social*	0,01
privar	24,27	*emparentado*	3,69	*llamar*	0,01
instrucciones	23,51	*abrazado*	3,53	*pasar*	0,01
instrucción	23,51	*vinculado*	3,30	*permanecer*	0,01
ministerio	20,97	*tener*	3,22	*tratar*	0,01
administración	20,54	*emparentar*	3,21	*humanar*	0,01
tener	20,47	*abrazar*	3,19	*entrelazado*	0,00

Tabla 2-53 Resultados ZChebyshev, adverbio estrechamente + Verbo

Max MAD		Punto de corte MAD		Min MAD	
unir	165,66	*bajar*	5,00	*suficiencia*	0
abrazar	120,99	*guardar*	5,00	*inválido*	0
unido	110,99	*ideo*	5,00	*público*	0
ligar	107,99	*general*	5,00	*quieto*	0
relacionar	103,33	*señor*	5,00	*público*	0
ligado	92,666	*conectar*	4,74	*conminar*	0
ligado	92,666	*puntar*	4,66	*etiquetar*	0
abrazar	90,500	*sentido*	4,66	*mortificar*	0
ligar	80,749	*poblar*	4,66	*semillar*	0
relacionado	71,999	*apretar*	4,60	*público*	0

Tabla 2-54 Resultados MAD, adverbio estrechamente + Verbo

Max *Box-Plot*		Min *Box-Plot*	
unir	39,97	*preguntar*	-1,27
relacionar	33,93	*resultar*	-1,27
ligar	33,45	*esperar*	-1,27
vincular	24,17	*dirigir*	-1,29
abrazar	21,14	*pensar*	-1,30
enlazar	9,08	*sentar*	-1,30
vigilar	8,05	*escribir*	-1,32
asociar	6,69	*salir*	-1,36
aliar	5,55	*vezar*	-1,39
ceñir	4,44	*llevar*	-1,45

Tabla 2-55 Resultados *Box-Plot*, adverbio estrechamente + Verbo

		ZChebyshev
frugalmente	*comer*	1,76
frugalmente	*desayunar*	0,72
frugalmente	*almorzar*	0,38
frugalmente	*vivir*	0,38
frugalmente	*desayunar*	0,33
frugalmente	*cenar*	0,27
frugalmente	*cenar*	0,24
frugalmente	*almorzar*	0,24
frugalmente	*vivir*	0,24
frugalmente	*comer*	0,20

Tabla 2-56 Resultados ZChebyshev, adverbio frugalmente + Verbo

		Box-Plot
frugalmente	*comer*	72,30
frugalmente	*cenar*	68,09
frugalmente	*desayunar*	45,83
frugalmente	*almorzar*	21,45
frugalmente	*vivir*	7,56

Tabla 2-57 Resultados Box-Plot, adverbio frugalmente + Verbo

		MAD
frugalmente	*comer*	18,99
frugalmente	*desayunar*	3
frugalmente	*comer*	1,399
frugalmente	*cenar*	1
frugalmente	*almorzar*	0,666
frugalmente	*vivir*	0,642
frugalmente	*desayunar*	0,500
frugalmente	*cenar*	0,199
frugalmente	*vivir*	0
frugalmente	*almorzar*	0

Tabla 2-58 Resultados MAD, adverbio frugalmente + Verbo

Para finalizar, cabe resaltar en este apartado la necesidad del tratamiento diferencial de los datos según la semántica de la palabra, ya que incide directamente en su uso. Véase en los ejemplos que palabras con valor funcional como el verbo *dar*, o el adverbio *estrechamente* abarcan un espectro más amplio de combinaciones, frente a *propinar, omiso* o *frugalmente*, que no solo son palabras de menor frecuencia de uso, sino que en el *DRAE*. solo disponen de una acepción (Tabla 2-59).

	Acepciones	Frecuencia
dar	53	1.118.012
propinar	3	2.970
contrato	1	12.244
tierra	10	286.387
público	8	109.763
omiso	1	1.581
estrechamente	4	3.188
frugalmente	1	73

Tabla 2-59 Acepciones y frecuencias en el corpus de los casos analizados.

Cuando se contrastan las combinaciones mejor posicionadas según cada indicador analizado se manifiestan nuevamente las diferencias entre palabras cuya frecuencia de uso es dispar. Las tablas Tabla 2-60 y Tabla 2-61 muestran que en general los casos mejor posicionados coinciden cuando se trata de una palabra cuya frecuencia no es alta y presenta pocas acepciones. Por el contrario, la información mutua arroja resultados similares a las mayores frecuencias relativas evaluadas respecto al otro elemento de la combinación. El resto de los indicadores aportan resultados similares.

	FrecRel	**FrecRel**	***InfMutua***	***ZScore***	***t-Score***	***ZChebyshev***	***MAD***	***Box-Plot***
dar	*adestrador*	*vuelto*	*adestrador*	*vuelta*	*vuelto*	*vuelto*	*vuelta*	*vuelto*
	moravedí	*vuelta*	*moravedí*	*vuelto*	*vuelta*	*vuelta*	*vuelto*	*vuelta*
	carpetazo	*hombre*	*carpetazo*	*palmada*	*hombre*	*golpe*	*palmado*	*hombre*
	traspié	*señora*	*traspié*	*palmado*	*dios*	*pasa*	*palmada*	*señor*
	palmada	*señor*	*palmada*	*salto*	*diosa*	*paso*	*gracia*	*señora*
	palmado	*dios*	*palmado*	*golpe*	*pasa*	*gracia*	*pasa*	*diosa*
	respingo	*diosa*	*respingo*	*gracia*	*paso*	*muestra*	*paso*	*dios*
	golpecito	*mujer*	*golpecito*	*golpecito*	*señora*	*salto*	*priso*	*mujer*

Tabla 2-60 Colocaciones según el indicador, verbo dar + Sustantivo

	Frec. Rel.	Frec. Rel. propinar	*InfMutua*	*ZScore*	*t-Score*	*ZChebyshev*	*MAD*	*Box-Plot*
propinar	*paliza*	*golpe*	*paliza*	*paliza*	*golpe*	*golpe*	*golpe*	*golpe*
	puntapié	*paliza*	*puntapié*	*puntapié*	*paliza*	*paliza*	*paliza*	*paliza*
	puñetazo	*bueno*	*puñetazo*	*puñetazo*	*patada*	*bueno*	*paliza*	*bueno*
	puñete	*patada*	*puñete*	*puñete*	*bueno*	*patada*	*bueno*	*patada*
	patada	*buena*	*patada*	*patada*	*puntapié*	*buena*	*patada*	*buena*
	codazo	*puntapié*	*codazo*	*golpe*	*buena*	*puntapié*	*buena*	*puntapié*
	bofetada	*puñete*	*bofetada*	*bofetada*	*puñetazo*	*puñetazo*	*puntapié*	*puñetazo*
	bofetón	*puñetazo*	*bofetón*	*codazo*	*puñete*	*puñete*	*puñete*	*puñete*

Tabla 2-61 Colocaciones según el indicador, verbo propinar + Sustantivo

	Frec. Rel_{Adj}	Frec. Rel_{tierra}.	*Inf. Mutua*	*ZScore*	*t-Score*	*ZChebyshev*	*Box Plot*	*MAD*
tierra	*temporejado*	*tanto*	*temporejado*	*firme*	*baja*	*firme*	*baja*	*baja*
	lejas	*baja*	*lejas*	*fértil*	*bajo*	*cielo*	*bajo*	*bajo*
	estercolado	*bajo*	*estercolado*	*apisonado*	*mucho*	*firmar*	*nuevo*	*firme*
	labrantío	*mucho*	*labrantío*	*natal*	*grande*	*adentrar*	*bien*	*medio*
	apisonado	*grande*	*apisonado*	*baldío*	*nuevo*	*hacer*	*padre*	*vido*
	lantánido	*nuevo*	*lantánido*	*labrantío*	*tanto*	*mar*	*firme*	*indio*
	cultivable	*señor*	*cultivable*	*lejas*	*firme*	*cultivar*	*medio*	*alto*
	alquiladizo	*señora*	*alquiladizo*	*poblado*	*señora*	*tener*	*vido*	*visto*

Tabla 2-62 Colocaciones según el indicador, sustantivo tierra + Adjetivo

	Frec. Rel_{Adj}	Frec. Rel contrato	*Inf Mutua*	*ZScore*	*t-Score*	*ZChebyshev*	*Box Plot*	*MAD*
contrato	*rescindido*	*poder*	*rescindido*	*matrimonial*	*social*	*social*	*social*	*social*
	sinalagmático	*hacer*	*sinalagmático*	*firmado*	*derecha*	*derecha*	*derecha*	*derecha*
	prenupcial	*firmar*	*prenupcial*	*rescindido*	*derecho*	*derecho*	*derecho*	*derecho*
	revisable	*tener*	*revisable*	*celebrado*	*público*	*público*	*público*	*público*
	consensual	*celebrar*	*consensual*	*usurario*	*firmado*	*firmado*	*firmado*	*firmado*
	usurario	*decir*	*usurario*	*social*	*hecho*	*hecho*	*hecho*	*hecho*
	enfitéutico	*social*	*enfitéutico*	*consensual*	*celebrado*	*nuevo*	*nuevo*	*nuevo*
	canjeado	*dar*	*canjeado*	*sinalagmático*	*nuevo*	*celebrado*	*celebrado*	*celebrado*

Tabla 2-63 Colocaciones según el indicador, sustantivo contrato + Adjetivo

2.7.2. Resultados obtenidos sobre el corpus de Galdós

Se centra la discusión en las mismas formas utilizadas para ilustrar los resultados del epígrafe anterior, se incluye en la Tabla 2-64 sus frecuencias calculadas sobre el Corpus de Galdós.

	Frecuencia	Combinaciones
dar	9628	972
propinar	15	0
tierra	689	88
contrato	20	1
público	461	66
omiso	3	3
estrechamente	9	1
frugalmente	1	0

Tabla 2-64 Frecuencias en el corpus de Galdós de los ejemplos

Si se utiliza la información mutua o el valor de *Box-plot* en los valores máximos se encuentran las mismas colocaciones mejor posicionadas. Se observa que este último es más eficaz usándose en corpus menos extensos, ya que no surgen las distorsiones originadas por las palabras de uso muy frecuente frente a otras menos habituales. La palabra *carpetazo* aparece 3 veces en este corpus, todas ellas con el verbo *dar* (Tabla 2-65).

	ZCh		MAD		Box-Plot		InfM		Frec R	
dar	*vuelta*	10,07	*vuelta*	85,50	*carpetazo*	23,71	*carpetazo*	2,91	*vuelta*	0,03
	vuelto	10,07	*vuelto*	85,50	*atraque*	23,71	*atraque*	2,69	*vuelto*	0,03
	pasa	8,09	*paso*	69,25	*atraco*	19,07	*atracón*	2,69	*pasa*	0,03
	paso	8,09	*pasa*	69,25	*atracón*	19,07	*atraco*	2,69	*paso*	0,03
	gana	5,41	*gana*	47,25	*morrada*	12,79	*morrada*	2,59	*dar*	0,02
	diosa	5,01	*diosa*	44,00	*grima*	11,86	*grima*	2,59	*gana*	0,02
	dios	5,01	*dios*	44,00	*palmada*	11,07	*palmado*	2,56	*diosa*	0,02
	mano	4,67	*mano*	41,25	*palmado*	11,07	*palmada*	2,56	*dios*	0,02
	manos	4,64	*manos*	41,00	*pábulo*	11,00	*pábulo*	2,55	*mano*	0,02
	cuenta	4,34	*cuenta*	38,50	*traste*	10,29	*traste*	2,53	*manos*	0,02

Tabla 2-65 Colocaciones en el corpus de Galdós según el indicador, verbo dar

	ZCh		MAD		Box-Plot		InfM		Frec.	
tierra	*cielo*	5,67	*cielo*	38,00	espuerta	52,61	*espuerta*	4,49	*cielo*	0,060
	echar	3,12	*echar*	22,00	entraña	37,37	*entraña*	3,20	*echar*	0,037
	ver	3,12	*ver*	22,00	entraño	37,37	*entraño*	3,20	*ver*	0,037
			hacer	13,00	expediente	37,05	*expediente*	3,17	*hacer*	0,024
			dar	13,00	expediente	37,05	*expediente*	3,17	*dar*	0,024
			venir	12,00	cielo	29,61	*expedientar*	3,17	*haber*	0,023
			haber	12,00	fecundo	29,06	*entrañar*	3,05	*venir*	0,023
			bien	10,00	montón	18,23	*cielo*	2,54	*tener*	0,02
			tener	10,00	mortal	17,14	*fecundo*	2,49	*ir*	0,02
			ir	10,00	centro	13,56	*rodrigar*	2,44	*bien*	0,02

Tabla 2-66 Colocaciones en el Corpus de Galdós, sustantivo tierra

	ZCh			MAD		Box Plot		Inf.		FrecR
	orden	4,11	*órdenes*	23	*orden*	*5,4*	*privada*	4,80	*órdenes*	0,06
	órdenes	4,11	*orden*	23	*órdenes*	*5,4*	*instrucciones*	4,66	*orden*	0,06
			público	13			*instrucción*	4,66	*hacer*	0,039
			público	11			*implorar*	4,49	*ver*	0,03
público			*público*	11			*notorio*	3,97	*dar*	0,03
			instrucciones	8			*vía*	3,94	*instrucción*	0,02
			instrucción	8			*privado*	3,93	*instrucciones*	0,02
			decir	7			*sitios*	3,27	*decir*	0,02
			hombre	7			*privar*	3,00	*hombre*	0,02
			voz	6			*hacienda*	2,93	*voz*	0,02

Tabla 2-67 Colocaciones en el Corpus de Galdós, adjetivo público

Se manifiesta la dificultad a la hora de encontrar muestras de las combinaciones que se ajusten a estructuras colocacionales en español cuando el corpus es significativamente menos extenso: *contrato* y *estrechamente* solo producen una, *omiso* dos y *frugalmente* ninguna. Es por ello que *ZChebyshev*, *MAD* y *Box-Plot* no pueden seleccionar combinaciones de uso preferente, puesto que la ausencia de elementos donde evaluar la variación impide captar anomalías en las frecuencias relativas (Tabla 2-68).

		Inf. Mutua	Frec. Relativa
contrato	*hacer*	-0,22	0,15
omiso	*caso*	5,82	1
omiso	*casar*	3,26	1
omiso	*hacer*	2,512	1
estrechamente	*abrazar*	6,65	0,44

Tabla 2-68 Ejemplos en los que no se encuentra variación

3. Detección de nexos con grupos semánticos

3.1. Análisis cualitativo de los datos y selección léxica

El análisis estadístico del corpus permite extraer una gran cantidad de combinaciones que representan las propiedades combinatorias de una amplia muestra del español. Las estrategias seguidas para la obtención de las colocaciones se han centrado en determinar aquellas combinaciones que reflejan el uso preferente de las mismas frente a cualquier otra combinación posible. Sin embargo, se hace patente en los conjuntos extraídos para cualquier forma canónica, que se pueden establecer grupos o clases a partir de criterios semánticos. Se observa que las bases con las que combina un determinado verbo, adverbio o sustantivo en las colocaciones *verbo* + *sustantivo*, *adverbio* + *verbo* o *sustantivo* + *adjetivo* pertenecen a grupos semánticos, los elementos recuperados mediante las técnicas de detección de outliers representan los de uso preferente. Se presenta como ejemplo de este fenómeno algunos de los sustantivos que se obtienen con los verbos *sentir* y *palpitar,* los adverbios *enérgicamente* y *abundamente* y los adjetivos *brusco* y *doloroso* (Tabla 3-1).

Estos resultados se pueden explicar a partir del fenómeno de la selección léxica. En Bosque (2 001) se resalta la interpretación de las colocaciones como un caso particular del concepto más amplio de selección léxica. Una determinada base de forma aislada no constituye el único caso compatible para un determinado colocativo. En contraposición, éste la selecciona por sus rasgos semánticos, que son compartidos por otros sustantivos en el caso *verbo* + *sustantivo*, otros verbos para el tipo *adverbio* + *verbo* o bien sustantivos cuando se trata del caso *sustantivo* + *adjetivo.* Este enfoque también está presente en (Serra, 2 009), quien defiende la idea de que los grupos que se aprecian en el conjunto de bases dado un colocativo corresponden a casos de la relación *predicado - argumentos*. El DCECR, utilizado como conjunto de ensayo en este trabajo, presenta las propiedades combinatorias de verbos, adjetivos y adverbios de forma intencional y extensional. La información que proporciona este diccionario, correspondería a los predicados, y no se limita a la pura enumeración de combinaciones de uso más o menos frecuente, sino que se muestran grupos de argumentos integrantes de una clase semántica, junto con los rasgos semánticos que comparten. La frecuencia en este caso se considera irrelevante. Si se consulta en el DCECR sobre los verbos *sentir* y *palpitar* se encuentra la información en la Tabla 3-2 y la Tabla 3-3 –aparecen resaltados en negrilla los casos detectados como colocaciones según los criterios estadísticos que se han aplicado en el capítulo precedente de este trabajo.

Forma Canónica	Grupos
sentir	*náusea, escalofrío, hambre.*
	roce, caricia.
	estómago, garganta, pierna.
	pánico, terror, horror.
	alivio, cansancio, presión, angustia.
	compasión, lástima.
	envidia, desprecio.
	simpatía, admiración, orgullo, respeto, confianza, felicidad, alegría.
	remordimiento, vergüenza.
palpitar	*pulso, latido.*
	compás, ritmo.
	sien, vena, vientre, entraña, mejilla, cuello.
	motor.
	miedo, esperanza.
	placer, pasión.
enérgicamente	*protestar, decir.*
abundantemente	*proveer, producir, surtir.*
	llorar, sudar.
	comer, beber.
brusco	*gesto, movimiento.*
	modo, manera.
doloroso	*experiencia, momento, vida, días.*
	corazón, alma.

Tabla 3-1 Colocaciones agrupadas por la semántica

Identificador	Rasgo Semántico	Clase
(I)	SUSTANTIVOS QUE DESIGNAN SENTIMIENTOS DE INCLINACIÓN APRECIO PREFERENCIA SATISFACCIÓN Y OTROS ANÁLOGOS	***admiración,*** *afecto, amor, aprecio, cariño, consideración, debilidad, estima,* ***orgullo,*** *predilección,* ***respeto,*** *simpatía*
(II)	SUSTANTIVOS QUE DENOTAN PASIÓN O DEVOCIÓN	*adoración, delirio, devoción, fascinación, locura, pasión, veneración*
(III)	SUSTANTIVOS QUE DENOTAN ATRACCIÓN HACIA ALGO O ALGUIEN O DESEO DE CONOCER O EXPERIMENTAR ALGUNA COSA	*afición, atracción, curiosidad, interés, seducción*
(IV)	SUSTANTIVOS QUE DENOTAN ODIO, RECHAZO O DESEO DE ALEJAMIENTO	*aversión, desprecio, distanciamiento, fobia, odio, rechazo*
(V)	SUSTANTIVOS QUE DENOTAN TRISTEZA O CONMISERACIÓN	***compasión,*** *conmiseración,* ***lástima,*** *pena, tristeza*
(VI)	SUSTANTIVOS QUE DENOTAN PREOCUPACIÓN, INCERTIDUMBRE O RESERVA	*inquietud, miedo, preocupación, reparo, reserva, temor,* ***terror,*** *zozobra*
(VII)	SUSTANTIVOS QUE DENOTAN CULPA	*culpa, culpabilidad*

Tabla 3-2 Clases semánticas en el DCECR para el verbo sentir

Listados similares se obtienen en las entradas de los adverbios *enérgicamente* y *abundantemente*, así como el adjetivo brusco, que quedan recogidos en la Tabla 3-4, Tabla 3-6 y Tabla 3-5, respectivamente. Sin embargo, no hay entrada para el adjetivo *doloroso*.

Esta teoría resulta especialmente útil de cara a afrontar la tarea de establecer grupos semánticos entre las combinaciones recuperadas para una forma canónica dada, puesto que explica el fenómeno observado. Por una parte, determina la orientación que se debe dar a la explotación de los resultados, indica qué elemento se debe fijar, y en qué categoría gramatical establecer los grupos. Son tres los casos que se contemplan: fijar un verbo y agrupar los sustantivos con los que se combina, fijar un adverbio y agrupar los verbos a los que modifica, o fijar un adjetivo y agrupar sus sustantivos. Por tanto, se hace necesario diseñar alguna estrategia

que permita agrupar mediante criterios semánticos las bases; se trata de establecer el nexo entre el predicado y sus argumentos de forma extensional.

El problema consiste en delimitar el subconjunto formado por aquellas que responden al fenómeno de la selección léxica dentro del conjunto de combinaciones posibles para un predicado dado. Desde un punto de vista lingüístico, el subconjunto de las colocaciones se encuentra en un nivel más profundo de las restricciones combinatorias. Si se traslada este esquema a la automatización del proceso, se parte de todas las combinaciones extraídas del corpus, entre las que se incluyen grupos semánticos y en el nivel más profundo, los casos que destacan por su uso frecuente o outliers (Figura 3-1).

Rasgo Semántico	**Clase**
SUSTANTIVOS QUE DESIGNAN EMOCIONES Y SENTIMIENTOS PRINCIPALMENTE LOS DE NATURALEZA PASIONAL	*amor, deseo, emoción, pasión, pulsión*
SUSTANTIVOS QUE DENOTAN MIEDO ODIO O PREOCUPACIÓN TAMBIÉN CON ALGUNOS QUE DESIGNAN OTRAS ACTITUDES QUE EXPRESAN LA INTRANQUILIDAD MANIFESTADA CON RELACIÓN A ALGO	*angustia, ansiedad, desconfianza, **miedo**, odio, preocupación, resentimiento, temor*
SUSTANTIVOS QUE DENOTAN FUERZA O VITALIDAD	*fuerza, vida, vitalidad*
SUSTANTIVOS QUE DESIGNAN ARTES TÉCNICAS Y TENDENCIAS ARTÍSTICAS ASÍ COMO ALGUNAS DE SUS MANIFESTACIONES	*acorde, creación, cubismo, flamenco, literatura, melodía, obra, partitura, poesía, tendencia*
SUSTANTIVOS QUE DENOTAN LUGAR CASI SIEMPRE USADOS EN SENTIDO METONÍMICO CON LA INTERPRETACIÓN DE GRUPO HUMANO	*banquillo, ciudad, país*

Tabla 3-3 Clases semánticas en el DCECR para el verbo palpitar

Rasgo Semántico	Clase
VERBOS QUE DENOTAN OPOSICIÓN O RECHAZO A MENUDO MANIFESTADAS VERBALMENTE	*condenar, criticar, denunciar, descartar, oponerse,* ***protestar****, rechazar, reprobar, repudiar*
LOS VERBOS DEFENDER Y APOYAR	*apoyar, defender*
VERBOS QUE DENOTAN LA ACCIÓN DE AFIRMAR O NEGAR ALGUNA COSA	*afirmar, asegurar, desmentir, negar*
VERBOS QUE DENOTAN SOLICITUD	*demandar, exigir, pedir, preguntar, reclamar*
VERBOS QUE DENOTAN ENFRENTAMIENTO O CONFRONTACIÓN, A MENUDO NO FÍSICA	*combatir, discutir, luchar*
VERBOS QUE DESIGNAN DIVERSAS MANIFESTACIONES VERBALES CARACTERIZADAS POR LA REITERACIÓN O EL ÉNFASIS	*recalcar, reiterar, repetir*
VERBOS QUE DENOTAN AMENAZA O ADMONICIÓN	*advertir, amenazar, recomendar*

Tabla 3-4 Clases Semánticas en el DCECR. para el adverbio enérgicamente

Figura 3-1 Esquema del problema en el nivel lingüístico y en el nivel computacional.

Rasgo Semántico	**Clase**
SUSTANTIVOS QUE DENOTAN CAMBIO REFORMA O MODIFICACIÓN TAMBIÉN CON OTROS QUE DESIGNAN DIVERSOS ESTADOS SUCESIVOS DE ALTERNANCIA	*alteración, altibajo, corrección, fluctuación, mutación, oscilación, reforma, transformación, vaivén, variación*
SUSTANTIVOS QUE DENOTAN AUMENTO O ASCENSO	*alza, ascensión, ascenso, aumento, crecimiento, elevación, incremento, subida*
SUSTANTIVOS QUE DENOTAN DISMINUCIÓN O DESCENSO	*bajada, bajada, caída, derrumbe, descenso, disminución, rebaja, recorte, reducción*
SUSTANTIVOS QUE DENOTAN MOVIMIENTO, Y A MENUDO CAMBIO RÁPIDO DE DIRECCIÓN	*giro, maniobra, movimiento, quiebro, viraje, volantazo, vuelco*
SUSTANTIVOS QUE DENOTAN DETENCIÓN O VARIACIÓN REPENTINA DE VELOCIDAD	*aceleración, acelerón, desaceleración, detención, frenada, frenazo, paralización, parón*
SUSTANTIVOS QUE DENOTAN DIVISIÓN, CORTE O RUPTURA	*corte, escisión, interrupción, quiebra, rotura, ruptura, separación*
EL SUSTANTIVO GESTO Y CON OTROS QUE DESIGNAN ACCIONES Y MOVIMIENTOS CORPORALES, A MENUDO RÁPIDOS O REPENTINOS	*abrazo, ademán, brinco, cabezada, gesto, pirueta*
SUSTANTIVOS QUE DENOTAN GOLPE, CHOQUE, ENCUENTRO VIOLENTO O MOVIMIENTO IMPULSIVO O IMPETUOSO	*chapuzón, choque, embestida, empujón, encontronazo, golpe, remezón, sacudida, tirón, vapuleo, zapatillazo, zarpazo*
SUSTANTIVOS QUE DENOTAN APARICIÓN, INICIO O SURGIMIENTO DE ALGO, O DESIGNAN ALGUNAS DE LAS ACCIONES QUE LLEVAN A CABO ESOS PROCESOS	*aparición, apertura, arrancada, arranque, eclosión, irrupción, rebrote, surgimiento*

Tabla 3-5 clases semánticas en el DCECR. para el adjetivo brusco (I)

Rasgo Semántico	Clase
VERBOS DE CAMBIO DE ESTADO ESPECIALMENTE LOS QUE DESIGNAN LA ACCIÓN DE PROPORCIONAR HACER ADQUIRIR O LOGRAR QUE ALGO O ALGUIEN PASE A ESTAR EN DISPOSICIÓN DE ALGUNA COSA	*dotar, engrasar, financiar, pintar, premiar, proporcionar, suministrar, untar*
VERBOS QUE DENOTAN CONSUMICIÓN GENERALMENTE DE ALIMENTOS	*abrevar, beber, cenar, comer, consumir, leer*
VERBOS QUE DENOTAN EXISTENCIA O INCREMENTO DE ALGO	*almacenar(se), crecer, dar(se), encontrar(se), existir, extender(se), practicar(se), presentar(se), prodigar(se), producir(se), realizar(se), sobrar*
VERBOS QUE DESIGNAN EL PROCESO DE SURGIR O FLUIR UN LÍQUIDO U OTRAS MATERIAS QUE SE LE ASIMILAN FÍSICA O FIGURADAMENTE TAMBIÉN CON OTROS VERBOS QUE EXPRESAN LA ACCIÓN QUE DESENCADENA ESE PROCESO	*brotar, correr, desbordar, fluir, llover, manar, nevar, perder sangre, regar, remojar, rociar, sangrar, sudar, verter*
VERBOS QUE DESIGNAN LA ACCIÓN DE EXPRESAR O MANIFESTAR ALGO GENERALMENTE COMO MEDIOS VERBALES PERO TAMBIÉN CON OTROS	*anunciar, citar, comentar, contar, describir, dialogar, discutir, escribir, hablar, informar, mencionar*
VERBOS QUE DESIGNAN LA ACCIÓN DE PROBAR ALGO	*demostrar, probar, dar propaganda*
VERBOS QUE DESIGNAN LA ACCIÓN DE REPRODUCIR DIFUNDIR O HACER PÚBLICA ALGUNA COSA	*exhibir, filmar, grabar, ilustrar, representar, reproducir, reseñar, traducir, versionear*
VERBOS QUE DESIGNAN LA ACCIÓN DE USAR ALGO	*emplear, recurrir, usar, utilizar*

Tabla 3-6 Clases Semánticas en el DCECR. para el adverbio abundantemente.

3.2. Similitud semántica

Se propone aplicar técnicas que permitan automatizar el proceso de generación de clases léxicas; deben estar formadas por palabras que estén próximas en el espacio semántico, de manera que se puedan construir grupos formados por elementos similares en cuanto a su significado. La generación automática de grupos semánticos obliga a cuantificar las relaciones semánticas existentes entre las palabras del texto que se analiza. Se debe hacer uso de medidas que reflejen la relación entre palabras que representan un mismo concepto, y así conformar grupos que hagan referencia a las mismas entidades conceptuales.

El problema de cuantificar las relaciones semánticas ha sido ampliamente estudiado por su utilidad en distintas tareas del Procesamiento del Lenguaje Natural. Se emplean diferentes términos concernientes a métricas que evalúan tanto la existencia de una relación semántica entre dos palabras como su fuerza. En este sentido se pueden encontrar en la literatura referencias a *relación*, *similitud* o *distancia semántica*.

3.2.1. Técnicas para determinar la similitud semántica

Desde el punto de vista algorítmico, los métodos comparan los términos en la representación topológica de algún recurso léxico que incluya relaciones semánticas, o bien se analizan contextos en los que aparecen –en muchas ocasiones se representan mediante vectores.

En cualquier caso, una medida de similitud debe verificar los siguientes requisitos (Lin, 1 998):

i. Cuando las dos unidades son idénticas, el valor de la similitud debe ser máximo.
ii. Dos unidades son más similares cuantos más elementos comparten. Por tanto, la similitud está relacionada con los elementos en común.
iii. Dos unidades son menos similares cuantas más diferencias haya entre ellos. La similitud está relacionada con las diferencias entre ellas.

3.2.1.1. Métodos basados en Bases de Conocimientos léxicas.

Se utilizan bases de conocimiento léxicas, ya sean diccionarios, tesauros, ontologías o redes semánticas: algún recurso léxico con relaciones semánticas bien de forma explícita o bien que se puedan inferir como en el caso de los diccionarios convencionales.

Kozima y Furugori (1 993) evalúan la similitud semántica a partir de una red neuronal en la que los nodos son entradas del diccionario. Dos entradas tienen relación si una de ellas se usa para definir a la otra. Sobre esta red se calcula la similitud a partir de la activación de los nodos.

Otros autores como Morris y Hirst (1 998) utilizan un tesauros, que agrupa las palabras sin especificar qué tipo de relación motiva el agrupamiento. Dos palabras se consideran cercanas semánticamente si comparten alguna categoría, están en la misma subcategoría, etc.

Existe un grupo importante de medidas de similitud que toman como espacio topológico en el que proyectar las palabras a *WordNet* –red semántica de la lengua inglesa. Se agrupan las palabras en conjuntos de sinónimos o *SynSets*, que proporcionan definiciones cortas y generales, y almacenan las relaciones semánticas entre estos conjuntos de sinónimos. De esta base de datos también se ha creado una versión multilingüe para varios idiomas europeos (Holandés, Italiano, Español, Alemán, Francés, Checo y Estonio) llamada *EuroWordNet*.

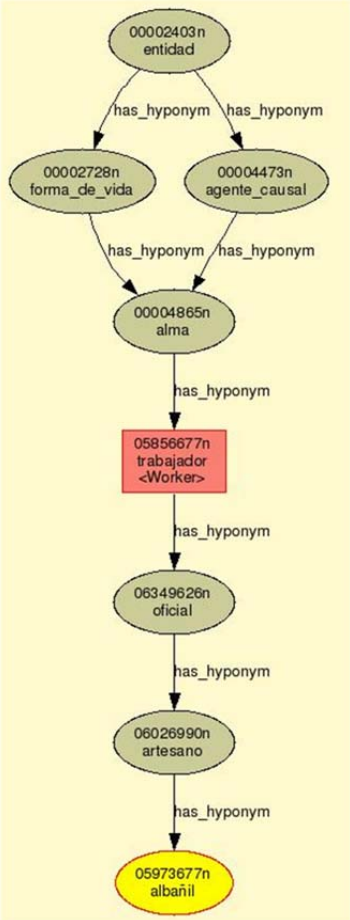


Figura 3-2 Grafos de las consultas albañil y empresa en EuroWordNet.

Si bien este grupo de técnicas considera la similitud semántica en función de la longitud del camino que las une en el grafo que representa la red, esta aproximación es demasiado simple. Considerar exclusivamente el número de arcos que unen los synsets equivale a asumir que todas las taxonomías registradas son equidistantes, sin embargo, la estructura de la red no cumple con tal suposición. Es necesario determinar pesos que reequilibren su naturaleza no homogénea, de forma que los parámetros a contemplar incluyan tanto la longitud del camino

como la especificidad de los sentidos de las palabras en el camino (Agirre, 1995). Este problema se ilustra en (Resnik, 1999) en donde se presentan los enlaces:

Rabbit Ears is-a Television Antenna, o

Phytoplankton is-a Living Thing.

Para resolver este problema existen diferentes propuestas que incorporan datos adicionales, como la profundidad del concepto en la jerarquía. A profundidades mayores se deben obtener mayores puntuaciones; también deben ser más próximos semánticamente hablando aquellos que se encuentren en zonas de mayor concentración de conceptos frente a los que estén en zonas dispersas (Agirre, 1995).

3.2.1.2. A partir de los contextos de uso

El segundo grupo de técnicas para la detección de similitud semántica se basa en la información que se puede extraer de las palabras a partir de los contextos en que se emplean. Su uso exige como requisito indispensable disponer de corpus extensos para que las medidas sean fiables y precisas.

Se evalúa la relación de las palabras mediante distancias que representan la similitud o la fuerza de la relación existente –se basan en la frecuencia de aparición conjunta en los textos. En general, la idea subyacente es que la relación semántica es más fuerte en aquellos grupos de palabras cuyos contextos estén próximos.

Se proyecta la información lingüística sobre un espacio vectorial N dimensional, habitualmente vectores de coocurrencias en el corpus. Sobre dicho espacio se calculan distancias geométricas entre sus elementos: distancia euclídea, distancia de Manhattan, distancia del coseno, etc. Autores como Manning y Schütze (1999) indican una representación de vectores binarios, a los que se pueden aplicar la distancia de Dice, Jaccard, etc.

Además de poder aplicar las distancias anteriores, cuando los vectores representan distribuciones de probabilidad de los objetos lingüísticos en el corpus, surgen otras distancias que explotan las propiedades inherentes a las distribuciones de probabilidad: la Divergencia de Jensen-Shannon, el coeficiente de correlación de Pearson, la divergencia de Kullback-Leibler, o el coeficiente de τ-Kendall, entre otras.

3.3. Obtención de clases léxicas

Se pretende construir los grupos semánticos a partir de los datos registrados en la BDD de combinaciones léxicas. El problema que se pretende resolver sobre nuestros datos es el

siguiente*: dada una forma canónica, agrupar en clases semánticas el conjunto de formas canónicas con las que aparece en alguna de las estructuras admisibles para colocaciones del español*. Por ejemplo, dada una forma canónica con categoría gramatical verbo, se deben obtener, en caso de que existan, grupos dentro de la bolsa de sustantivos con los que aparece en las combinaciones *verbo + sustantivo*. El criterio que regirá su formación es la similitud semántica entre sus miembros. Otra posibilidad, que consideraremos de forma independiente sería realizar agrupamientos entre los adverbios con los que se ha registrado en el corpus.

3.3.1.Agrupamiento mediante técnicas estadísticas

En este epígrafe revisamos dos técnicas estadísticas que permiten generar grupos semánticos, analizando si su aplicación sobre los datos que manejamos es recomendable. Por una parte se presenta el análisis clúster, dentro de las técnicas convencionales de clasificación estadística, y por otro el análisis de semántica latente.

3.3.1.1. Análisis clúster

La generación de grupos semánticos o clases léxicas puede resolverse a partir de técnicas basadas en análisis *clúster*, tradicionalmente utilizadas para la creación de clasificaciones. Se conoce como *análisis clúster* al conjunto de técnicas multivalentes diseñadas para extraer grupos o *clústeres* de un conjunto de observaciones de un conjunto de variables aleatorias. Estos métodos dividen un conjunto de objetos en grupos o clústeres con cohesión interna, es decir, los objetos en un mismo grupo serán muy similares entre sí. Por otra parte, también se impone la restricción de aislamiento externo del grupo, es decir, que los elementos en clústeres distintos sean muy diferentes.

En esta técnica se hace fundamental la selección de la medida de similitud, siendo importante que se determine en función de una distancia $d: \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^+$ debe verificar (Rodríguez, 2003):

$$d(\overrightarrow{x_i}, \overrightarrow{x_j}) \geq 0$$
$$d(\overrightarrow{x_i}, \overrightarrow{x_i}) = 0$$
$$d(\overrightarrow{x_i}, \overrightarrow{x_j}) = d(\overrightarrow{x_j}, \overrightarrow{x_i})$$
$$d(\overrightarrow{x_i}, \overrightarrow{x_j}) \leq d(\overrightarrow{x_i}, \overrightarrow{x_k}) + d(\overrightarrow{x_k}, \overrightarrow{x_j})$$

Generalmente, los métodos de análisis clúster se clasifican en *jerárquicos* o no *jeráquicos*. En el primer caso se presupone una estructura jerárquica en la clasificación, además los resultados son muy sensibles a los outliers. Dadas las características de la distribución de las combinaciones en las que interviene una forma canónica, totalmente asimétrica y con presencia de valores extremos hace que se descarten estas técnicas en este problema. En el segundo, no se

impone ningún tipo de estructura de antemano, pero se debe prefijar el número de clases que se van a generar. Lo más común en estos casos es aplicar algún método de reasignación: se asignan individuos a una configuración inicial de grupos y a lo largo de las sucesivas iteraciones se permite reasignar individuos a grupos, de manera que se optimice el criterio de selección; el proceso acaba cuando no existan objetos que al ser reasignados provoquen mejoras. Es necesario por tanto, determinar algún mecanismo de evaluación de la bondad de la configuración de los grupos que permita establecer la reasignación de individuos orientada a la optimización. Con este fin se define el *centroide* del clúster como representante de la clase en cada iteración, consistiendo en el vector formado por la media aritmética de los valores en el grupo de cada dimensión. Cada componente del *centroide* del grupo *i* es el promedio de los valores de la variable correspondiente a esa componente para todos los elementos que conforman dicho grupo. Es decir:

$$\overrightarrow{c_l} = \frac{1}{n_l}\sum_{i=0}^{n_l}\overrightarrow{x_l}$$

Cuánto de bueno es el agrupamiento se mide a partir de este vector, los elementos del grupo tienen que tener distancias pequeñas al *centroide* y los grupos deben estar lo más diferenciados posible, lo cual se puede lograr maximizando las distancias entre *centroides*. En cualquier caso, se hace necesario medir las distancias entre los objetos que se pretende agrupar, para lo que se dispone de diferentes funciones tanto para variables cuantitativas como categóricas.

En este caso se establece como contexto de una forma canónica dada, todas las formas canónicas con las que aparece en alguna combinación en el corpus. La Tabla 3-7 ilustra los contextos de algunos sustantivos que se revelan como colocaciones del verbo *sentir.* Cada entrada corresponde a la frecuencia de coocurrencia de las dos formas canónicas, también se pueden utilizar las de frecuencias relativas. En el ejemplo seleccionado, la construcción de la matriz de contextos conlleva extraer todas las bases del verbo *sentir,* éstas constituirán las filas, es decir, los objetos a clasificar. Las columnas, o variables utilizadas para clasificarlas, corresponden a cada uno de los verbos o colocativos con los que aparecen esas bases. Algoritmos de análisis clúster sobre esta matriz permiten obtener agrupamientos bajo el principio de que sustantivos similares se usan en contextos similares: los grupos se construyen bajo la premisa de que *sustantivos similares combinan con verbos similares*.

Sustantivo	*abajar*	*abalanzar*	*abandoner*	*abarcar*	*abatir*	*abismar*	*ablandar*	*abordar*	…
escalofrío	12								
miedo	79	11	322	4	38	29	4	6	
náusea			6						
terror	34	3	111		9	35			

Tabla 3-7 Extracto de vectores de contexto de algunos sustantivos en las colocaciones sentir + Sust.

La matriz de vectores de contexto del verbo *sentir*, restringida a los sustantivos que se detectan como colocaciones desde el punto de vista frecuentista, es de 12545×7494; cuando se trata del verbo *palpitar*, la matriz es de rango 661×6955. Si se restringen los casos a aquellos que aparecen al menos 10 veces en el corpus se obtiene una matriz de 8285×5606 y 224×4955 respectivamente.

La construcción de dicha matriz a partir de la BDD es inviable en tiempo real, para cada caso es necesario extraer la bolsa de palabras base y cada una de las formas canónicas que combinan con ellas, luego se asignará la frecuencia de coocurrencia en caso de que la combinación exista en el corpus o 0 en caso contrario. Además se aconseja aplicar alguna técnica de reducción de dimensiones para poder aplicar los algoritmos de análisis clúster, en caso contrario sobre matrices tan extensas sería muy costoso computacionalmente hablando. Nuevamente surgen las restricciones marcadas por el gran volumen de datos que se maneja, lo que obliga a elegir una técnica viable desde el punto de vista del tratamiento computacional. Dado que el objetivo es obtener clases léxicas se opta por valorar el Análisis de Semántica Latente, que es una técnica de reducción de dimensiones capaz de extraer conceptos y agrupar términos en torno a ellos.

3.3.1.2. Análisis de Semántica Latente

El Análisis de Semántica Latente (LSA, del inglés Latent Semantic Analysis) utiliza la representación de contextos de términos o palabras para determinar la similaridad de significados de palabras y conjuntos de palabras (Landauer, T. K., Foltz, P.W., Laham, 1998). Esta técnica, diseñada para aplicarse a grandes corpus textuales, revela conceptos implícitos en los datos, y permite representar relaciones semánticas entre documentos, palabras o palabras con documentos. En este caso documentos hace referencia al contexto que se considere para cada término o palabra.

En LSA se parte de la matriz de términos-documentos o matriz de contextos, en la que cada fila representa las frecuencias de aparición de un término en el contexto (documento)

correspondiente. Cada columna por tanto representa un documento o contexto, y los valores en la columna la frecuencia de aparición de cada uno de los términos contenidos en él. Los elementos que se tomen como contextos pueden variar: pueden ser concordancias extraídas para una palabra, frases, párrafos o documentos enteros (Lancia, 2007). En cualquier caso, serán considerados como documentos cuando se aplica el algoritmo y los términos hacen referencia a las palabras que se analizan. En la presente apuesta, los documentos corresponden a las líneas de concordancia simuladas por la aplicación desarrollada para recuperar la información que se ha registrado en la BD y los términos son las formas canónicas registradas en ella. De esta manera, dada una FC se puede reconstruir la matriz de contextos que se obtendría. Las combinaciones en la Tabla 3-7 sería el punto de partida si se quiere obtener los conceptos o grupos de palabras entre las bases del verbo *sentir*. La extracción de conceptos se realiza aplicando la descomposición en valores singulares (SVD, del inglés Singular Value Decomposition) de la matriz de contextos.

3.3.1.3. Descomposición SVD

La descomposición SVD se aplica a matrices rectangulares, dada una matriz $A_{m\times n}$ con $m > n$ se determinan las matrices S_1, S_2, Σ tales que:

$$A = S_2 \Sigma S_1^T$$

La matriz Σ es la matriz diagonal con los autovalores de $A^T A$, existe un índice *r* para el que las raíces cuadradas de los primeros *r* autovalores ordenados decrecientemente: $\sigma_1, \sigma_2, \cdots, \sigma_r$ son positivos, y el resto es cero, generalmente coincide con *n*. Este conjunto de valores se denominan **valores singulares**. S_2 es la matriz de autovectores de $A^T A$ y S_1 la de autovectores de AA^T, cada entrada de $A^T A$ es el número de palabras en común entre los contextos *i, j*. Cada entrada de AA^T es el número de contextos en los que aparecen conjuntamente las palabras *i, j*. La descomposición de *A* se denomina **descomposición en valores singulares**. Para reducir la dimensión, a *k* factores o conceptos ($k<n$) se hacen 0 los correspondientes valores singulares, es decir, los autovalores $\sigma_{k+1}, \sigma_{k+2}, \cdots, \sigma_r$ se obtiene una aproximación de *A* utilizando la nueva versión de Σ para su reconstrucción (de Jorge Botana, G., 2010):

$$A \approx A_k = S_{2k} \Sigma_k S_{1k}^T$$

La elección de *k* se realiza empíricamente, estableciendo qué valores singulares: $\sigma_r, \sigma_{r-1}, \cdots, \sigma_{k+1}$ se consideran despreciables respecto a los restantes *k*. Al estar ordenados los valores singulares de mayor a menor, se considerarán las *k* primeras filas en la matriz. Estas *k*

dimensiones seleccionadas representan los diferentes conceptos o grupos que se extraen de los datos.

En la nueva matriz reconstruida, cada entrada de coocurrencias observadas es sustituida por valores estimados que marcan las diferencias entre las clases de palabras, de forma que las puntuaciones que obtienen serán mayores o menores, incluso se infiere la relación de palabras con contextos que no son plausibles en la matriz original, pasando de tener frecuencia de coocurrencia 0 a valores mayores que 0.

Se ha aplicado LSA para clasificar las bases que combinan con el verbo *perpetrar,* que tiene un único significado, *palpitar* con 4 acepciones en el DRAE y el verbo *sentir*, que es polisémica con numerosas acepciones y todas ellas con distintos rangos de frecuencia en el corpus. Se ha utilizado el paquete *lsa* del software estadístico R (2012).

3.3.1.4. Semántica Latente en las bases de perpetrar:

En el corpus se encuentran 23 sustantivos que combinan con el verbo *perpetrar*, la matriz de contextos es de dimensión 23×4046.

Los autovalores que se obtienen se listan a continuación, su representación gráfica ayuda a elegir la cantidad de dimensiones, que se traduce en los conceptos o clases léxicas que se extraen de las bases:

```
[1] 1367.100581  552.573877  274.523182  242.149528  209.576341  185.234976
[7]  179.162088  165.281073  164.306348  149.203851  144.902200  140.766047
[13]  138.140623  131.904095  114.926038  111.266192   88.563222   60.356442
[19]   37.929552   20.078836    4.319880    1.031975
```

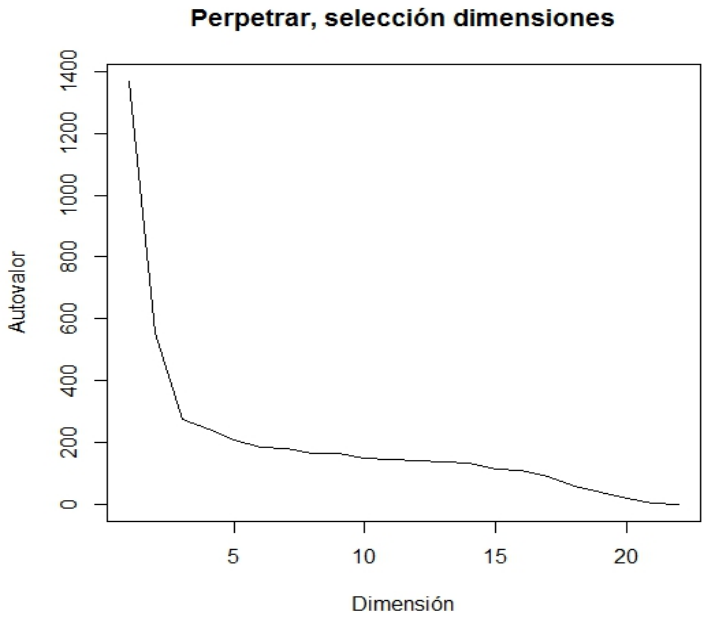


Gráfico 3-18 Autovalores en la descomposición SVD del verbo perpetrar

Se observa el peso diferenciado de la primera dimensión respecto a las restantes. En el Gráfico 3-19 se muestra la representación de los términos en el espacio semántico latente, utilizando la matriz: $S_{2k}\Sigma_k$, se seleccionó *k=2*. Se observa el grupo. *acto, asesinato, ataque* y el resto de sustantivos como *masacre, delito, robo*, etc. que se ven entremezclados con sustantivos de uso muy frecuentes como *hombre, mujer*.

3.3.1.5. Semántica Latente en las bases de palpitar:

Se repite el experimento sobre las bases del colocativo *palpitar*, la construcción de la matriz de contextos consume un intervalo de tiempo que se mide en horas. Aún así se ha procesado la matriz de contextos generada por este verbo, los resultados se muestran en el Gráfico 3-20.

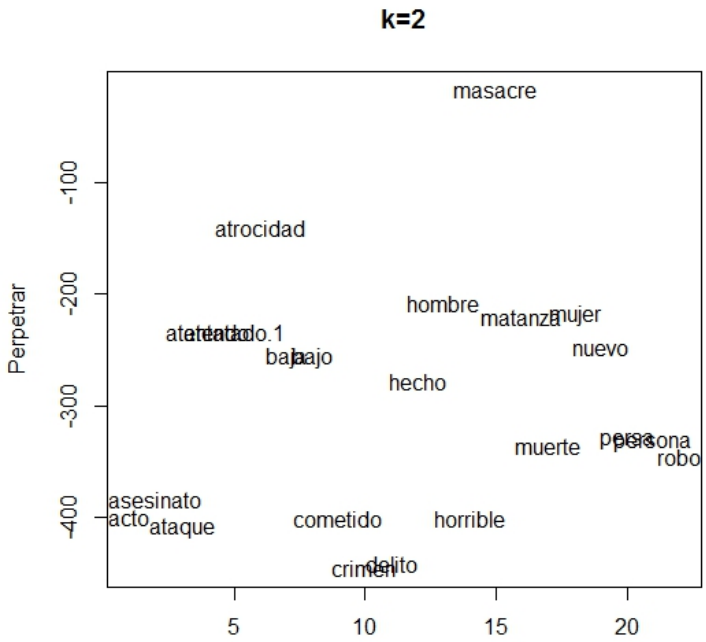


Gráfico 3-19 Representación de los términos en el espacio semántico latente del verbo perpetrar.

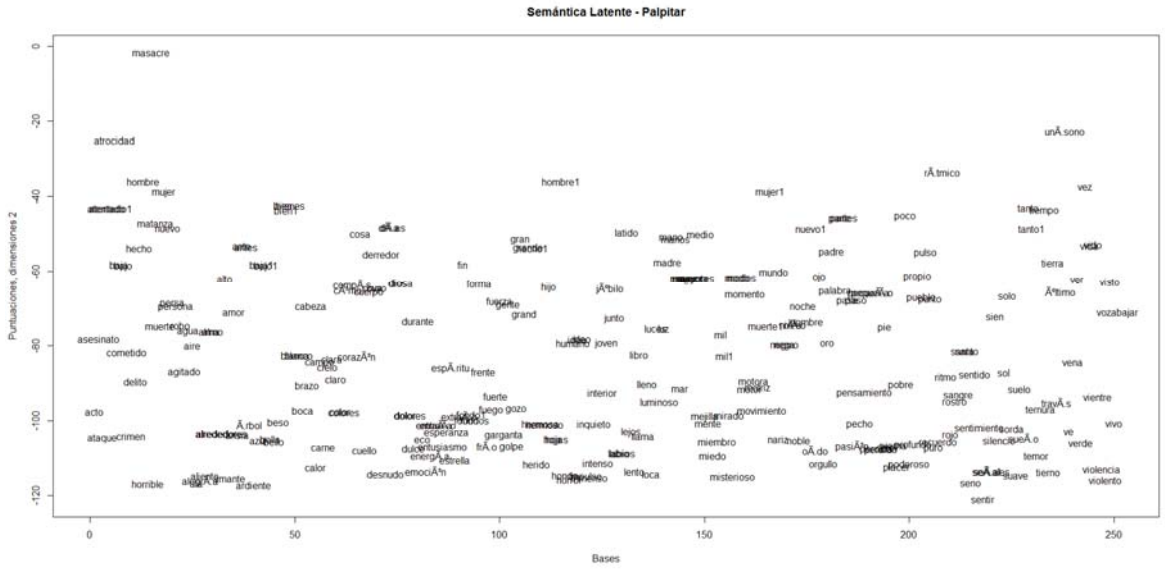


Gráfico 3-20 Representación de los términos en el espacio semántico latente del verbo palpitar.

Las formas canónicas cercanas en este gráfico reflejan similitud semántica (Jorge-Botana, G. *et al*., 2010), (Olmos, R. *et al*, 2014); para apreciar los resultados se ha realizado un zoom del Gráfico 3-20 que se presenta en 3 piezas separadas en zonas de izquierda a derecha del gráfico original: Gráfico 3-21, Gráfico 3-22 y Gráfico 3-23. Sobre ellos se han marcado conjuntos de bases cuya cercanía en el espacio semántico latente revela cierta similitud, o más bien relación semántica. La definición en el DRAE del verbo *palpitar* pone de manifiesto su vínculo con sustantivos relacionados con *corazón* y *afectos* o *pasiones*, también es lógico encontrar

fuertes vínculos con sustantivos *referidos a alguna parte del cuerpo*. Los grupos que se han marcado reflejan el mismo problema que se detectó en el análisis de *perpetrar*, bases de frecuencia muy alta como *hombre, mujer, cosa, tiempo, vez, ...* se entremezclan con elementos en los grupos en los que hay una motivación semántica en su configuración. Se han marcado 5 grupos:

G1: *ataque, crimen, asesinato, muerte, delito, matanza, hombre, mujer, hecho,...*
G2: *boca, brazo, cuello, corazón, ...*
G3: *espíritu, esperanza, entusiasmo, emoción, fuerte, fuego, gozo, impulso, dolor, ...*
G4: *unísono, rítmico, tiempo, vez, ...*
G5: *pasión, orgullo, placer, ternura, violencia, profundo, puro, ...*

G3 se ajusta al grupo del DCECR:

SUSTANTIVOS QUE DENOTAN FUERZA O VITALIDAD.

G4 se ajusta al grupo del DCECR:

SUSTANTIVOS QUE DESIGNAN EMOCIONES Y SENTIMIENTOS PRINCIPALMENTE LOS DE NATURALEZA PASIONAL.

DRAE palpitar.

(Del lat. *palpitāre*).

1. *intr.* Dicho del corazón: Contraerse y dilatarse alternativamente.
2. *intr.* Dicho del corazón: Aumentar su palpitación natural a causa de una emoción.
3. intr. Dicho de una parte del cuerpo: Moverse o agitarse interiormente con movimiento trémulo e involuntario.
4. *intr.* Dicho de algún afecto o pasión: Manifestarse vehementemente. *En sus gestos y palabras palpita el rencor.*

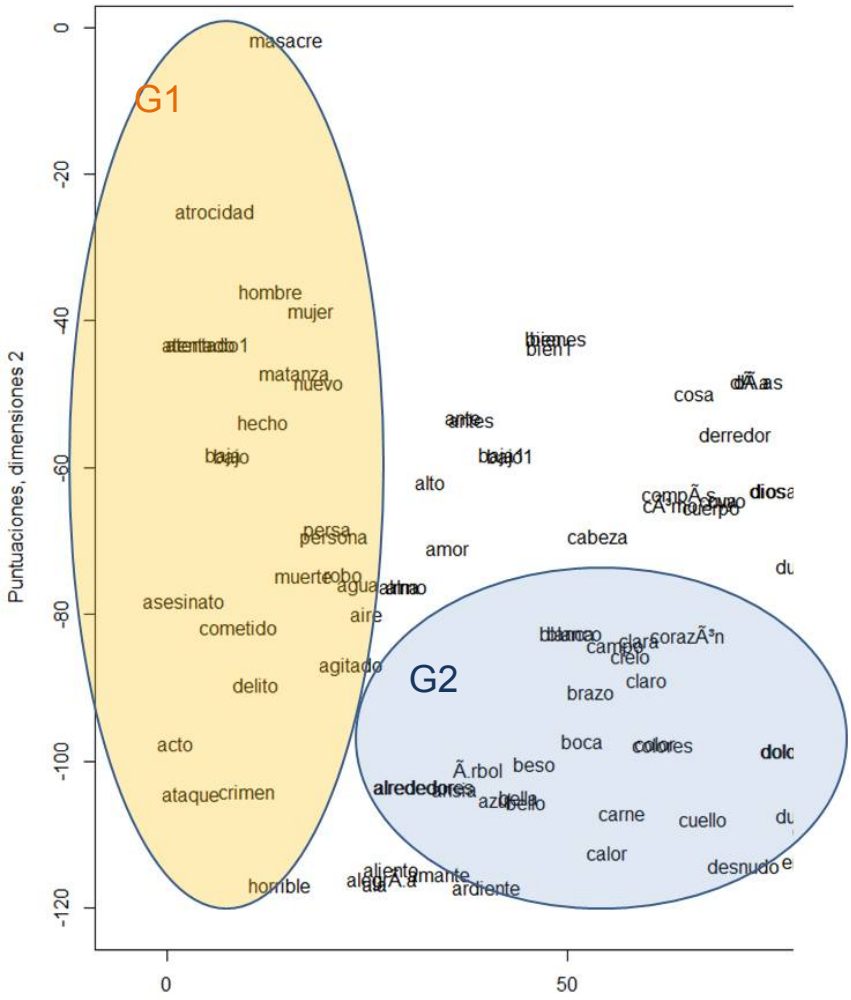


Gráfico 3-21 Zoom 1, resultados de palpitar.

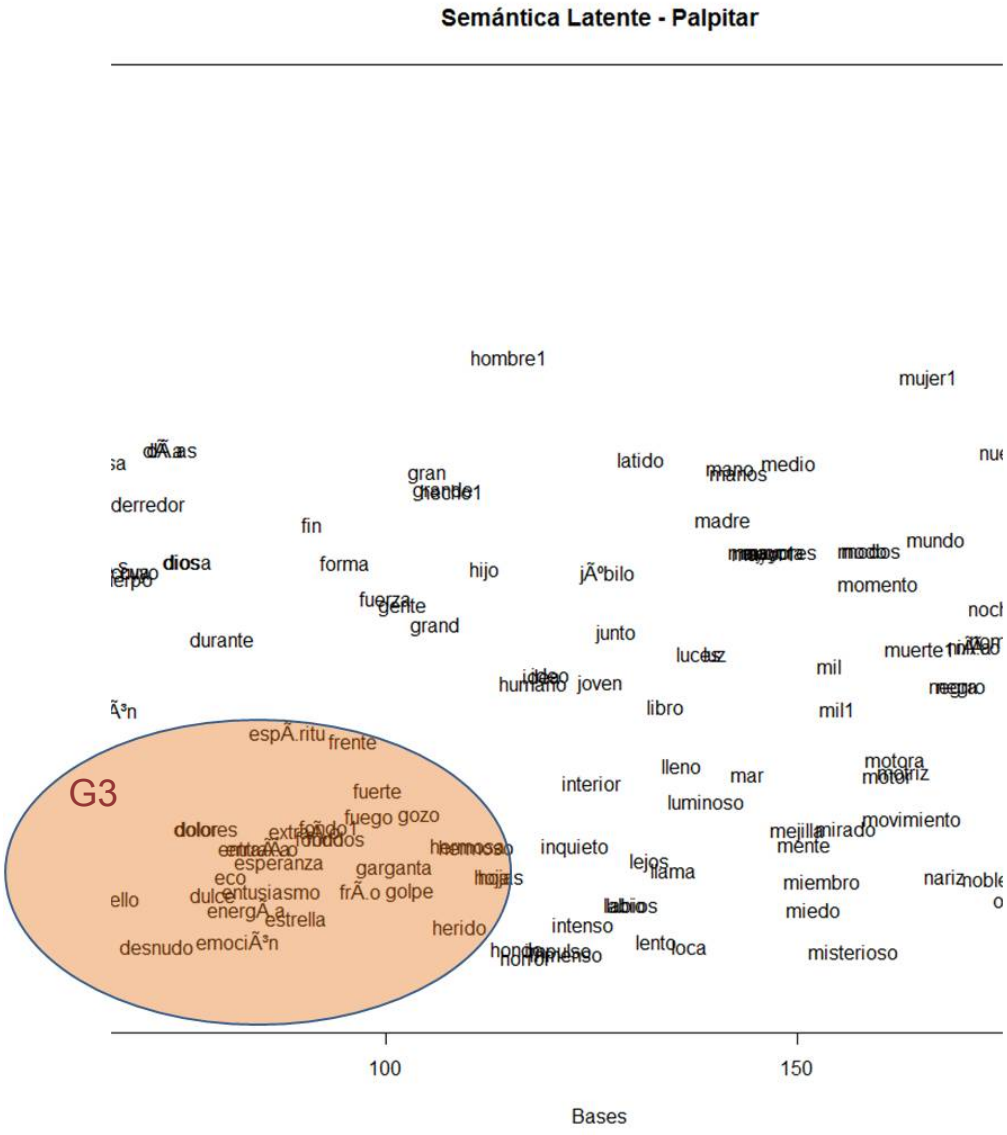


Gráfico 3-22 Zoom 2, resultados de palpitar

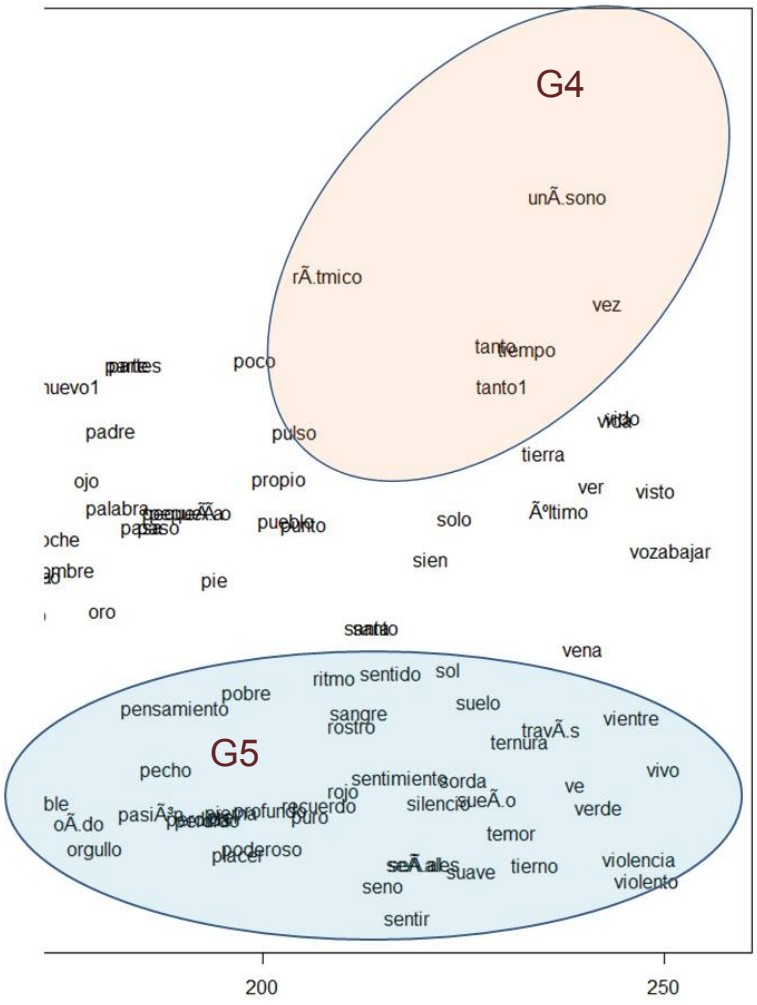


Gráfico 3-23 Zoom 3, resultados de palpitar

Si bien en los grupos obtenidos, el método del análisis de la semántica latente resulta adecuado, hay varias cuestiones que no aconsejan aplicarlo sobre los registros extraídos del corpus. Por una parte, generar la matriz de datos que se requiere consume un tiempo inasumible en una aplicación de consulta de información de características combinatorias del español. Por otra parte, los resultados conllevan la supervisión manual para establecer la configuración de grupos.

3.3.2.Obtención de grupos semánticos a partir de la relación predicado-argumentos

La obtención de clases léxicas mediante las técnicas expuestas anteriormente requiere un tiempo de cómputo y recursos de memoria excesivos, además, se hace necesaria la toma de decisiones como el número de dimensiones o conceptos presentes, para lo que no hay un criterio claro. Abundando en los aspectos negativos, no quedan claramente marcados los grupos, por ejemplo, en el caso de *perpetrar* (Gráfico 3-19) se observan muy próximas las formas canónicas

hombre-mujer, *nuevo* y *matanza*, o *persona* y *robo*. También se suma a estas dificultades la necesidad de reconstruir en cada caso la matriz de contextos a partir de los datos registrados en la BD. Por esta razón se acomete la solución del problema a partir del modelo lingüístico de la **selección predicado-argumentos**, mediante las **preferencias de selección**.

En algunos trabajos de desambiguación semántica de palabras se utilizan medidas de las relaciones entre un predicado y sus argumentos. En Resnik (1 999) se define como asociación de selección una medida entre verbos y sustantivos que se formaliza según la teoría de la información. Utiliza probabilidades que se estiman a partir de las coocurrencias de un predicado dado con los elementos de una determinada clase de sustantivos. Las clases están formadas por los elementos en synsets de *WordNet* que contienen hipónimos de un sustantivo o por synsets correspondientes a hiperónimos del mismo. También en Agirre (1 994) se evalúa una medida similar, aunque establece la asociación entre clases de predicados con clases de argumentos.

El método de Resnik requiere disponer de un recurso léxico que provea de clases semánticas ya establecidas, para tal fin se propone utilizar algún Diccionario Ideológico del español como base de conocimiento léxico contrastada por los lingüistas.

3.3.3.Diccionarios Ideológicos, bases de conocimientos léxicas.

Un diccionario ideológico agrupa las palabras en torno a una idea, según un orden conceptual. Representan, por tanto, una fuente de información sobre relaciones semánticas similares a los tesauros. En español se dispone del Diccionario Ideológico Vox (DIV) (Vox, 1 995) y del Diccionario Ideológico de la Lengua Española de Julio Casares (DILE) (Casares, 1 959). Ya que se propone el uso de este tipo de recursos en la extracción de preferencias de selección en sustitución de *WordNet* o su equivalente *EuroWordNet* en español, en los siguientes epígrafes se describe la estructura de ambos diccionarios.

3.3.3.1. Estructura del Diccionario Ideológico Vox.

El Diccionario Ideológico Vox se estructura en varios niveles diferentes (Figura 3-3):

Cuadro General, recoge los grandes grupos de ideas: Relaciones Generales, Materia, Naturaleza, Seres Vivos, Individuo y Modos de Vida.

Parte Sinóptica: Incluye los cuadros, corresponde a otro nivel de agrupación y representan identificadores de títulos de grupos conceptuales (Tabla 3-8).

Parte Analógica: Aparecen todas las palabras ordenadas en grupos según un criterio conceptual. Dentro de los grupos pueden aparecer subgrupos, motivados por una mayor afinidad entre sus componentes y también marcadores que representan relaciones entre

grupos y subgrupos, sin especificar el tipo. Los grupos a su vez se dividen en bloques por categoría gramatical de los elementos (Tabla 3-9). Se presentan las palabras como se asocian en la mente del hablante (Vox, 1 995).

Cuadro	**Grupo**
dolor	*cantar, doler, entortijarse, entrepunzar, latir, punzar, rabiar, suposición, traspasar, ver las estrellas, sentir*
juicio	*cantar, doler, entortijarse, entrepunzar, latir, punzar, rabiar, suposición, traspasar, ver las estrellas, sentir*
juicio	*conceptualizar, conceptuar, conocerse, considerar, contar, creer, dar por, decir, encasillar, entender, estimar, juzgar, opinar , pasar, reputar, sentir, tenerse, tomar por, ver, reconocerse*
sentimiento	*afección, corazón, entretelas, podredumbre, psicología, sentir, sentimiento*
sentimiento	*acoger, cobrar, coger, concebir, dar, morir, anidar*
tristeza	*deplorar, llorar, sentir*
balbuceo	*declamar, monologar, recitar, responder, sentir*

Tabla 3-8 Cuadros y grupos en los que aparece sentir

Dolor
acritud, analgesia, clavo, dolor, puntada, punzada, ramalazo, rayo, sinalgia, tormento, yaya
agujetas, ciática, entuertos, ijada, lumbago, mastalgia, mialgia, ostealgia, retortijón
cálculos, cólico, cólico nefrítico, cólico renal, litiasis, mal de piedra
arenillas, bezoar, cálculo, piedra
ardor, comecome, comezón, escozor, fogaje, fuego, hormiguilla, hormiguillo, picazón, picor, piquiña, prurito, quemazón, rascazón
agudo, dolorido, doloroso, lacerante, lancinante, pedregoso, pruriginoso, terebrante, urente
indolente, indoloro
cantar, entortijarse, entrepunzar, latir, punzar, rabiar, sentir, traspasar, ver las estrellas
abrasar, comer, enardecerse, encender, escaldarse, escarabajear, irritar, picar, quemar

Tabla 3-9 Grupos en el Cuadro dolor

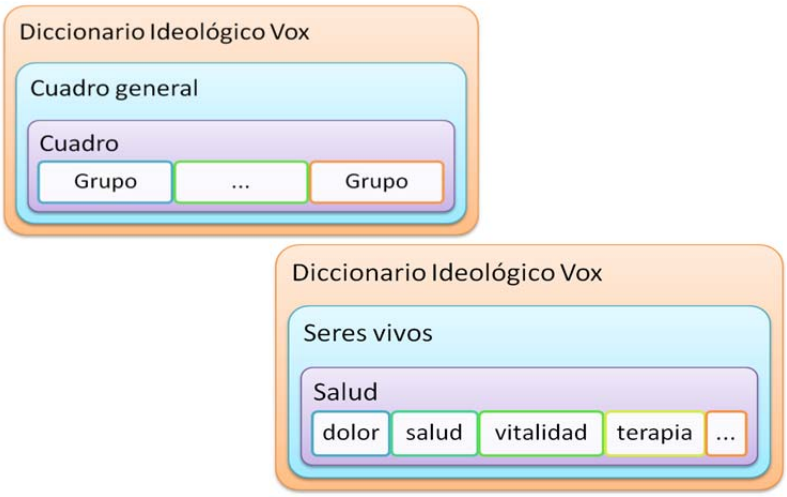


Figura 3-3 Estructura del Diccionario Ideológico VOX

3.3.3.2. Diccionario Ideológico Casares

En la parte sinóptica de este diccionario, se establecen los distintos niveles de agrupamiento que se emplean. En un primer nivel se tienen las grandes materias que estructuran la clasificación (Figura 3-4), donde se encuadran las 38 categorías básicas en las que se distribuyen las palabras; estas categorías se denominan cuadros y se subdividen a su vez en 2000 grupos de palabras conceptualmente relacionadas (Tabla 3-10).

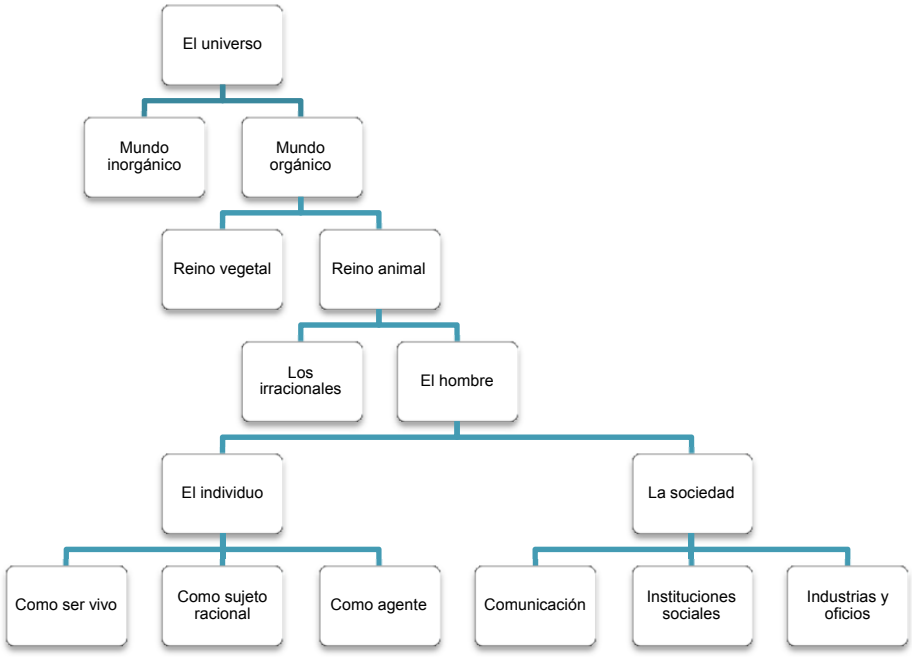


Figura 3-4 Grandes materias, Diccionario Ideológico Casares

Cuadro	Cuadro
Religión	*Colocación*
Física y Química	*Tiempo*
Geografía, Astronomía, Meteorología	*Cantidad*
Geología, Mineralogía	*Intelección*
Botánica	*Apreciación, juicio*
Zoología	*Voluntad*
Anatomía	*Conducta*
Fisiología	*Acción*
Medicina	*Lenguaje*
Alimentación	*Arte*
Vestido	*Estado, Nación*
Vivienda	*Costumbres*
Sensibilidad, Sentidos	*Derecho y Justicia*
Sentimientos	*Propiedad*
Existencia, Cambio	*Milicia*
Relación, Orden, Causalidad	*Comercio, Banca y Bolsa*
Espacio	*Agricultura, Zootecnia*
Forma	*Zootecnia*
Movimiento	*Transportes*

Tabla 3-10 Cuadros de la clasificación del DIC .

En la parte analógica se enumeran los grupos, divididos en subgrupos, y encabezados por un término, normalmente sustantivo, que corresponde con los que se detallan en los diferentes cuadros. Puede incluir remisiones a otros grupos relacionados conceptualmente a través de elementos marcados en negrilla para indicar que son otras cabeceras. También se diferencian subgrupos según la categoría gramatical de sus componentes: sustantivos, verbos, adjetivos, adverbios, modos adverbiales, preposiciones e interjecciones.

En la parte alfabética se recogen los significados de todas las palabras que aparecen en la parte analógica y se muestran referencias a los grupos a los que pertenece.

Cuadro	Encabezado	Grupo
Conducta, lenguaje	*ademán*	*accionar, amanerarse, chistar, cocar, enzainarse, expresar, gestear, gesticular, gesticular, guiñar, guiznar, hablar con los ojos, hablar por señas, hacer cocos, hacer del ojo, hacer figuras, hacer telégrafos, momear, pujar, remilgarse, sentir, timarse, torcer el gesto*
Sentimientos	*aflicción*	*ahogarse en poco agua, añorar, apurarse, atravesarse un nudo en la garganta, atribularse, caerse, cariñar, clamar a Dios, cubrirse el corazón, dar el pésame, deplorar, deprimirse, desolarse, dolerse, echar de menos, endecharse, engurruñarse, enmantarse, ensombrecerse, entristecer, lamentar, llorar lágrimas de sangre, llorar lágrimas de sangre, no caber el corazón en el pecho, nublársele el cielo, padecer, partirse el alma, pasar la pena negra, pasar las penas del purgatorio, penarse, poderse ahogar con un cabello, sentir , sentir de muerte*
Sentimientos, conducta	*arrepentimiento*	*acusar la conciencia, arrepentirse, compungir, compungirse, concomerse, corroer, deplorar, dolerse, escarabajear la conciencia, lamentar, llorar, llorar con lágrimas de sangre, morderse las manos, pesarle a uno, recomerse, sentir*
Sensibilidad, sentidos	*audición*	*abrir el oído, abrir los oídos, abrir tanto el oído, aguzar el sentido, aguzar las orejas, aplicar el oído, atender, auscultar, beber las palabras, dar oídos, dejarse de oír, entreoír, escuchar, estar colgado de los labios, estar pendiente de la boca, herir el oído, no perder ripio, oír, perderse, sentir, taparse los oídos, trasoír*
Sentimientos, costumbres	*compasión*	*ablandarse, adolecer, apiadarse, arrancársele a uno las entrañas, compadecer, compadecerse, compartir, compungirse, condolerse, conmoverse, contristarse, deplorar, dolerse, emblandecerse, lamentar, lastimarse, sentir*

Tabla 3-11 Cuadros y Encabezados de Grupo para el verbo sentir en D.I.C (I)

Cuadro	Encabezado	Grupo
Intelección, conducta	*conciencia*	*acusar la conciencia, advertir, ajustarse con la conciencia, apercibir, argüir la conciencia, caer en la cuenta, cargar la conciencia, comprender, conocerse, darse cuenta, encargar la conciencia, enconar, entenderse, escarabajear, escarbar, escrupulizar, percibir, reflexionar, remorder, sentir, tener conciencia de*
Intelección	*experiencia*	*advertir, aprender, ensayar, experimentar, foguear, notar, sentir, sufrir, tocar, ver*
Apreciación-Juicio	*juicio*	*apreciar, calificar, censurar, conocer de, conocerse, considerar, creer, criticar, dictaminar, discernir, echar el fallo, emitir, enjuiciar, entender en, estimar, existimar, fallar, hacer distinción, informar, juzgar, librar, opinar, parecerle a uno, reputar, sentir, someter, tener por, tomarle las medidas (a uno), valorar*
Acción	*pasividad*	*aceptar, aguantar, experimentar, expiar, incurrir en, lastar, padecer, pasar, pasar por las picas, penar, percibir, permitir, recibir, sentir, ser objeto de, soportar, sufrir, tener buenas espaldas, tolerar, tomar, tomarle a uno*
Intelección, conducta	*previsión*	*agorar, antever, anunciar, aprevenir, augurar, barruntar, darle el aire de, darle el corazón, decirle el corazón, haberse tragado, mirar las cosas con anteojo de aumento, mirar las cosas con anteojo de larga vista, ominar, preconocer, predecir, prenotar, presentir, presumir, prevenir, prever, pronosticar, remusgar, sentir, sospechar, tenerse tragado, ver, ver las cosas con anteojo de aumento, ver las cosas con anteojo de larga vista*
Fisiología, Sensibilidad-sentidos, sentimientos	*sensibilidad*	*advertir, apreciar, entrar en, experimentar, impresionarse, notar, observar, padecer, percibir, sentir, sufrir*
Sentimientos	*sentimiento*	*abrigar, cobrar, coger, concebir, experimentar, profesar, sentir, tener, tomar*

Tabla 3-12 Cuadros y Encabezados de Grupo para el verbo sentir en D.I.C (II)

3.3.4. Extracción de preferencias de selección

En este apartado se expone la medida de asociación de selección (Resnik, 1 993), su adaptación a las Bases de Conocimiento sobre las que se trabaja y los resultados obtenidos.

3.3.4.1. Asociación de selección

La medida de asociación de selección de Resnik cuantifica las restricciones de selección de un predicado respecto a una clase de argumentos. Este valor surge a partir de la entropía de una variable aleatoria que se utiliza como herramienta para determinar qué cantidad de incertidumbre se tiene sobre los argumentos que pudieran aparecer junto a un predicado dado, es decir, qué "ha sucedido".

Este artificio se construye sobre la Teoría de la Información, que se fundamenta en la idea de que un suceso poco probable aporta más información que uno altamente probable. Por ejemplo, el verbo *palpitar* aporta más información sobre qué palabras se pueden esperar en su contexto que el verbo *dar*. La **cantidad de información** de un suceso dada por:

$$I(X = x) = \log_2 \frac{1}{p(x)}$$

La **entropía** de una variable aleatoria es la información media que aporta la misma:

$$H(X) = \sum_x p(x) \cdot \log_2 \frac{1}{p(x)}$$

La **entropía relativa**, dadas dos distribuciones de probabilidad p, q:

$$D(p \parallel q) = \sum_x p(x) \cdot \log_2 \frac{p(x)}{q(x)}$$

También es conocida como la distancia de Kullback-Leibler y verifica:

$$D(p \parallel q) \geq 0$$

$$D(p \parallel q) = 0 \Leftrightarrow p, q \text{ idénticas}$$

$$D(p \parallel q) = \sum_x p(x) \cdot \left[\log_2 \frac{1}{q(x)} - \log_2 \frac{1}{p(x)}\right] = \sum_x p(x) \cdot \left[I_q(X = x) - I_p(X = x)\right]$$

es el promedio de diferencias de cantidad de información.

Se interpreta como el coste, en bits de información, que tiene el hecho de asumir que la distribución de probabilidad es q cuando en realidad es p.

Resnik recurre a la entropía relativa para evaluar si existen diferencias entre la distribución de probabilidades de una clase léxica de argumentos C, supuesto que ha sucedido el predicado p_i, frente a la probabilidad de la clase C sin tal supuesto, es decir:

$$D\big(p(C|p_i) \parallel p(C)\big) = \sum_C p(C|p_i) \cdot \log_2 \frac{p(C|p_i)}{p(C)}$$

Siendo:

$p(C)$ la distribución de la clase C, o **distribución a priori**

$p(C|p_i)$ la distribución de la clase C, conocido p_i, *o* **distribución a posteriori**.

Si un predicado restringe sus argumentos, selecciona una clase léxica, por tanto, las distribuciones a priori y posteriori deben diferir. Además, el rango de selección puede ser amplio o estrecho según la fuerza de la restricción que imponga a sus argumentos. En caso de rangos estrechos, la diferencia entre las dos distribuciones será grande, la distribución a priori debe ser muy diferente a la distribución a posteriori. Por este motivo se denomina **fuerza de la selección** a $D\big(p(C|p_i) \parallel p(C)\big)$, y es el coste de asumir que C no tiene ninguna motivación para aparecer en el contexto de p_i.

Las clases C a las que hace referencia son synsets de hipónimos o hiperónimos del lexicón. Sin embargo, un sustantivo consta en un conjunto de synsets en la taxonomía, correspondientes a los diferentes sentidos.

La **asociación de selección**, *A(p$_i$,C)*, de un predicado, p_i, se define para cada clase, *C*, como la contribución de la clase en la fuerza de la selección, ponderada por el factor de escala respecto al total –normaliza el indicador y permite comparar los valores independientemente de la fuerza de la selección.

$$A(p_i, C) = \frac{p(C|p_i) \log_2 \frac{p(C|p_i)}{p(C)}}{\sum_C p(C|p_i) \cdot \log_2 \frac{p(C|p_i)}{p(C)}}$$

El numerador corresponde a cada sumando de $D\big(p(C|p_i) \parallel p(C)\big)$ y el denominador es dicho valor.

En la práctica, en lugar de trabajar con distribuciones de probabilidad, se emplean sus estimaciones obtenidas a partir de las frecuencias registradas en el corpus:

$$\hat{p}(C|p_i) = \frac{\sum_{w \in C} \frac{frec(p_i, w)}{frec(w)}}{frec(p_i)}$$

$$\hat{p}(C) = \frac{\sum_{w \in C} \frac{1}{|\mathbb{C}_w|}}{N}$$

siendo: $p_i$el predicado, C la clase, w una palabra, $frec(p_i, w)$ la frecuencia conjunta en el corpus del predicado con la palabra *w,* $frec(p_i)$ la frecuencia del predicado en el corpus, y $\mathbb{C}_W$ el conjunto de todas las clases en las que aparece *w* en la clasificación, *N* es el número de palabras en la clasificación.

3.3.4.2. La asociación de selección sobre el Diccionario Ideológico Vox

El Diccionario Ideológico Vox, DIV, proporciona las clases de palabras a partir de las que inducir grupos semánticos en los argumentos que seleccionan los predicados en el corpus. Dada la estructura de ambos diccionarios del mismo modo se podría utilizar el DILE, sin embargo, al disponer de la información de la categoría de las palabras del grupo se escoge el primero para realizar los cálculos. Las suposiciones que se hacen son las siguientes:

- Las clases que se utilizarán son los grupos en el nivel de la parte analógica.
- Se diferencian las tres estructuras:
    - *Verbo* + *Sustantivo*: El predicado tiene que ser un verbo.
    - *Sustantivo* + *Adjetivo*: El predicado tiene que ser un adjetivo.
    - *Verbo* + *Adverbio*: El predicado tiene que ser un adverbio.

En cada caso, se consideran solo los grupos del ideológico cuya categoría gramatical coincida con la de los argumentos.

Se ilustran los cálculos utilizando como predicado el verbo *sentir*, y el grupo de sustantivos**:**

$$G = \begin{Bmatrix} analgesia, sinalgia, yaya, acritud, rayo, clavo, puntada, \\ ramalazo, tormento, punzada \end{Bmatrix}$$

incluido en la cabecera dolor y que tiene elementos que aparecen en el corpus en el contexto del verbo *sentir*.

$$A(sentir, G) = \frac{\hat{p}(sentir|G) \cdot \log_2 \frac{\hat{p}(sentir|G)}{\hat{p}(G)}}{\sum_{G_{Sust_{sentir}}} \hat{p}\left(G_{Sust_{sentir}}\right) \cdot \log_2 \frac{\hat{p}\left(sentir|G_{Sust_{sentir}}\right)}{\hat{p}\left(G_{Sust_{sentir}}\right)}}$$

$$\hat{p}(sentir|G) = \frac{\sum_{palabra \in G} \frac{frec(sentir, palabra)}{frec(palabra)}}{frec(sentir)} = \sum_{palabra \in G} \frac{frecRel_{palabra}(sentir)}{frec(sentir)}$$

$$\hat{p}(G) = \frac{\sum_{palabra \in G} \frac{1}{\left|\mathbb{C}_{palabra}\right|}}{N}$$

$\left|\mathbb{C}_{palabra}\right|$ es el cardinal del conjunto de todos los grupos en los que aparece *palabra.*

Los grupos en la cabecera *dolor* que aparecen en el contexto del verbo *sentir*, y el orden inducido por la asociación de selección respecto al total de grupos de sustantivos de los que hay muestras en el corpus en combinación con el verbo *sentir* se muestran en la Tabla 3-13.

Orden	**Grupo**	***A(sentir, G)***
2	*acritud, analgesia, clavo, dolor, puntada, punzada, ramalazo, rayo, sinalgia, tormento, yaya*	0.00084056
5	*ardor, comecome, comezón, escozor, fogaje, fuego, hormiguilla, hormiguillo, picazón, picor, piquiña, prurito, quemazón, rascazón*	0.00056183
49	*agujetas, cefalalgia, cefalea, ciática, entuertos, ijada, jaqueca, lumbago, mastalgia, mialgia, migraña, neuralgia, odontalgia, ostealgia, otalgia, retortijón, silbido de oídos, torticolis, tortícolis*	0.00025726
2640	*arenillas, bezoar, cálculo, piedra*	0.000019488

Tabla 3-13 Resultados para sentir + "los grupos de la cabecera dolor"

La Tabla 3-14 recoge los 25 grupos con mayor asociación de selección cuando ésta se evalúa en grupos de sustantivos en el DIV Se incluyen todos los elementos del grupo que han sido registrados en el corpus –se resaltan en negrilla los que previamente se catalogaron como colocaciones, mostrándose también las cabeceras. Nótese que se pueden obtener diferentes grupos de una misma cabecera con desiguales valores de asociación de selección. Se muestra la capacidad de la asociación de selección para detectar grupos de argumentos que no incluyen elementos identificados como colocaciones –a pesar de que no existe una marcada preferencia individual, el conjunto de combinaciones aporta la suficiente fuerza de selección para hacer destacar al grupo.

Forma Canónica	Cabecera	*A(sentir, G)*
*agitación, alarma, ansia, ansiedad, batalla, combate, comezón, conflicto, desasosiego, **guerra**, hormigueo, inquietud, intranquilidad, malestar, pesadilla, preocupación, taco, zozobra*	*intranquilidad*	*0,00103141*
*ardor, comezón, escozor, **fuego**, picazón, picor, prurito, quemazón*	*dolor*	*0,00084056*
mojada, picadura, pinchazo, puntada, punzada	*herida*	*0,00061050*
*aflicción, agonía, ahogo, amargura, **angustia**, apuro, congoja, desconsuelo, **dolor**, duelo, hiel, **pasión**, **pena**, **pesar**, quebranto, **sentimiento**, sufrimiento*	*sufrimiento*	*0,00056807*
clavo, dolor, puntada, punzada, ramalazo, rayo, tormento, yaya	*dolor*	*0,00056182*
arcada, asco, basca, escrúpulo, fatiga, náusea, repugnancia	*indigestión*	*0,00053228*
desazón, disgusto, pesadumbre, punzada, puñalada, sinsabor, tribulación	*sufrimiento*	*0,00052641*
enfado, enojo, escozor, indignación, mosca, ofensa, picazón	*enojo*	*0,0005178*
*afán, agonía, anhelo, ansia, apetito, **deseo**, empeño, emulación, envidia, **gana**, **hambre**, pelota, pío, prurito, sed, voto*	*deseo*	*0,00050270*
*azote, calamidad, catástrofe, desastre, desdicha, desgracia, desventura, drama, fatalidad, infortunio, **mal**, miseria, ramalazo, rayo, suceso, suspenso, tormenta, tragedia, trago, **través***	*revés*	*0,00047396*
*asco, **horror**, náusea, repugnancia, repulsión*	*desagrado*	*0,00047321*
disgusto, enojo, fastidio, gravedad, incomodidad, incomodo, mareo, molestia	*desagrado*	*0,00045289*
aguijón, aguijonazo, espuela, pinchazo	*inducción*	*0,00044326*
aberración, delirio, desatino, desvarío, devaneo, enajenación, locura, luna, ramalazo, vértigo	*demencia*	*0,00044148*

Tabla 3-14 Los 25 grupos con mayor asociación de selección para el verbo sentir (I)

Forma Canónica	Cabecera	*A(sentir, G)*
*alborozo, **alegría**, dicha, euforia, felicidad, gozo, **humor**, júbilo, **placer**, regocijo*	*alegría*	*0,00042495*
*aprensión, cuidado, desconfianza, escrúpulo, espina, inseguridad, malicia, **miedo**, **pensamiento**, prejuicio, recelo, sospecha, **temor***	*recelo*	*0,00041306*
aguijonazo, dardo, picante, quemazón	*ironía*	*0,00040895*
*añoranza, depresión, melancolía, nostalgia, soledad, **vacío***	*tristeza*	*0,00038088*
animadversión, animosidad, antipatía, aversión, discordia, enemistad, rivalidad	*enemistad*	*0,00036861*
abatimiento, aturdimiento, hormigueo	*enfermedad*	*0,00036835*
agrado, deleite, delicia, dulzura, fruición, gloria, goce, gozo, gusto, placer, regalo, satisfacción	*placer*	*0,00036707*
*aleteo, contracción, **golpe**, latido, palpitación, pulsación, pulso, salto*	*agitación*	*0,00036645*
*cicatriz, deje, **efecto**, excitación, huella, impaciencia, impacto, **impresión**, regusto, sabor, **sensación***	*excitación*	*0,00036569*
*afección, **afecto**, **amor**, apego, cariño, debilidad, querencia, querer, ternura*	*amor*	*0,00036494*
*efusión, **emoción**, entusiasmo, exaltación, expansión, inspiración, intensidad, lirismo, **pasión**, tensión, vehemencia*	*exaltación*	*0,00036076*

Tabla 3-15 Los 25 grupos con mayor asociación de selección para el verbo sentir (II)

La representación gráfica de los resultados que se registran para cada uno de los ejemplos tratados en este capítulo sigue un mismo patrón. Para la mayor parte de los grupos que contienen alguna de las bases con las que combina el colocativo, se obtiene una puntuación casi nula; sin embargo, en todos los casos se observa un conjunto de grupos del DIV que presentan un valor alto de asociación de selección.

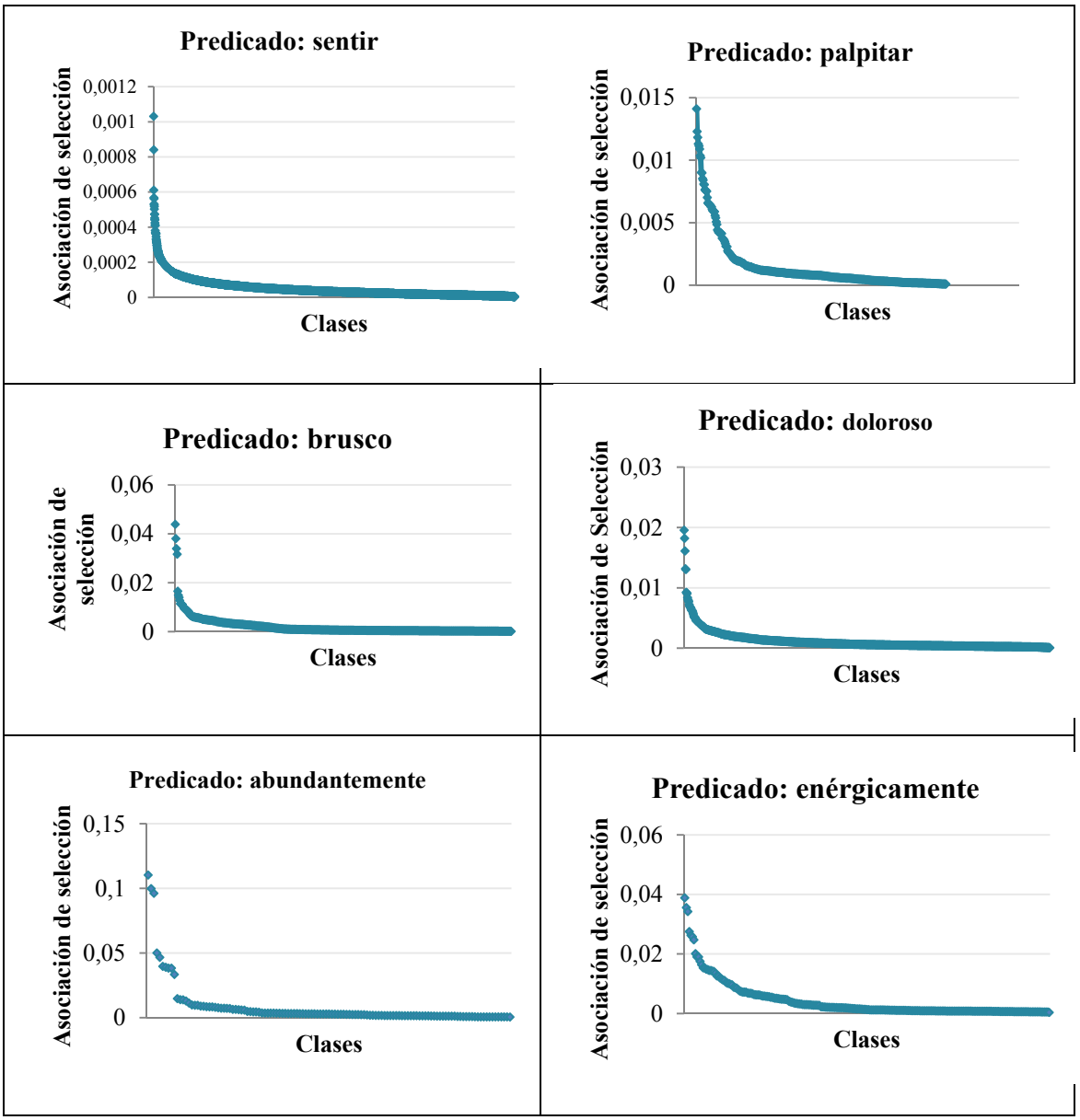


Gráfico 3-24 Asociación de Selección para los ejemplos tratados.

.

Los elementos pertenecientes a los grupos mejor puntuados para las restantes formas canónicas que han servido de ejemplo en este capítulo se enumeran en el Anexo 6.

3.4. Evaluación de la asociación de selección en el DCECR.

Los resultados obtenidos en los ejemplos anteriores muestran el buen comportamiento de la asociación de selección. Como indicio de la idoneidad del método se aportan las clases léxicas que describen las capacidades de selección de argumentos del verbo *sentir* descritas en DCECR. (Tabla 3-2) junto con las clases léxicas del DIV con mejores puntuaciones, se han resaltado las coincidencias en los grupos. En el primer caso se puede considerar que el grupo

proporcionado por el DCECR es prácticamente un subconjunto del grupo del DIV (Tabla 3-16), salvo un elemento todos están incluidos en él. En el segundo caso, Tabla 3-17, se recurre a las definiciones de las palabras en cada grupo para verificar que sus elementos responden al rasgo semántico *tristeza* que define al grupo del DCECR. Todos los elementos del grupo tienen una acepción con significado de *tristeza* o *aflicción*.

	DCECR.**: SUSTANTIVOS QUE DENOTAN PREOCUPACIÓN, INCERTIDUMBRE O RESERVA = **{inquietud, miedo, preocupación, reparo, temor, terror, zozobra}
DIV	*agobio, alarma, ansia, comecome, comezón, conflicto, desabrimiento, hormigueo, **inquietud**, malestar, mareta, **miedo**, pesadilla, **preocupación**, rebato, reparo, suspense, **temor**, **terror**, torozón, **zozobra***

Tabla 3-16 Clases léxicas para sentir en el DCECR. y DIV.

	DCECR** SUSTANTIVOS QUE DENOTAN TRISTEZA O CONMISERACIÓN = **{compasión, conmiseración , lástima, pena, tristeza}
DIV	*aflicción, agonía, angustia, congoja, desconsuelo, dolor, duelo, **pena**, pesar, sentimiento*

Tabla 3-17 Clases léxicas para sentir en el DCECR y D.I.V.

aflicción (DEA)

1. Pesadumbre o tristeza

agonía (DRAE)

1. f. Angustia y congoja del moribundo; estado que precede a la muerte.

2. f. Pena o aflicción extremada.

3. f. Angustia o congoja provocadas por conflictos espirituales.

angustia. (DRAE)

(Del lat. *angustĭa* 'angostura', 'dificultad').

1. f. Aflicción, congoja, ansiedad.

5. f. Dolor o sufrimiento.

angustia (DEA)

1. Aflicción o congoja

congoja (DRAE)

1. f. Desmayo, fatiga, angustia y aflicción del ánimo.

desconsuelo. (DRAE)

1. m. Angustia y aflicción profunda por falta de consuelo.

desconsuelo (DAE)

1 Pena muy profunda y que se siente como insuperable

dolor. (DRAE)

(Del lat. dolor, -*ōris*).

2. m. Sentimiento de pena y congoja.

duelo (DAE)

1. Pena o dolor por la muerte de alguien

2. En gral: Pena o dolor

pena (DAE)

1. Tristeza (estado de ánimo)

2. Compasión o lástima

pesar[2]. (DRAE)

1. m. Sentimiento o dolor interior que molesta y fatiga el ánimo.

sentimiento (DUE)

1. («de») m. Estado afectivo de la clase que se expresa: 'Un sentimiento de angustia, de soledad, de abandono, de insatisfacción'.

2. Estado de ánimo de los que consisten en sentir atracción o aversión por una persona, o en sentir *alegría o tristeza.

Utilizando como medida de similitud entre los grupos del DCECR y el DIV el coeficiente de Jaccard:

$$similiud(G_i, G_j) = \frac{|G_i \cap G_j|}{|G_i \cup G_j|}$$

Se obtienen los grupos con $similitud(G_i, G_j) > 0$ ordenados por este valor, se consideran grupos similares los que presentan cumplen $similitud(G_i, G_j) > 0{,}1$. De los 3685 grupos del DIV que aparecen con el verbo *sentir* en el corpus (Gráfico 3-25), para cada caso se encuentran entre 1 y 3 grupos similares en el DIV según este criterio. En todos los casos alguno de ellos está entre los 25 con mayor asociación de selección cuando se toma como espacio semántico el DIV.

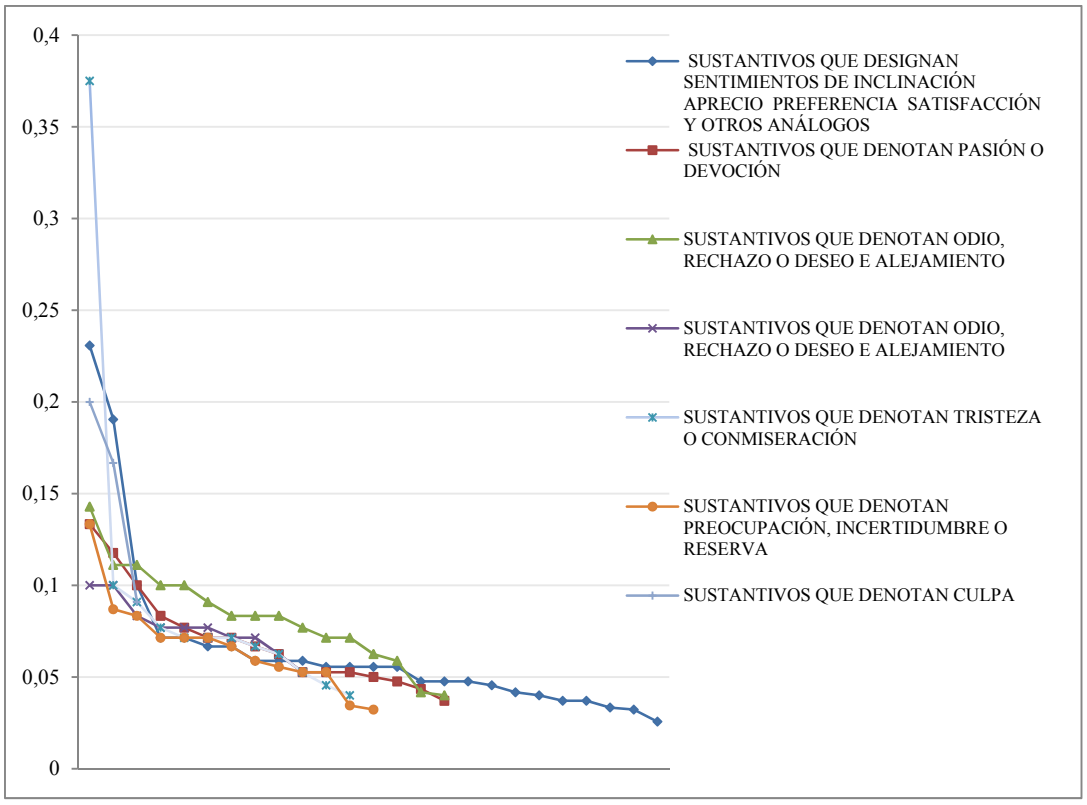


Gráfico 3-25 Grupos en el DIV similares a los grupos para el verbo sentir.

Con objeto de ratificar la validez de este indicador de forma objetiva, se realiza el experimento consistente en evaluarlo sobre el espacio semántico definido por las clases léxicas del DCECR. De este modo, se espera obtener valores altos de la asociación de selección entre un determinado predicado y las clases léxicas que selecciona como argumentos según el DCECR. Se ilustra la discusión con los ejemplos utilizados a lo largo del capítulo que tienen entrada en el citado diccionario: *sentir, palpitar, brusco, abundantemente y enérgicamente*.

En el Gráfico 3-26 se representan las puntuaciones que obtienen los grupos del DCECR cuando se utiliza éste como espacio semántico de interés. Se resaltan en las clases semánticas que combinan con el verbo *sentir* –se observa que la mayoría se posiciona entre las más altas.

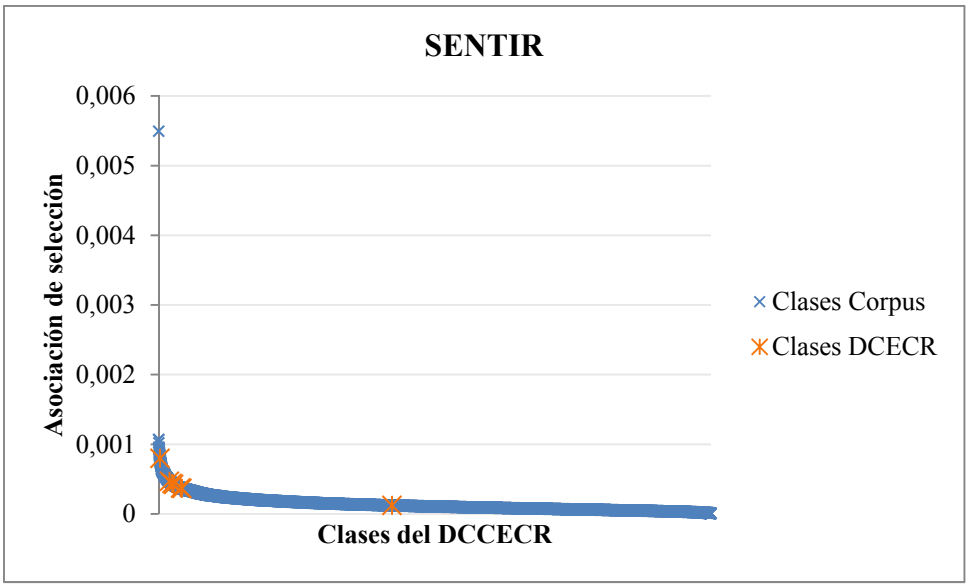


Gráfico 3-26 Asociación de Selección sobre el DCECR. para el verbo sentir

Además, otras clases con puntuaciones elevadas corresponden a grupos que hacen referencia a sustantivos que comparten rasgos semánticos con los enumerados para el verbo *sentir*. Se presentan los mejores casos en la Tabla 3-18 y Tabla 3-19 ordenados de mayor a menor asociación de selección. Se ha calculado el coeficiente de Jaccard entre éstas y las del verbo sentir en el DCECR. Se ha encontrado que en 6 de los casos se cumple el criterio establecido para considerarlas similares, indicándose los casos que lo verifican entre paréntesis. Los grupos en este espacio semántico están motivados por las capacidades combinatorias de los predicados que los determinan, y no por una clasificación semántica de las palabras del español, hecho que determina su solapamiento. De esta forma, se explica el resultado obtenido, clases léxicas seleccionadas por el verbo *sentir* entre las mejores y con una alta relación con las mejor posicionadas en el ranking.

Clase DCECR. en el corpus	
SUSTANTIVOS QUE DESIGNAN DIVERSOS SENTIMIENTOS SENSACIONES Y ESTADOS DE ÁNIMO DE SIGNO NEGATIVO EN ESPECIAL LOS REFERIDOS A LA EXCITACIÓN NERVIOSA EL ENOJO LA AVERSIÓN Y EL DESASOSIEGO TAMBIÉN CON OTROS QUE EXPRESAN ALGUNAS DE SUS MANIFESTACIONES	*nervios, ánimo, ira, cólera, inquietud, excitación, angustia, descontento, disgusto, cabreo, odio, rencor, sollozo, sufrimiento, estrés, enfado, rabia, ansiedad, crispación, enojo, indignación*
SUSTANTIVOS QUE DESIGNAN DIVERSAS AFLICCIONES DEL ÁNIMO MÁS FRECUENTEMENTE SI SE RELACIONAN CON EL DISGUSTO LA PESADUMBRE O LA AVERSIÓN HACIA ALGO (IV)	*pesar, dolor, sufrimiento, rencor, rechazo, odio, pesimismo, desinterés, disgusto, malestar, desagrado, remordimiento, resentimiento*
SUSTANTIVOS QUE DENOTAN AVERSIÓN RENCOR RECHAZO Y OTROS SENTIMIENTOS HOSTILES EN DIVERSOS GRADOS (IV)	*odio, enemistad, hostilidad, antipatía, animadversión, rechazo, inquina, rivalidad, resentimiento, rencor, aversión, desprecio, aborrecimiento, repudio*
SUSTANTIVOS QUE DENOTAN ANSIEDAD TEMOR Y OTRAS FORMAS DE DESEQUILIBRIO EMOCIONAL (VI)	*ansiedad, nerviosismo, escalofrío, estupor, tensión, temor, miedo, irritabilidad*
SUSTANTIVOS QUE DESIGNAN SENTIMIENTOS O ESTADOS DE ÁNIMO DE NATURALEZA NEGATIVA MÁS FRECUENTEMENTE SI MANIFIESTAN AGITACIÓN IRRITACIÓN Y OTRAS FORMAS DE INESTABILIDAD	*nerviosismo, emoción, angustia, furia, celos, irritación, rabia, nervios, ira, ansiedad, desesperación, excitación, histeria, pánico, miedo*
SUSTANTIVOS QUE DESIGNAN EL DOLOR LA PREOCUPACIÓN Y OTROS ESTADOS AFLICTIVOS (V)	*preocupación, dolor, consternación, depresión, desilusión, disgusto, malestar, pena, tristeza, angustia, cansancio, crispación, desaliento*

Tabla 3-18 Clases léxicas del DCECR. con mayor asociación de selección para sentir (I).

Clase DCECR. en el corpus	
SUSTANTIVOS QUE DESIGNAN OTRAS ALTERACIONES PSÍQUICAS ASÍ COMO ALGUNAS AFECCIONES LIGADAS A ELLAS MÁS FRECUENTEMENTE SI SE RELACIONAN CON LA ANSIEDAD LA INSATISFACCIÓN O EL DESABRIMIENTO	*desazón, tensión, presión, malestar, preocupación, remordimiento, soledad, pena, amargura, neura, estrés, angustia*
SUSTANTIVOS QUE DENOTAN AVERSIÓN REACCIÓN HOSTIL Y OTRAS FORMAS DE RECHAZO TAMBIÉN CON OTROS QUE DESIGNAN LOS SENTIMIENTOS QUE SE SUELEN ASOCIAR CON TALES REACCIONES (IV)	*enemistad, antipatía, desprecio, odio, hostilidad, desconfianza, animadversión, desafecto, recelo, rechazo, oposición, envidia, ira, anatema, ojeriza*
SUSTANTIVOS QUE DENOTAN MALESTAR, DISGUSTO Y OTRAS FORMAS DE PESADUMBRE, MÁS FRECUENTEMENTE LAS QUE DESIGNAN EL EFECTO ANÍMICO QUE PROVOCA LO QUE SE MALOGRA. TAMBIÉN CON OTROS QUE DESIGNAN EL QUE PRODUCE LA PÉRDIDAD DE LAS FUERZAS O LA VOLUNTAD	*frustración, decepción, desánimo, desgana, amargura, cansancio, debilidad, agotamiento, inapetencia, aburrimiento, malestar*
SUSTANTIVOS QUE DENOTAN INCLINACIÓN DEL ÁNIMO A MENUDO INTENSA Y FAVORABLE HACIA ALGUNA COSA (I)	*predilección, fascinación, pasión, afición, simpatía, amor, cariño, adicción, debilidad, parcialidad, admiración, respeto, identificación*

Tabla 3-19 Clases léxicas del DCECR. con mayor asociación de selección para sentir (II).

El análisis de los resultados del verbo *palpitar* indica un comportamiento que se ajusta al descrito para el verbo *sentir,* a pesar de que en el gráfico aparecen numerosas clases léxicas mejor puntuadas que las propuestas para este verbo en el diccionario (Gráfico 3-27). Por ejemplo, la clase léxica en el catálogo de *palpitar*:

> SUSTANTIVOS QUE DESIGNAN EMOCIONES Y SENTIMIENTOS, PRINCIPALMENTE LOS DE NATURALEZA PASIONAL

que la forman los sustantivos: *amor, deseo, emoción, pasión* y *pulsión*, tiene correspondencia con la clase:

> SUSTANTIVOS QUE DENOTAN ENERGÍA, AFÁN O EMPEÑO PUESTOS EN ALGUNA COSA, ASÍ COMO OTRAS CUALIDADES Y ACTITUDES

NECESARIAS PARA LLEVAR A CABO LO QUE SE CONSIDERA DIFÍCIL O ARRIESGADO

Entre sus elementos se encuentran los sustantivos: *corazón, arrojo, gana.* Por una parte, *deseo* y *pasión* se encuentran entre los sustantivos sinónimos de *afán, deseo* también es sinónimo de *empeño*. Por otro lado, *corazón* es sinónimo de: *amor* y *sentimiento, gana* lo es de: *deseo, anhelo* y *arrojo* es sinónimo de: *admiración, pasión, arrebato.* Algo similar sucede entre SUSTANTIVOS QUE DENOTAN FUERZA O VITALIDAD formada por *fuerza, vitalidad, vida* y

SUSTANTIVOS QUE DESIGNAN DIVERSOS RASGOS SOBRESALIENTES DE LAS PERSONAS O LAS COSA, MÁS FRECUENTEMENTE SI SE RELACIONAN CON EL ÍMPETU, LA ENERGÍA Y OTRAS FORMAS EN QUE PUEDE MANIFESTARSELA CAPACIDAD DE MOVERSE, PORFIAR, IMPULSAR O ATRAER.

formado por los sustantivos: *atracción, corazón, energía, estilo, fuerza, impulso, obstinación, pasión, personalidad, vena populista,* entre los que se halla el sustantivo *fuerza* y el sustantivo *energía* que figura como sinónimo de *vitalidad* y de *vida*.

Del mismo modo, existe cierto solapamiento entre:

SUSTANTIVOS QUE DESIGNAN ARTES, TÉCNICAS Y TENDENCIAS ARTÍSTICAS, ASÍ COMO ALGUNAS DE SUS MANIFESTACIONES

conformado por: *poesía, literatura, melodía, obra, creación, cubismo, flamenco, partitura, acorde, tendencia*

y

SUSTANTIVOS QUE DESIGNAN MANIFESTACIONES MUSICALES, TAMBIÉN CON OTROS QUE EXPRESAN LA REPETICIÓN REGULAR DEL SONIDO

ritmo, melodía, canción, música, compás, estribillo, pop, rap, latido. Se entienden las manifestaciones musicales como un caso de manifestaciones artísticas, además los sustantivos del primer caso: acorde, melodía, partitura encajan perfectamente en este segundo grupo.

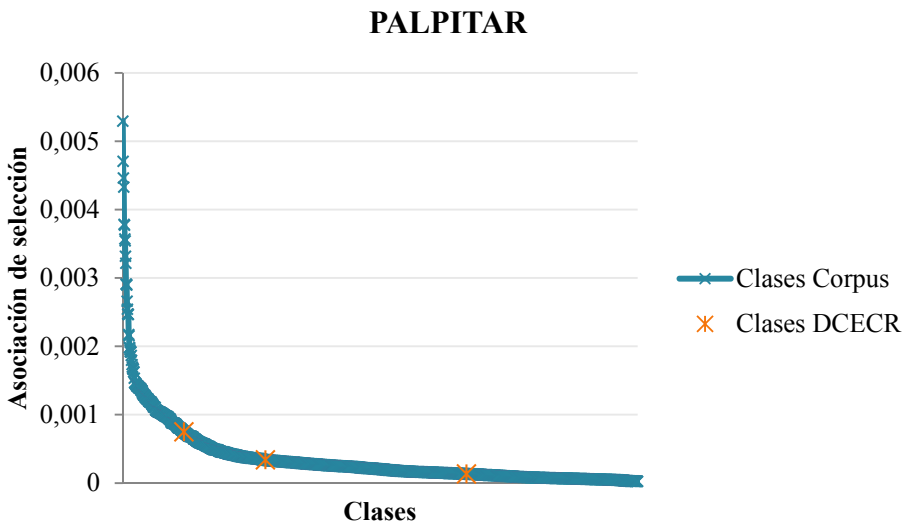


Gráfico 3-27 Asociación de selección sobre el DCECR. para el verbo palpitar.

Clase DCECR. en el corpus	
SUSTANTIVOS QUE DESIGNAN MANIFESTACIONES MUSICALES, TAMBIÉN CON OTROS QUE EXPRESAN LA REPETICIÓN REGULAR DEL SONIDO	*compás, melodía, latido,…*
SUSTANTIVOS QUE DESIGNAN LAS PARTES INTERIORES O ESENCIALES DE ALGO O DE ALGUIEN	*corazón, entraña, interior,…*
SUSTANTIVOS QUE DENOTAN ENERGÍA, AFÁN O EMPEÑO PUESTOS EN ALGUNA COSA, ASÍ COMO OTRAS CUALIDADES Y ACTITUDES NECESARIAS PARA LLEVAR A CABO LO QUE SE CONSIDERA DIFÍCIL O ARRIESGADO	*corazón, emoción,…*
EL SUSTANTIVO RITMO Y CON OTROS QUE DESIGNAN ACCIONES O MOVIMIENTOS, ESPECIALMENTE SI SON REGULARES	*latido, ritmo,…*

Tabla 3-20 Clases léxicas del DCECR. con mayor asociación de selección para palpitar.

Clase DCECR. en el corpus	
SUSTANTIVOS QUE DESIGNAN DIVERSOS RASGOS SOBRESALIENTES DE LAS PERSONAS O LAS COSA, MÁS FRECUENTEMENTE SI SE RELACIONAN CON EL ÍMPETU, LA ENERGÍA Y OTRAS FORMAS EN QUE PUEDE MANIFESTARSELA CAPACIDAD DE MOVERSE, PORFIAR, IMPULSAR O ATRAER	*energía, fuerza, pasión…*
SUSTANTIVOS QUE DESIGNAN CAPACIDADES Y ATRIBUTOS DE LAS PERSONAS RELATIVOS A SU FORMA DE SER, DE PENSAR O DE ACTUAR MENTAL O EMOCIONALMENTE, TAMBIÉN CON OTROS QUE DESIGNAN ALGUNOS ÓRGANOS EN LOS QUE SUPUESTAMENTE RADICAN	*mente, corazón,…*
SUSTANTIVOS QUE DESIGNAN PARTES, ELEMENTOS, ASPECTOS O COMPONENTES DEL SER HUMANO QUE REPRESENTAN SENTIMIENTOS, SENSACIONES O CAPACIDADES	*alma, espíritu,…*

Tabla 3-21 Clases léxicas del DCECR. con mayor asociación de selección para palpitar**.**

Clase DCECR. en el corpus	
	energía, fuerza, pasión…
SUSTANTIVOS QUE DESIGNAN DIVERSOS RASGOS SOBRESALIENTES DE LAS PERSONAS O LAS COSA, MÁS FRECUENTEMENTE SI SE RELACIONAN CON EL ÍMPETU, LA ENERGÍA Y OTRAS FORMAS EN QUE PUEDE MANIFESTARSELA CAPACIDAD DE MOVERSE, PORFIAR, IMPULSAR O ATRAER	
SUSTANTIVOS QUE DESIGNAN CAPACIDADES Y ATRIBUTOS DE LAS PERSONAS RELATIVOS A SU FORMA DE SER, DE PENSAR O DE ACTUAR MENTAL O EMOCIONALMENTE, TAMBIÉN CON OTROS QUE DESIGNAN ALGUNOS ÓRGANOS EN LOS QUE SUPUESTAMENTE RADICAN	*mente, corazón,…*
SUSTANTIVOS QUE DESIGNAN PARTES, ELEMENTOS, ASPECTOS O COMPONENTES DEL SER HUMANO QUE REPRESENTAN SENTIMIENTOS, SENSACIONES O CAPACIDADES	*alma, espíritu,…*

Tabla 3-22 Clases léxicas del DCECR con mayor asociación de selección para palpitar**.**

El resto de ejemplos ratifica los resultados obtenidos en los casos *sustantivo + verbo*. El adjetivo *brusco* obtiene las máximas puntuaciones en las clases léxicas que selecciona según el DCECR. Si el análisis se centra en los adverbios *abundantemente* y *enérgicamente*, se observa que la mayor parte de las clases bien puntuadas tienen un alto grado de solapamiento y que tienen correspondencia con alguna de las aportadas para tales adverbios por el DCECR como clases léxicas que seleccionan como argumentos. Para el adverbio *enérgicamente* destaca la

gran variedad de clases similares que se aportan, como se puede verificar en la Tabla 3-25 y la Tabla 3-26

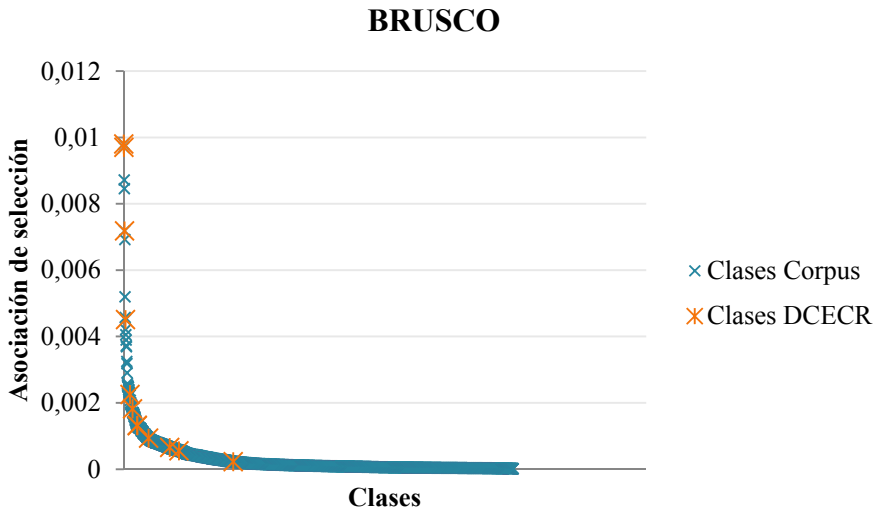


Gráfico 3-28 Asociación de selección sobre el DCECR. para el adjetivo brusco.

.**Clase DCECR. en el corpus**	
SUSTANTIVOS QUE DENOTAN MOVIMIENTO, Y A MENUDO CAMBIO RÁPIDO DE DIRECCIÓN	*maniobra, movimiento, viraje,…*
SUSTANTIVOS QUE DENOTAN DETENCIÓN O VARIACIÓN REPENTINA DE VELOCIDAD	*aceleración, frenazo,…*
SUSTANTIVOS QUE DENOTAN DETENCIÓN O DISMINUCIÓN DE LA VELOCIDAD, DERIVADOS FORMALMENTE DE LOS VERBOS DEL APARTADO A	*frenazo, parada,…*
OTROS SUSTANTIVOS QUE DENOTAN REFORMA, EVOLUCIÓN, CAMBIO O TRANSFORMACIÓN DE ALGO	*cambio, viraje,…*
SUSTANTIVOS QUE DENOTAN GOLPE, CHOQUE, ENCUENTRO VIOLENTO O MOVIMIENTO IMPULSIVO O IMPETUOSO	*golpe, tirón, sacudida,…*
SUSTANTIVOS QUE DENOTAN CAMBIO DE ESTADO, MÁS FRECUENTEMENTE SI SUPONE PROGRESO, MEJORA, INCREMENTO, DESCENSO O REDUCCIÓN	*aumento, cambio, viraje,…*
SUSTANTIVOS QUE DENOTAN CAMBIO DE ESTADO, DE DIRECCIÓN O DE POSICIÓN	*aumento, transición, viraje,…*

Tabla 3-23 Clases léxicas del DCECR. con mayor asociación de selección para brusco.

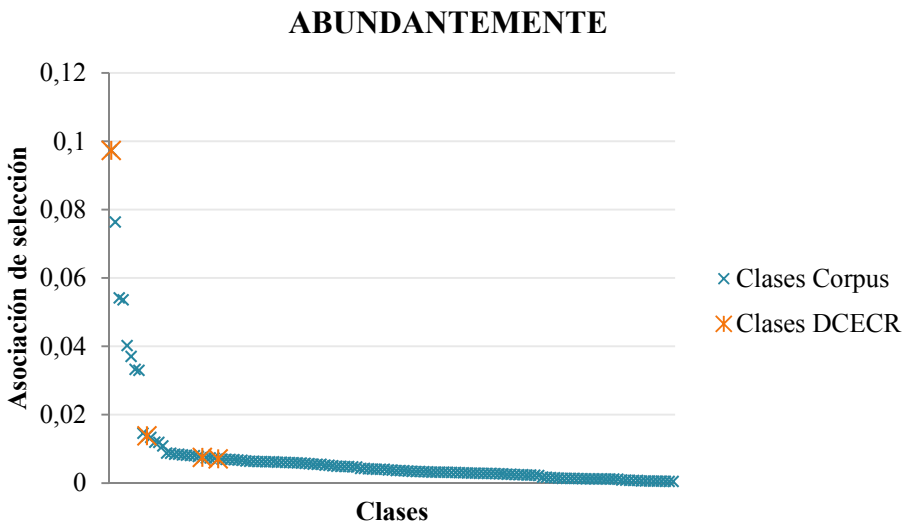


Gráfico 3-29 Asociación de selección sobre el DCECR. para el adverbio abundantemente.

Clase DCECR. en el corpus	
VERBOS QUE DESIGNAN EL PROCESO DE SURGIR O FLUIR UN LÍQUIDO U OTRAS MATERIAS QUE SE LE ASIMILAN FÍSICA O FIGURADAMENTE, TAMBIÉN CON OTROS VERBOS QUE EXPRESAN LA ACCIÓN QUE DESENCADENA ESE PROCESO	*correr, regar, sangrar,…*
VERBOS QUE DESIGNAN DIVERSAS FORMAS DE PONER ALGO A DISPOSICIÓN DE ALGUIEN	*proveer,…*
EL VERBO SUDAR Y CON OTROS QUE DENOTAN EXPULSIÓN O DERRAMAMIENTO DE LÍQUIDOS, FLUIDOS U OTRAS MATERIAS QUE PUEDEN ASIMILÁRSELES, SE USAN A VECES EN SENTIDO FIGURADO	*llorar, sudar,…*
OTROS VERBOS QUE DESIGNAN LA ACCIÓN DE FLUIR, DERRAMARSE O ESPARCIRSE ALGO	*correr, llorar, sudar,…*
VERBOS QUE DENOTAN ESFUERZO O DIFICULTAD PARA EL DESEMPE O DE UNA ACTIVIDAD	*sudar,…*
VERBOS QUE DENOTAN REPARTO, ENTREGA O DISTRIBUCIÓN DE ALGO, A MENUDO IMPRUDENTES O EXCESIVAS	*dar, regar,…*
VERBOS QUE DENOTAN EXPULSIÓN O DERRAMAMIENTO DE UN LÍQUIDO	*llorar, sudar,…*
VERBOS QUE DENOTAN INGESTIÓN	*beber, comer,…*

Tabla 3-24 Clases léxicas con mayor asociación de selección para abundantemente.

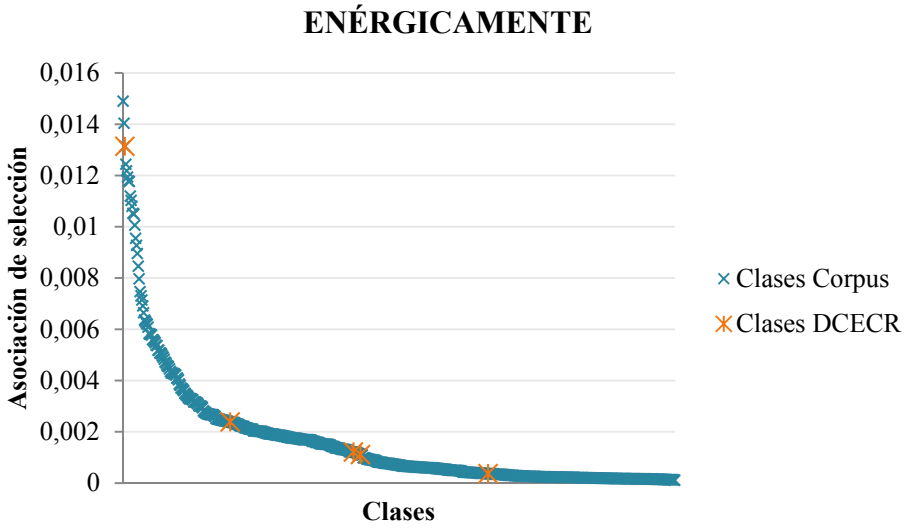


Gráfico 3-30 Asociación de selección sobre el DCECR. para el adverbio enérgicamente.

Clases DCECR. en el corpus	
	protestar,…
VERBOS QUE DESIGNAN MANIFESTACIONES DE OPOSICIÓN, DISENSO O DISCONFORMIDAD EN DIVERSOS GRADOS Y FORMAS, EN OCASIONES EXPRESADAS CON VEHEMENCIA O AGRESIVIDAD	
VERBOS QUE DENOTAN INCUMPLIMIENTO Y CON OTROS QUE DESIGNAN DIVERSAS FORMAS DE RECHAZO DE OBSTACULIZACIÓN DE UN PROCESO	*protestar,…*
VERBOS QUE DENOTAN OPOSICIÓN O RECHAZO, A MENUDO MANIFESTADAS VERBALMENTE	*condenar, protestar, rechazar,…*
VERBOS QUE DENOTAN LANZAMIENTO (A MENUDO HOSTIL, PERO NO NECESARIAMENTE). TAMBIÉN CON OTROS QUE DESIGNAN, POR EXTENSIÓN, CIERTAS MANIFESTACIONES VERBALES DE TONO AGRESIVO	*protestar,…*
VERBOS QUE DESIGNAN ACCIONES VERBALES QUE EXPRESAN CRÍTICA, OPOSICIÓN O RECHAZO EN DIVERSOS GRADOS	*protestar, condenar,…*
VERBOS QUE DESIGNAN DIVERSAS FORMAS DE RESISTIRSE	*condenar, contestar, protestar, reaccionar,…*

Tabla 3-25 Clases léxicas con mayor asociación de selección para enérgicamente (I).

Clases DCECR. en el corpus	
VERBOS QUE DENOTAN NEGACIÓN, OPOSICIÓN O RECHAZO, TAMBIÉN CON OTROS QUE DESIGNAN LA ACCIÓN DE ENFRENTARSE A ALGO O A ALGUIEN O LA DE CAUSARLE ALGÚN PERJUICIO	*negar, combatir, oponer,…*
VERBOS QUE DENOTAN PARTICIPACIÓN ACTIVA, A MENUDO EN ACCIONES CONTRARIAS A ALGO O A ALGUIEN	*protestar,…*
VERBOS QUE DENOTAN ENFRENTAMIENTO U OPOSICIÓN, TAMBIÉN CON ALGUNOS QUE DESIGNAN OTRAS ACCIONES DE CARÁCTER HOSTIL	*atacar, protestar,…*
VERBOS QUE DESIGNAN DIVERSAS MANIFESTACIONES VERBALES EXPOSITIVAS O DECLARATIVAS, MÁS FRECUENTEMENTE SI SE REALIZAN ANTE ALGUIEN CON AUTORIDAD O A PETICIÓN DE OTROS. TAMBIÉN CON ALGUNOS VERBOS QUE EXPRESAN LA ACCIÓN DE PRESENTARSE O ACUDIR A REALIZAR ESAS	*declarar, intervenir, hablar, pronunciar,…*
VERBOS QUE DENOTAN EXPRESIÓN O MANIFESTACIÓN VERBAL	*afirmar, contestar, decir, declarar,…*

Tabla 3-26 Clases léxicas con mayor asociación de selección para enérgicamente (II).

Llegados a este punto, se plantea la necesidad de filtrar de forma objetiva y automática los grupos de argumentos del diccionario ideológico con los que los predicados establecen un nexo preferente. Dado que se pretende captar la relación predicado-argumentos, entre una determinada forma canónica en el rol de predicado y los posibles grupos del diccionario ideológico, se deduce que estamos ante casos originados mediante un mecanismo diferente al que surgen la mayoría de los grupos que combinan con un predicado fijado. Es decir, esta relación genera restricciones de selección, que se reflejarán en los valores de la asociación de selección. Este razonamiento lleva a utilizar los mismos criterios aplicados a las combinaciones simples de formas canónicas, en las que el interés se centraba en aquellas que se pudieran considerar de uso preferente utilizando técnicas de detección de outliers. Todo ello conduce a aplicar tanto el valor *MAD* como *Box-Plot* sobre los valores de la asociación de selección obtenidos por los grupos registrados con la forma canónica sobre la que se haga la consulta. Se exponen los valores máximos de *MAD* para los ejemplos utilizados en este capítulo: *sentir, palpitar, brusco, doloroso, enérgicamente, abundantemente.*

Predicado	Grupo	Asociación de selección	Argumentos
SENTIR	INTRANQUILIDAD	63,48358912	*mar de fondo, intranquilidad, polvareda, revuelo, repunta, taco, inquietud, desasosiego, ansia, zozobra, ansiedad, agitación, estrés, preocupación, pesadilla, malestar, rebato, alarma, agobio, suspense, conflicto, torozón, hormigueo, comecome, comezón, desabrimiento, tártago, susidio, combate, guerra, batalla, mareta*
	DOLOR	57,53682436	*fuego, hormiguillo, hormiguilla, piquiña, fogaje, ardor, picazón, picor, quemazón, prurito, comezón, comecome, rascazón, escozor*
	HABLA	47,67730043	*sibilítico, sibilino, ambiguo, impreciso, figurado, traslaticio, lato, recto, literal*
	DOLOR	39,47676191	*tormento, punzada, dolor, analgesia, sinalgia, yaya, acritud, rayo, clavo, puntada, ramalazo*
	DESAGRADO	38,70660407	*fregado, mangangá, cosijoso, lipidioso, corrumpente, secante, molesto, importuno, latoso, pesado, engorroso, inconfortable, incómodo, chinchoso, endemoniado, duro, cansino, cansado, fastidioso, gravoso, oneroso, trabajoso, premioso, pijotero, fuñique, grave, jodido, dichoso, prolijo, matador, mortal, hediondo, guerrero, infernal, pajolero, odioso, insoportable*
	INDUCCIÓN	32,72506461	*aguijonazo, aguijón, anzuelo, espolada, espuela, pinchazo*
	SUFRIMIENTO	32,53837553	*pensión, desplacer, punzada, patatús, martelo, arestín, disgusto, puñalada, varapalo, sinsabor, desazón, tribulación, pesadumbre*
	ENOJO	31,69693666	*enfado, enojo, entripado, picazón, pique, escozor, ofensa, atufo, mosca, mohína, indignación*
	DESAGRADO	31,60851849	*repelencia, horror, repulsión, hámago, náusea, asco, repugnancia, grima, dentera*
	SUFRIMIENTO	31,39363334	*sufrimiento, hiel, agrá, agraz, acíbar, amargura, aflicción, pena, dolor, angustia, pesar, congoja, duelo, desconsuelo, sentimiento, apuro, ahogo, quebranto, presura, pasión, agonía*

Tabla 3-27 Grupos de selección preferente según MAD para el verbo sentir

Predicado	Grupo	Asociación de selección	Argumentos
PALPITAR	VALENTÍA	95,52566745	*hígados, redaños, arrestos, agallas, pecho, corazón*
	AGITACIÓN	88,7261206	*aleteo, salto, sístole, diástole, contracción, pulsación, palpitación, golpe, latido, pulso*
	CENTRO	85,46242544	*yema, ombligo, vientre, riñón, corazón, entraña*
	BENEVOLENCIA	84,92569175	*benevolencia, entraña, corazón, ternura, filantropía, humanidad*
	SENTIMIENTO	74,3173042	*sentimiento, podredumbre, psicología, entretelas, corazón, afección, sentir*
	DEDO	63,9634519	*pulgar, dedo, meñique, auricular, anular, corazón, medio, cordial, índice, dátil*
	VOLUNTAD	63,3948113	*voluntad, corazón, sentido, intención, merced, elección, albedrío, opción, facultad, potestad, libre albedrío, libertad, arbitrio*
	CORAZÓN	62,2941331	*corazón, pericardio, endocardio, válvula tricúspide, válvula mitral, ventrículo, aurícula, miocardio*
	PECHO	61,7956846	*pigidio, esternón, balancín, escólex, telson, escutelo, busto, pecho, seno, mama, teta, aréola, pezón, tetilla, ubre, chiche, mamila, metatórax, mesotórax, pronoto, protórax, cefalotórax, tórax*
	VENA	60,3168244	*hacecillo, vena, capilar, vaso, arteria*

Tabla 3-28 Grupos de selección preferente según MAD para el verbo palpitar

En los resultados extraídos respecto a los adjetivos brusco y doloroso se observa que ambos adjetivos se aplican a los mismos grupos, pero cada uno de ellos selecciona de forma preferente un único grupo que cumpla la restricción de $MAD \geq 4,5$ en el caso de brusco lo hace con: "*escaramuza, alborada, refriega, pugna, batalla, combate, liza, lucha, lid, choque, acción, fregado"* mientras que doloroso lo hace con: "*azote, flagelo, ramalazo, zurriagazo, tártago, rayo, desventura, infortunio, mal, desgracia, desdicha, desastre, tristezas, lacra, suceso, calamidad, fatalidad, malaventura, cataclismo, catástrofe, bancarrota, tragedia, drama, miseria, plaga, desmán, trago, tormenta, través, hecatombe"* (Tabla 3-29 y Tabla 3-30).

Predicado	Grupo	Asociación de selección	Argumentos	MAD
BRUSCO	GUERRA	0,00548033	*escaramuza, alborada, refriega, pugna, batalla, combate, liza, lucha, lid, choque, acción, fregado*	6,47377325
	MANO	0,00117275	*mano*	1,59933181
	TALLO	0,00073327	*tallo, carda, cladodio, pedicelo, estolón, bohordo, tronco, talo, lleta, guía, mástil, estipe, mata, pie*	1
	CORAZÓN	0,00044792	*corazón*	0,61084944
	REVÉS	0,00039014	*azote, flagelo, ramalazo, zurriagazo, tártago, rayo, desventura, infortunio, mal, desgracia, desdicha, desastre, tristezas, lacra, suceso, calamidad, fatalidad, malaventura, cataclismo, catástrofe, bancarrota, tragedia, drama, miseria, plaga, desmán, trago, tormenta, través, hecatombe*	0,53205471

Tabla 3-29 Grupos de selección preferente según MAD para el adjetivo brusco

Predicado	Grupo	Asociación de selección	Argumentos	MAD
DOLOROSO	REVÉS	0,00548033	*azote, flagelo, ramalazo, zurriagazo, tártago, rayo, desventura, infortunio, mal, desgracia, desdicha, desastre, tristezas, lacra, suceso, calamidad, fatalidad, malaventura, cataclismo, catástrofe, bancarrota, tragedia, drama, miseria, plaga, desmán, trago, tormenta, través, hecatombe*	4,46302398
	CORAZÓN	0,00117275	*corazón*	0,37723348
	GUERRA	0,00073327	*escaramuza, alborada, refriega, pugna, batalla, combate, liza, lucha, lid, choque, acción, fregado*	0,37723348
	TALLO	0,00044792	*tallo, carda, cladodio, pedicelo, estolón, bohordo, tronco, talo, lleta, guía, mástil, estipe, mata, pie*	1
	MANO	0,00039014	*mano*	1,1029913

Tabla 3-30 Grupos de selección preferente según MAD para el adjetivo doloroso

Predicado	Grupo	Asociación de selección	*Argumentos*
ABUNDANTEMENTE	SUDOR	482,4811368	*trasudar, resudar, sudar, transpirar*
	INDICIO	394,648837	*transpirar, transparentarse, adivinarse, revelar, repuntar, amanecer, dibujarse, perfilarse, indiciar, indicar, aflorar, asomar, salir, surgir, declararse, reflejarse*
	SUMINISTRAR	305,9141637	*alargar, proporcionar, dar, dotar, surtir, servir, proveer, suministrar*
	PREPARACIÓN	228,4413626	*aparejar, disponer, aprevenir, perdigar, mullir, poner, prevenir, preparar, armar, formatear, estructurar, organizar, proveer, reservar, apercibir, pertrechar, parar, conrear, aprestar, aviar, prevenir, preconcebir, sembrar, intentar, aprontar, avisar*
	FUERA	213,5840966	*trasvinar, revenir, instilar, rezumar, exudar, sudar, manar, efluir, fluir, brotar, destilar, resurgir, surgir, surtir, saltar, emanar*

Tabla 3-31 Grupos de selección preferente según MAD para el adverbio abundantemente

Predicado	Grupo	Asociación de selección	Argumentos
ENÉRGICAMENTE	COLOR	882,4697135	*sonrosar, purpurar, embijar, rubificar, embermejer, arrebolar, colorear, rojear, enrojecer*
		701,3689467	*invaginar, rebotar, redoblar, volver, trabar, triscar, engarabitar, engarabatar, escarzar, enarcar, ahorquillar*
	SUJECIÓN	700,4808225	*embrochalar, encofrar, entibar, apuntar, apuntalar, sujetar, atrancar, amacizar, lastrar, recibir, estacar, enramar, envigar, escorar, barretear, atarugar, enrabar, pinzar, enzarzar, encañar, ahorquillar, estaquear*
	NEGAR	170,062065	*repulsar, rehusar, denegar, desestimar, desaprobar, negar*
	REGOCIJO	88,92719291	*palmear, palmotear, loquear, troncharse de risa, desternillarse de risa, mondarse de risa, carcajear, sonreír, reír*

Tabla 3-32 Grupos de selección preferente según MAD para el adverbio enérgicamente

4. Aplicación desarrollada

4.1. Introducción

Se presenta en este capítulo la aplicación ColexWeb, recurso lexicográfico que hace accesible a través de la web el acceso a la información contenida en la base de datos que alberga toda la información combinatoria extraída del corpus. El diseño de ColexWeb permite consultar los datos respecto a las capacidades combinatorias de una determinada palabra y visualizarlos, tanto textual como gráficamente, obteniendo así una comprensión adecuada de los mismos. A través de ella también se pone a disposición de la comunidad un motor de extracción de colocaciones sobre cualquier texto que se proporcione, en el que se han implementado los procedimientos recomendados según la experiencia alcanzada. De esta manera se ofrece a través de internet el acceso a este caudal de información combinatoria del español, de especial utilidad en la enseñanza de la lengua, estudios de los lexicógrafos o tareas de eliminación de ambigüedad en el Procesamiento del Lenguaje Natural.

Se conjuga en la aplicación que aquí se presenta características de otras herramientas orientadas a la extracción de colocaciones como son Collocate, el Tollkit UCS o el diccionario Dice:

- *Collocate* es una aplicación de escritorio de pago que ha sido desarrollada para extraer colocaciones y términos a partir de un corpus textual, mediante el análisis estadístico de las mismas. Extrae las colocaciones de un corpus como un todo, permitiendo la selección del tamaño de ventana de análisis y la elección de los indicadores a calcular entre información mutua, t-score y la razón de verosimilitud.
- El *Tollkit UCS* desarrollado en Perl y su versión mejorada en el R-Project permiten el análisis estadístico de coocurrencia de datos. Extrae las combinaciones existentes en un texto, compuestas por dos palabras, asociadas a sus frecuencias de aparición, y almacenadas en ficheros de texto plano. Permite operaciones como la visualización e impresión de datos, manipulación, ordenación y asociación con un rango de medidas calculadas.
- *DiCE* es una aplicación web que representa un diccionario de colocaciones del español, con una nomenclatura limitada actualmente al campo semántico de los nombres asociados a sentimiento. Además de otra información como por ejemplo (cuasi)-sinónimos y (cuasi)-antónimos, se presentan para cada unidad léxica las colocaciones con información de las funciones léxicas.

ColexWeb abarca aspectos de todas ellas ya que permite la consulta de las capacidades combinatorias de las palabras del español (*DICE*), proporcionándola de forma extensional, enriqueciéndola con las posibilidades de ordenación según los valores de alguno de los indicadores (*Collocate)* y facilitando la interpretación y valoración de los datos a partir de técnicas de visualización (*Toolkit UCS*). Como novedades adicionales se incorporan los indicadores basados en la detección de Outliers además de información de las preferencias de selección entre un predicado y sus argumentos a través de los valores de la asociación de selección sobre el espacio semántico del D.I.V.

Los requisitos fundamentales en el desarrollo de esta herramienta radican en el funcionamiento a través de la web, el diseño de una interfaz de usuario amigable, la forma de acceso a la BDD a través de un servicio web dedicado, disponible para la integración en herramientas que desarrollen otros programadores y la optimización en términos de velocidad, intentando siempre obtener respuestas en un rango de tiempo adecuado.

4.2. Arquitectura del sistema

Los principales componentes del sistema se resumen en el diagrama de despliegue en la Figura 4-1

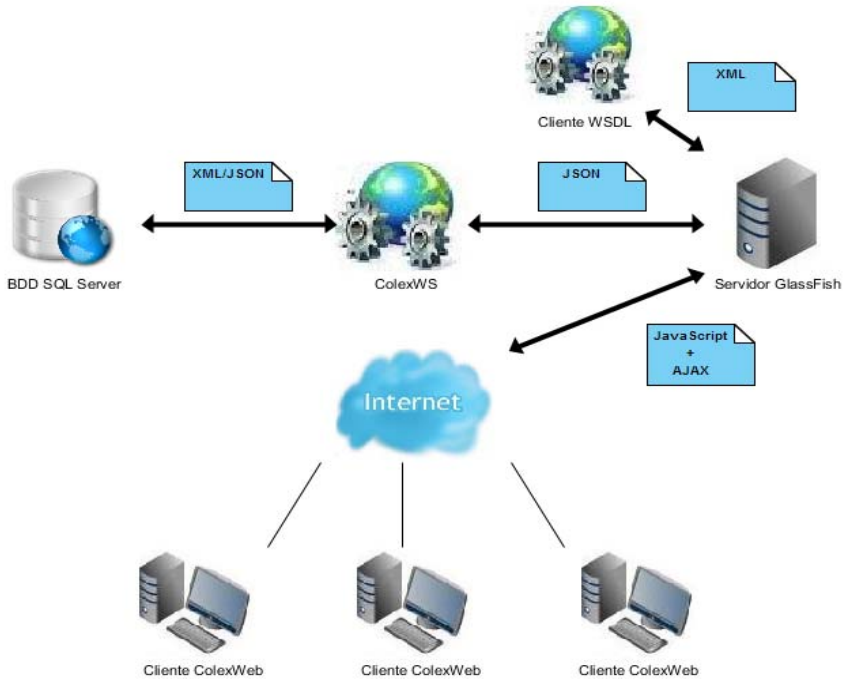


Figura 4-1 Diagrama de Despliegue

- **ColexWeb, aplicación web cliente.** Encargada de la interacción del usuario para realizar tanto el análisis de palabras sueltas como de un fichero de texto correctamente formateado.

- **ColexWS, Web Service que enlaza base de datos con ColexWeb.** Permite el acceso a la base de datos desde la aplicación web cliente, realiza operaciones de consulta para obtener los datos de indicadores seleccionados por el usuario. Sigue el protocolo REST.

- **Cliente Web Service Lematizador.** Servicio Web proporcionado por el Grupo de Estructura de Datos y Lingüística Computacional, GEDLC, de la Universidad de las Palmas de Gran Canaria. Se encarga de lematizar cualquier palabra del español identificando su forma canónica, categoría gramatical y la flexión o derivación que la produce, devolviendo las formas correspondientes. Es un web service de tipo SOAP, accedido a través del fichero WSDL, que a su vez genera las clases necesarias que pueden ser accedidas desde la aplicación web cliente.

- **Servidor contenedor de aplicaciones.** Actúa como contenedor de la aplicación web cliente y el web service que interactúa con la base de datos, además de procesar las peticiones realizadas por el cliente. Las operaciones de solicitud que realiza son las siguientes:

  - *Carga del fichero de texto involucrado en el análisis.*
  - *Análisis de combinaciones en un fichero de texto.*
  - *Consulta de indicadores generales.*
  - *Consulta de indicadores propios.*
  - *Cálculo de selección léxica.*

Sobre el servidor se ejecutan dos servlets: el primero establece el proxy de comunicación asíncrona que es el encargado de establecer la conexión con la aplicación web cliente, mientras que el segundo es un servlet común destinado a la carga del fichero a analizar cuando se realiza esta operación. El motivo de este segundo servlet se debe a la necesidad de realizar el parseo traducir del fichero en el lado servidor, ya que el cliente GWT no dispone de herramientas para ello.

- **Base de Datos.** Recoge todas las palabras del corpus analizadas en su contexto y clasificadas por tipos de combinaciones, con los datos de categorías, frecuencias e indicadores, entre otros, correctamente formateados, sobre el Sistema Gestor de Base de Datos Microsoft SQL Server 2008 R2.

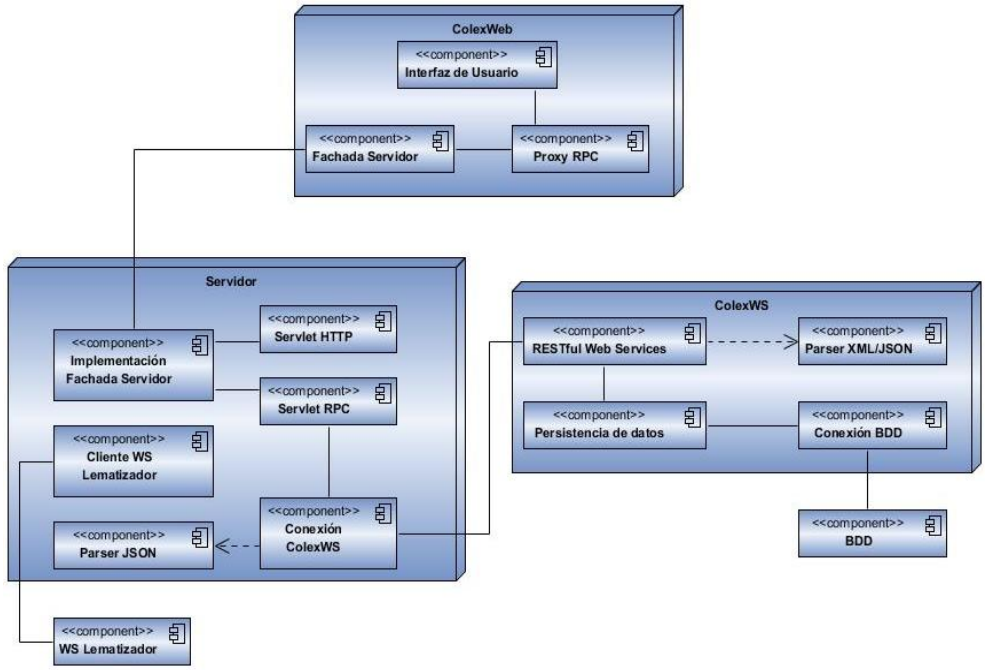


Figura 4-2 Diagrama de Componentes

El diagrama de componentes que aparece en la Figura 4-2 muestra la composición interna de cada uno de los elementos que intervienen en la arquitectura del sistema descrito:

4.3. Análisis de casos de uso

La aplicación web no precisa, hasta el momento, de un módulo dedicado a la identificación de usuarios para permitir su acceso ni restricción de procesos en base a perfiles. Por tanto, se considera un único actor que interactúa con la aplicación, denominado *lingüista*, el cual puede acceder a los procesos de análisis de palabras sueltas y procesado de un fichero de texto para extraer las combinaciones.

Teniendo en cuenta las consideraciones anteriores, se presenta el diagrama de casos de uso para la aplicación web cliente en la Figura 4-3.

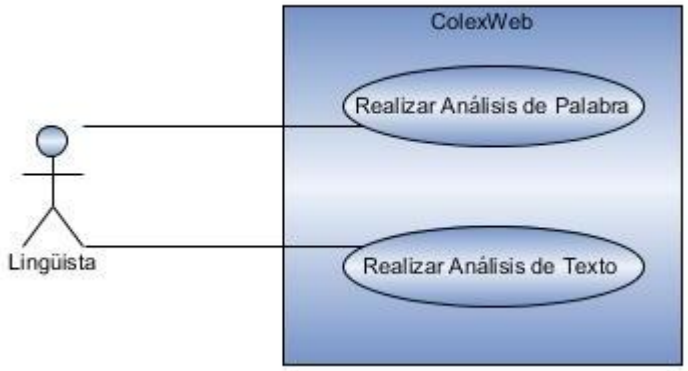


Figura 4-3 Diagrama de Casos de Uso

Se describen los casos de uso identificados:

- **Realizar análisis de palabra.** El lingüista introduce una determinada palabra del español a través de la aplicación, que solicita además seleccionar el indicador de consulta y el tipo de combinación a extraer de la base de datos de entre las permitidas por la palabra anterior. Se presenta así un conjunto de registros en la tabla con los datos solicitados, y una representación gráfica de los mismos, en el caso que se obtengan resultados en la base de datos.

- **Realizar análisis de un texto.** El lingüista indica el fichero que contiene el texto a analizar, además de especificar los indicadores que se desean calcular, el tipo de combinaciones a extraer y el tamaño de cinta para indicar la distancia máxima entre palabras. Tras realizar la carga del fichero en el servidor, se inicia el proceso de extracción. Se presenta así un conjunto de registros en la tabla con los datos solicitados, y una representación gráfica de los mismos, en el caso que se obtengan resultados con el tipo de combinación elegido.

4.4. ColexWeb: Aplicación web cliente

Este apartado está dedicado a la descripción de ColexWeb. Como se ha mencionado, representa la aplicación web cliente a la que tiene acceso el usuario lingüista y con la cual interactúa para obtener los resultados que desea en función del proceso lanzado.

Su construcción se ha realizado utilizando el *framework Google Web Toolkit*, *GWT*. GWT impone una división cliente-servidor, de forma que en el cliente se realice el diseño de la interfaz de usuario y posibles validaciones, y en el servidor se realicen operaciones relacionadas

con la interacción con el modelo de datos de la aplicación. Por tanto, se considera la división de ColexWeb mostrada en la Figura 4-4.

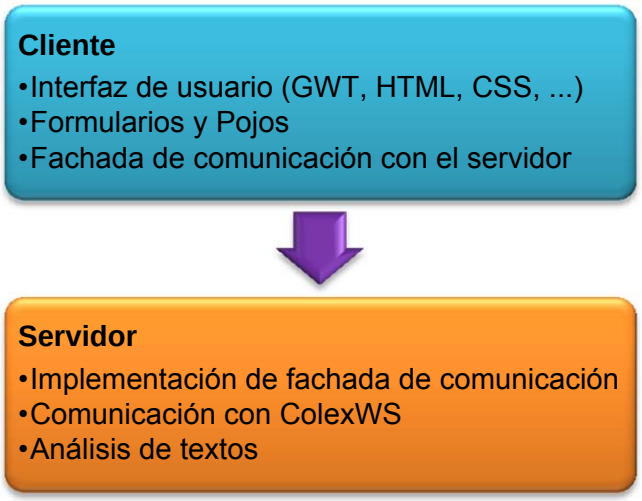


Figura 4-4 Estructura Cliente-Servidor

La interfaz de usuario que requiere ColexWeb no precisa de un gran despliegue, ya que requiere parámetros muy concretos para su funcionamiento en cualquiera de los procesos que ejecuta. Por ello, se busca simplicidad en el diseño para maximizar la fácil comprensión y usabilidad por parte del usuario. Dado que se ha orientado como una aplicación web en lugar de un conjunto de páginas, se dispone de una única pantalla principal sobre la que se realiza todo el trabajo que necesita el usuario, simulando así una aplicación de escritorio, pero compartida por la red. En la Figura 4-5 se observa pantalla inicial que puede visualizar un usuario.

Se identifican los elementos que se describen a continuación:

- Cabecera con título de la aplicación, situado en la parte superior del navegador.
- Menú desplegable ubicado en la zona central izquierda, justo debajo de la cabecera, con dos secciones diferenciadas para identificar los procesos: **Análisis de palabras** y **Análisis de un texto**. En la Figura 4-5 se muestra el correspondiente al análisis de palabras y en la Figura 4-6 al análisis del texto.
- Tabla que muestra los registros seleccionados, ubicada en la zona central derecha del navegador, justo debajo de la cabecera. Inicialmente, la tabla está vacía, se indica con el símbolo en gris la ausencia de resultados.

- Panel destinado a la representación de gráficas lineales para dibujar los datos presentados en la tabla anterior. Se ubica en la parte inferior de la pantalla, en blanco cuando ColexWeb ha sido iniciado.

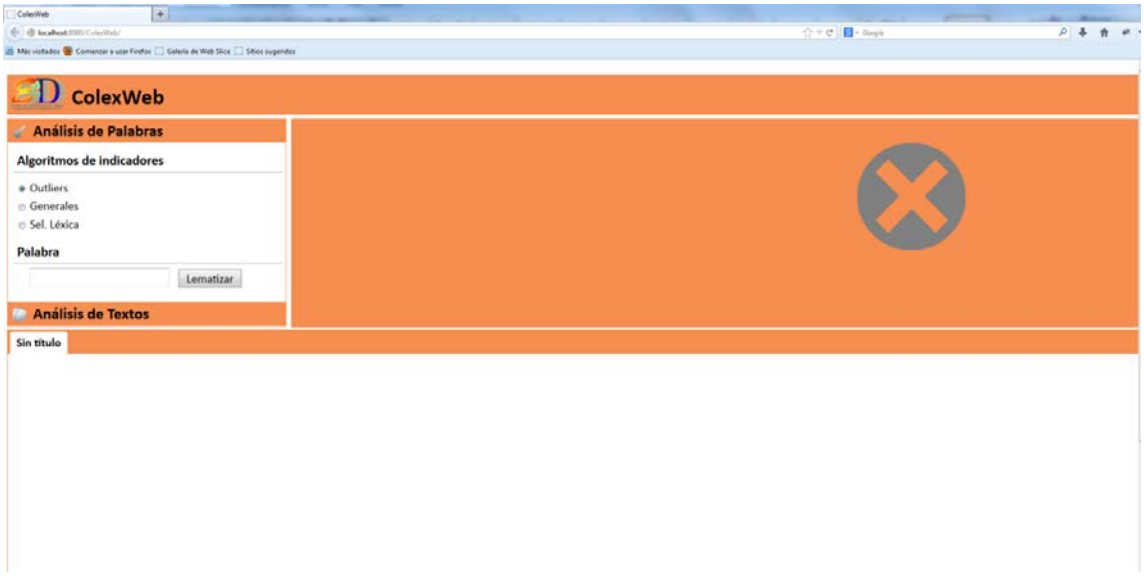


Figura 4-5 Pantalla inicial

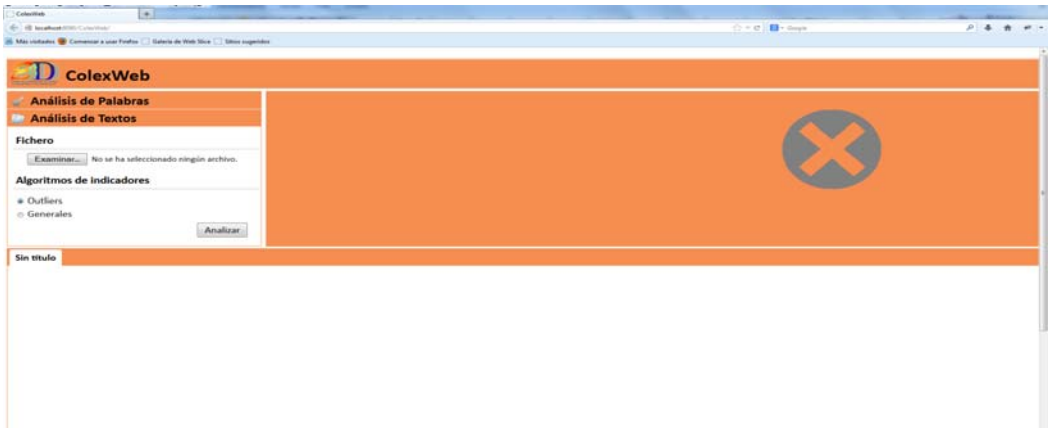


Figura 4-6 Panel que se muestra para el análisis de un texto.

4.4.1. Análisis de palabra

Se solicita la introducción de la palabra y el indicador o indicadores de interés. Se muestran los resultados de la lematización, especificando cada forma canónica resultante y su correspondiente categoría gramatical, y las categorías compatibles para construir una colocación sobre las que trabaja ColexWeb. El usuario puede realizar consultas sobre combinaciones cuyas formas canónicas tienen categoría gramatical: *verbo* + *sustantivo*, *sustantivo* + *adjetivo* y *verbo* + *adverbio*. Se permite seleccionar entre:

- Uno o varios indicadores de uso general: Frecuencia Relativa respecto a la base o el colocativo, Información Mutua, t-score y Z-Score (Figura 4-7).

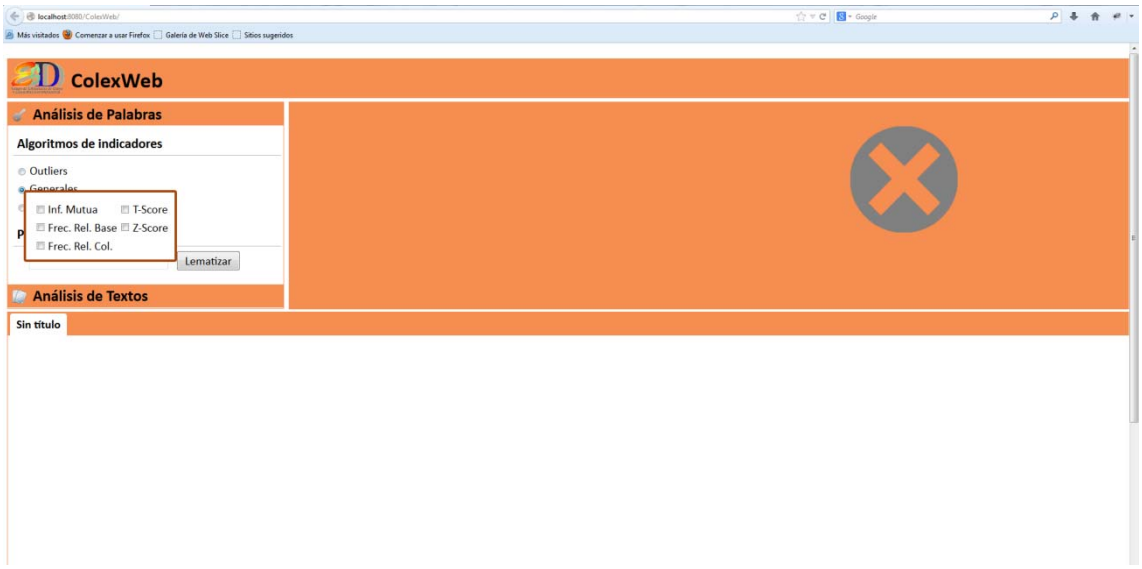


Figura 4-7 Análisis utilizando indicadores convencionales

- Indicadores basados en outliers: MAD y Box-Plot.
- Selección léxica.

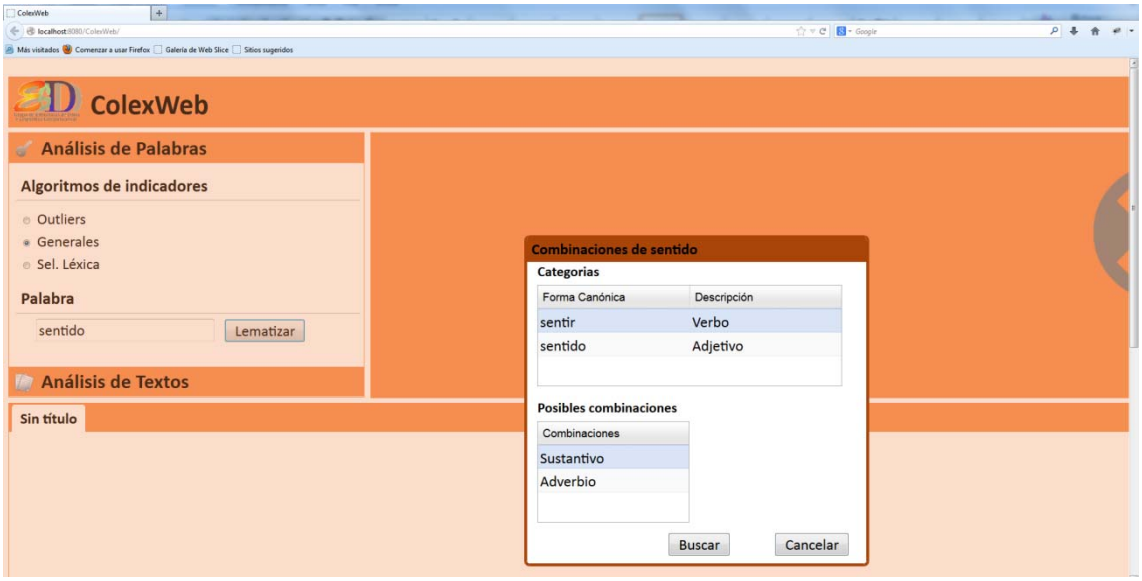


Figura 4-8 Lematización de la palabra y selección de combinaciones permitidas.

Una vez resuelta la consulta se muestra una tabla con las formas canónicas integrantes de la combinación y sus puntuaciones obtenidas del corpus según el indicador seleccionado. En

la parte inferior se muestra una gráfica con los resultados que facilita el análisis. El flujo se repite para la selección léxica y outliers, con pinceladas características de cada caso.

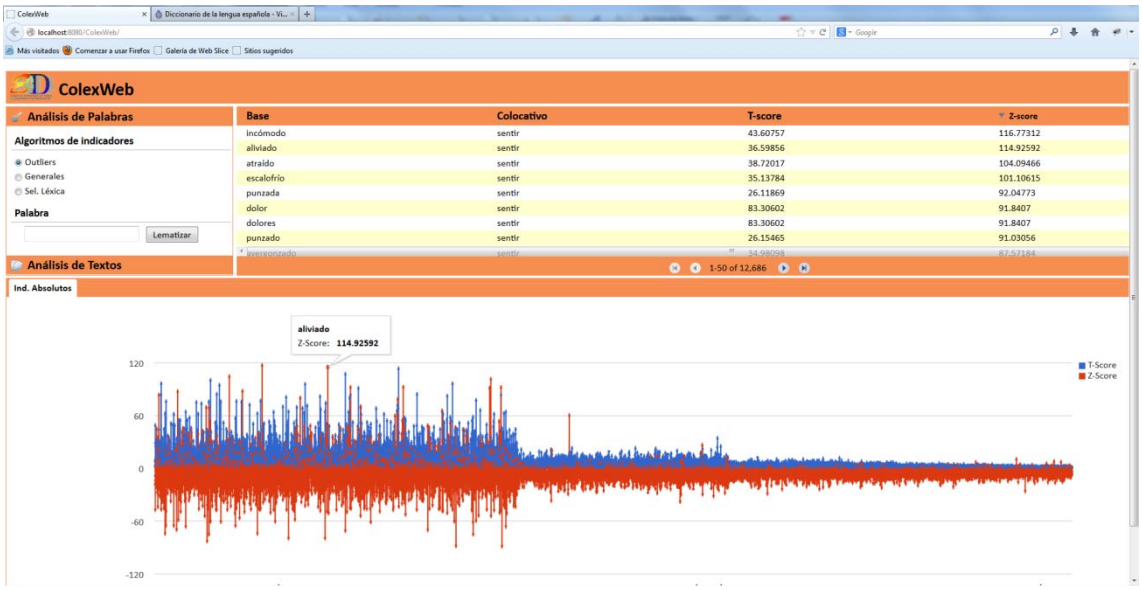


Figura 4-9 Resultados para la consulta sentido → sentir + sustantivo. Indicadores z-score y t-score

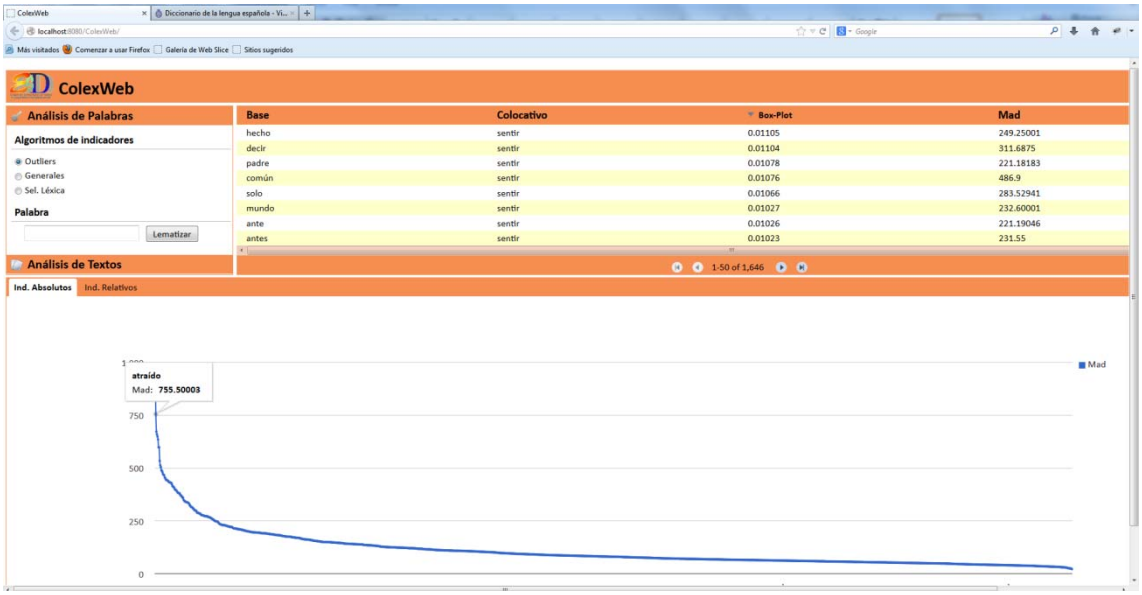


Figura 4-10 Resultados para la consulta sentido → sentir + sustantivo. Indicadores basados en outliers

Los resultados en la tabla se pueden mostrar ordenados por orden alfabético o por el valor numérico, según el usuario haga click sobre alguna de las columnas de formas canónicas o de indicadores. Si se sitúa el ratón sobre algún punto de la gráfica que se considere de interés se puede obtener la información de la combinación a la que hace referencia y su puntuación. En el caso de consultar sobre la Selección Léxica la información que se muestra en el globo es el conjunto de palabras que conforman el grupo semántico representado por ese punto.

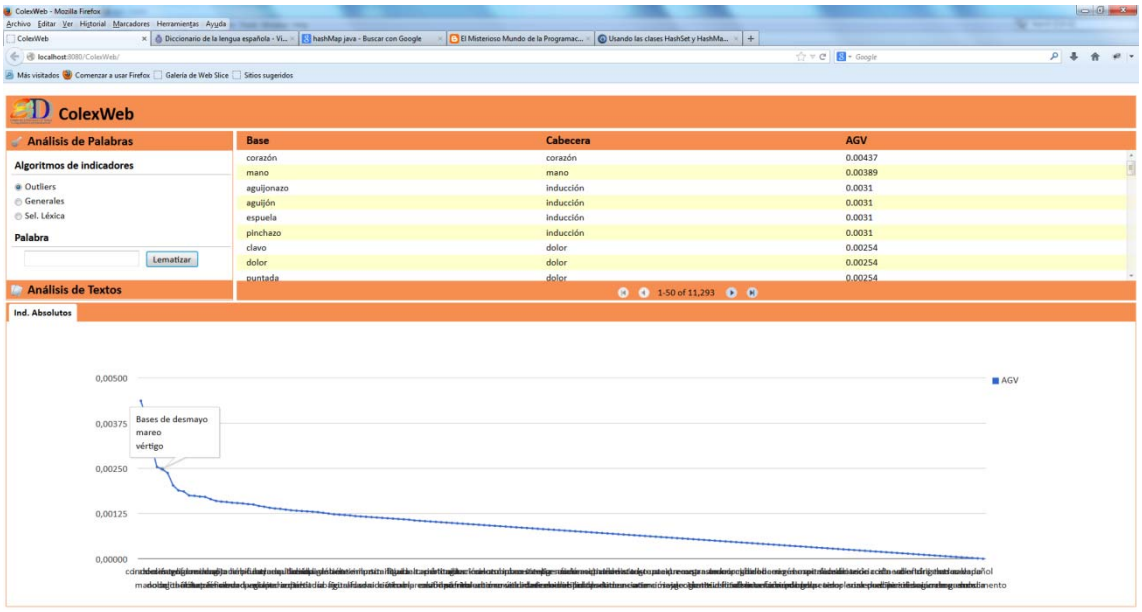


Figura 4-11 Ejemplo de consulta sobre selección léxica.

4.4.2. Análisis de texto

En el caso de análisis del texto, los parámetros solicitados son el fichero de texto a procesar y los indicadores que se desean extraer; en este caso, se permiten los indicadores generales: *Frecuencias Relativas, Información Mutua, t-score y Z-Score* y también los basados en outliers. A continuación, se abre una ventana en la que poder indicar el tipo de combinaciones a extraer y el tamaño de la concordancia. Las concordancias se restringen al ámbito de las oraciones (Figura 4-12, Figura 4-13). Al igual que el proceso anterior, se presenta la tabla de datos con los indicadores seleccionados y una ventana previa que indica el tiempo de ejecución empleado, total de combinaciones extraídas, número de palabras analizadas en el texto y total de veces que se ha invocado el web service lematizador.

En cuanto a la creación de gráficas, el funcionamiento es distinto. La diferencia radica en que no se tiene una palabra principal sobre la que generar el conjunto de resultados, sino que puede ser cualquiera que forme una colocación. De esta manera, se permite generar las gráficas correspondientes a la selección del usuario sobre las celdas de las columnas para base y colocativo. Por ejemplo, si se elige la palabra estrellar, colocativo en la combinación de tipo Sustantivo-Verbo (estrellar, presentado), se obtienen todas aquellas combinaciones en las que estrellar es colocativo y genera el conjunto de gráficas asociadas en función de los indicadores elegidos. El flujo, asociado al ejemplo descrito, es el siguiente:

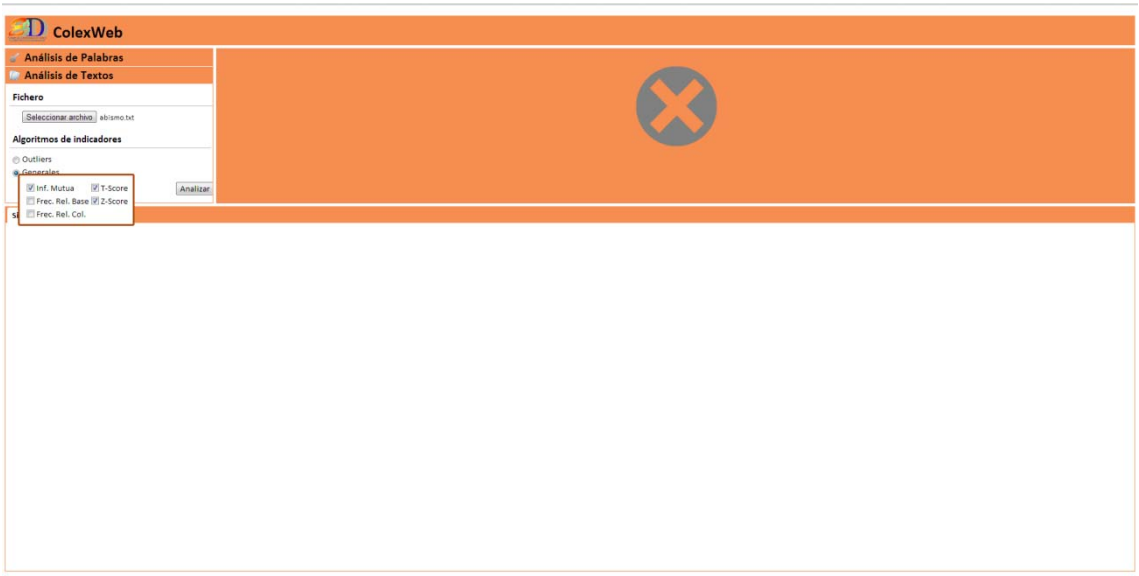


Figura 4-12 Inicio del análisis de textos.

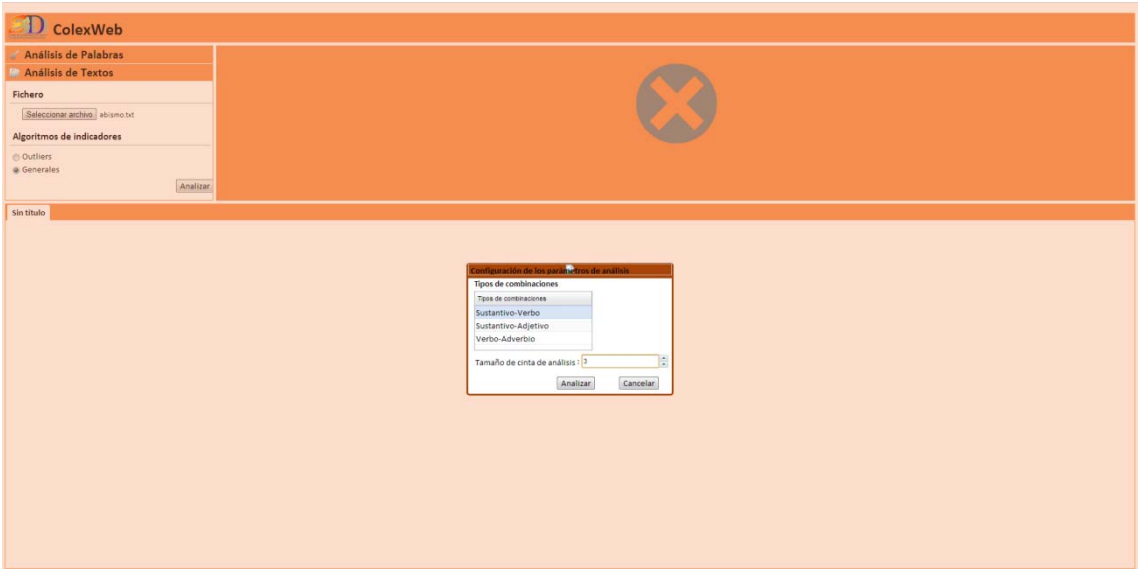


Figura 4-13 Configuración del análisis de textos.

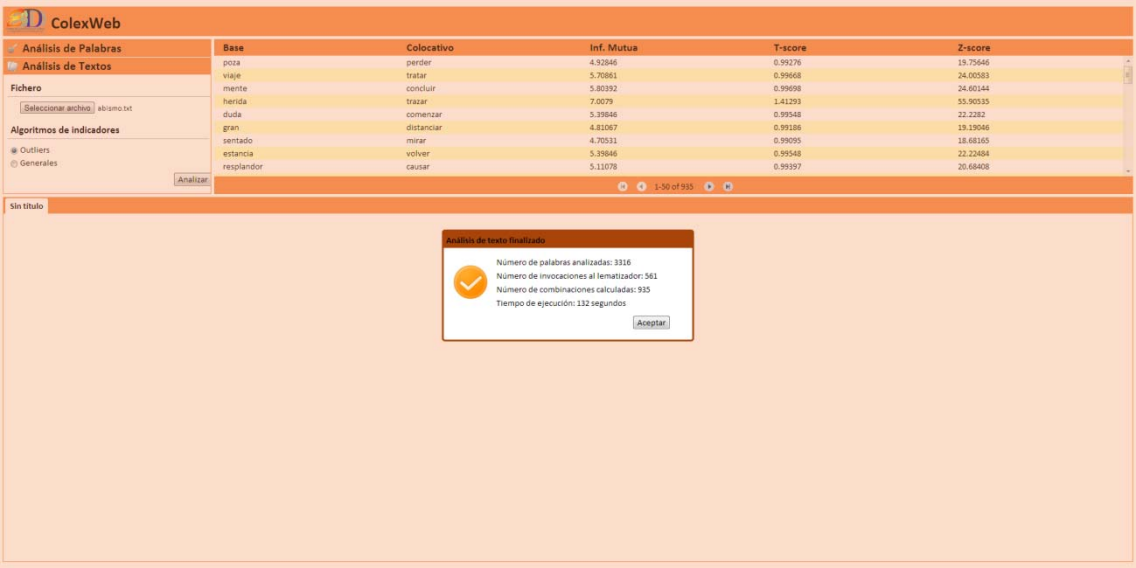


Figura 4-14 Resultados obtenidos del análisis de texto.

Durante la ejecución de los procesos, se presentan otras ventanas auxiliares. Dado que las operaciones de consulta y procesamiento de textos requieren un cierto tiempo de recuperación y presentación de datos, se visualizan diálogos de carga indicando el proceso realizado, que impide la interacción con la aplicación mientras se resuelve. También se realiza un control de errores sobre los datos introducidos, tales como la existencia de la palabra o del fichero de texto a analizar, se exige especificar como mínimo un indicador general si se solicita obtenerlos, se controla la imposibilidad de lematizar una palabra, etc.

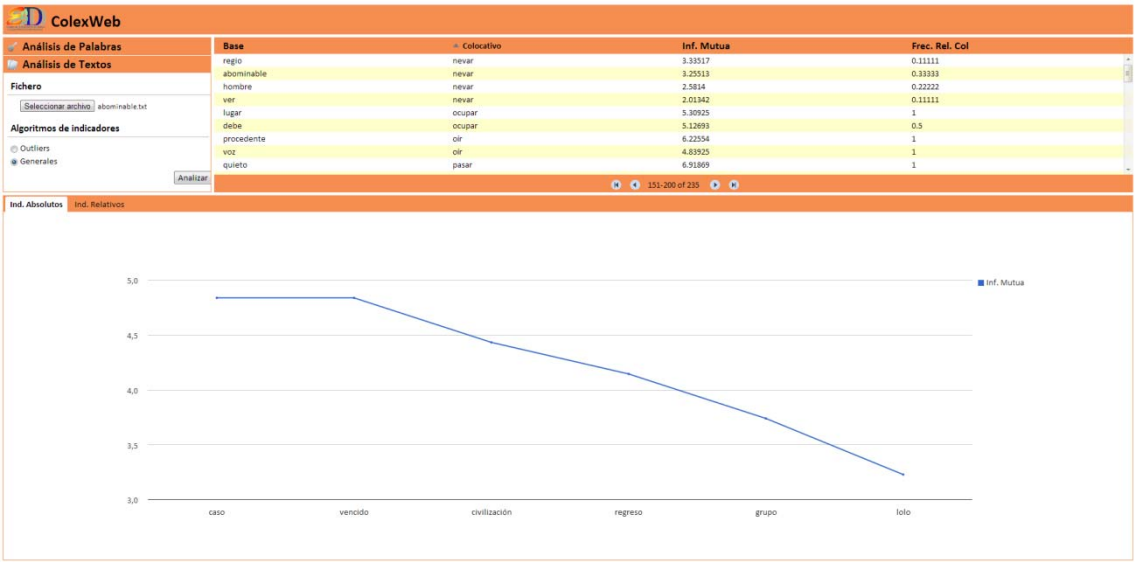


Figura 4-15 Resultados obtenidos del análisis de texto.

4.5. Elementos para acelerar el análisis de textos.

El análisis de los textos se lleva a cabo aplicando la misma metodología seguida para el procesamiento del corpus con el que se ha generado la BDD de combinaciones. Sin embargo, un factor crítico a tener en cuenta es el tiempo de respuesta en el procesamiento, por lo que además de unir estructuras con un acceso lo más rápido posible, se hace necesario construir un algoritmo que ofrezca una destacada eficiencia. A este fin, además de la lista de palabras vacías ya descrita en el capítulo 2, se ha generado un diccionario de palabras frecuentes.

Las palabras frecuentes consisten en un conjunto de palabras asociadas a una lista con todas las formas canónicas que produce su lematización, así como la categoría gramatical de la misma. Antes de lematizar una palabra se comprueba si ya existe en el diccionario, obteniendo los datos pertinentes de esta estructura local. El objetivo es evitar el coste computacional de ejecutar el lematizador cada vez que se necesite obtener las formas canónicas de una palabra. Debido al gran tamaño de este conjunto, su inicialización se delega al momento de inicio de la aplicación.

Este conjunto se ha extraído a partir de un corpus compuesto por 1.000.000 de palabras aproximadamente. Se han contabilizado las frecuencias de palabras no vacías utilizadas, extrayendo 62802 palabras diferentes, donde las 1000 más frecuentes representan el 50% del total empleado, y las 7000 más frecuentes cubren el 80%, como se ilustra en el gráfico de frecuencias acumuladas presentado en la Gráfico 4-1. Del mismo modo, a medida que se procesa un fichero de texto, se va actualizado la lista de palabras frecuentes con aquellas que aparecen en el corpus y que no forman parte del diccionario inicial.

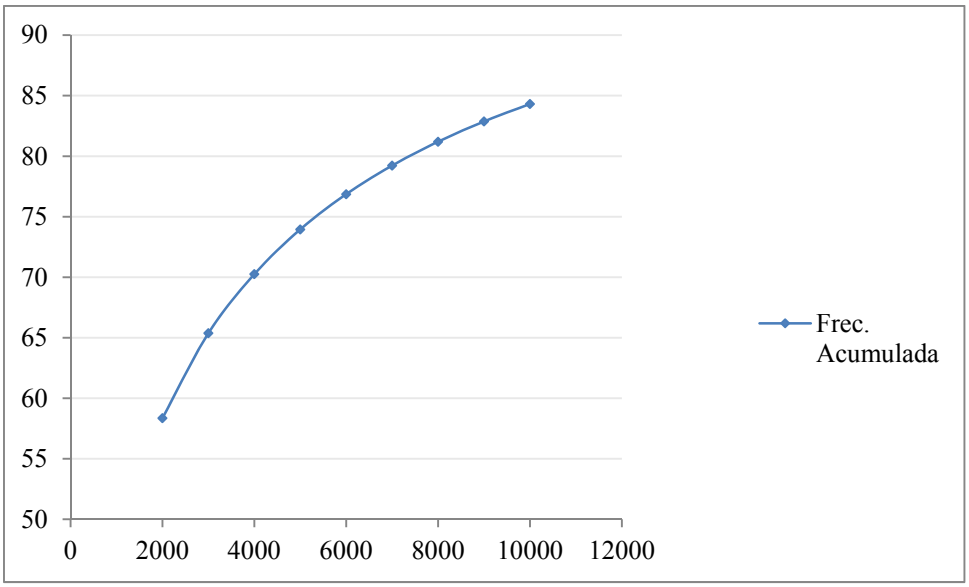


Gráfico 4-1 Frecuencias acumuladas para las palabras del corpus analizado

El Gráfico 4-2 y el Gráfico 4-3 muestran los tiempos invertidos en la ejecución del análisis de texto sobre varios ficheros de 6K, 12K, 113K, y 393K con una cantidad de 1011, 1955, 18724 y 64709 palabras respectivamente. Se observa que no necesariamente un mayor número de palabras conlleva más tiempo de procesamiento debido a la estrategia de incorporación de palabras frecuentes en tiempo real. Es decir, una vez que ha sido procesado un fichero, se mejora el rendimiento del programa, esto se debe por una parte, a que ya han sido cargados los recursos, establecido las conexiones, etc. Por otra parte, se disminuye la cantidad de invocaciones al lematizador, no solo por las palabras precargadas, sino por las que se van agregando en tiempo de ejecución. Finalmente, el tiempo de proceso lo consume la exploración de posibles combinaciones, ya que los tiempos de proceso no presentan diferencias extremas según la estructura de la colocación, pero hay una diferencia importante en la cantidad de combinaciones generadas según se trate de *Sustantivo + Verbo*, *Sustantivo + Adjetivo* o *Verbo + Adverbio*.

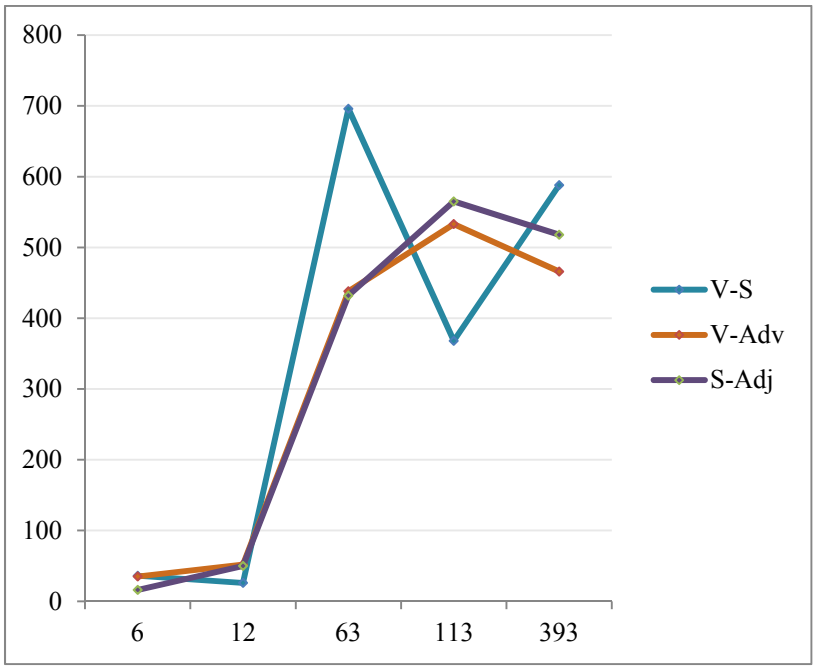


Gráfico 4-2 Tiempo de ejecución de ColexWeb para la Frecuencia Relativa sobre ficheros de distintos tamaños

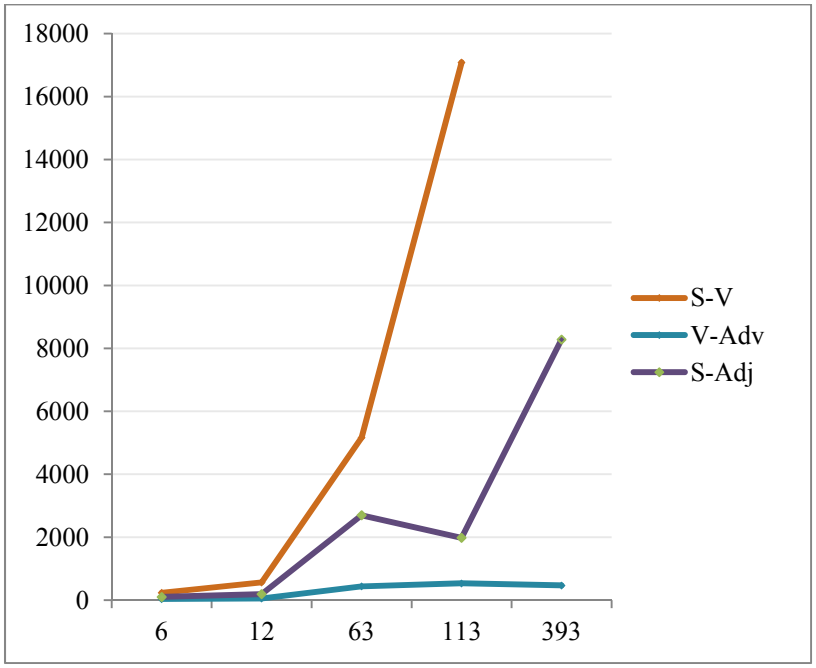


Gráfico 4-3 Combinaciones extraídas en ficheros de distintos tamaños.

5. Conclusiones

En esta tesis se ha abordado el estudio de la extracción automática de colocaciones léxicas del español a partir de corpus textuales extensos.

1. En primer lugar se ha realizado una recopilación de características lingüísticas de este tipo de combinaciones con objeto de utilizar aquellas que pudiesen repercutir en el tratamiento computacional del problema. En esta fase se concluyó que las palabras objetivo debían ser: nombres, adjetivos, verbos y adverbios, en las estructuras contenidas en la clasificación de colocaciones simples aportada por Koike (2002). Por otra parte, se considera esencial la flexibilidad formal, por lo que se opta por trabajar en el espacio de las formas canónicas en lugar de hacerlo en el espacio de las palabras. La característica de la preferencia, es decir, que la colocación se elija mayoritariamente por los hablantes frente a otras combinaciones posibles, se traduce en el enfoque basado en las frecuencias que tradicionalmente se ha dado al problema en cuestión. Esta última constituye el principal instrumento para el procesamiento de corpus textuales en busca de colocaciones. También destacan los numerosos casos en los que la asociación se produce entre una palabra y un grupo de palabras con rasgos semánticos comunes, característica directamente relacionada con la cualidad de los predicados de imponer restricciones sobre los posibles argumentos de los que predican, hecho que explica la teoría lingüística de la selección léxica.
2. Respecto a los recursos que se utilizaron para el desarrollo del trabajo, se compiló un corpus de textos sin ningún tipo de información lingüística de un volumen tal que pueda considerarse una amplia muestra del uso del español (300000000 de palabras). También se dispuso de otro corpus de menor extensión, formado por 2000000 de palabras que fue utilizado para realizar estudios comparativos. La transformación de ambos al espacio de las formas canónicas se realizó mediante el lematizador del GEDLC. Como datos de control, que permitan verificar los resultados que se obtienen, se recopilaron todos los ejemplos de colocaciones presentados en artículos lingüísticos sobre colocaciones en español, además de las combinaciones en el DCECR. Otro recurso relevante en este trabajo lo constituye el DIV, como base de conocimiento respecto a grupos semánticos.
3. El corpus fue procesado para obtener todas las combinaciones de formas canónicas que corresponden a alguna de las estructuras objeto de estudio mediante una aplicación en la que el contexto de una palabra se restringe a la frase limitando la línea de concordancia a una amplitud de 10. Se registraron las formas canónicas y sus frecuencias, las combinaciones de formas canónicas que al menos aparecen 3 veces y sus frecuencias

en un BDD, esta fue utilizada para las investigaciones de las técnicas computacionales y para el desarrollo de la aplicación.

4. La vertiente computacional del problema se orientó a la recopilación y evaluación de un amplio catálogo de indicadores basados en la frecuencia de uso, abundantemente utilizados en la literatura. Las medidas de asociación que se probaron incluyen la más simple, frecuencias relativas, la información mutua y las basadas en test estadísticos: z-score, t-score y test de Dunning. Las combinaciones analizadas pertenecen a alguno de los tres grupos que se pueden establecer a partir del lematizador del GEDLC y que se adaptan a las estructuras colocacionales presentes en el español: sustantivo + verbo, sustantivo + adjetivo y verbo + adverbio.
5. El análisis exhaustivo de los datos revela las dificultades de la automatización de este proceso tradicionalmente basado en rankings o en puntos de corte. Por un lado, se establece que las conclusiones son fiables en la medida en que se disponga de suficiente cantidad de muestras de las palabras en el corpus tratado. Por otro lado, la posición de una combinación en el ranking depende, en mayor o menor medida según el indicador que se fije, del uso más o menos elevado de la palabra, normalmente debido a su valor funcional o a la cantidad de acepciones de la misma o a la estructura gramatical de la colocación. Determinar la cantidad de ejemplares de una combinación que se exige que haya en el corpus se realiza manualmente, así como el punto de corte que delimita las combinaciones libres de las colocaciones.
6. De los experimentos analizados sobre el corpus de Galdós, se desprende la diferencia en los valores de frecuencias registrados y en las medidas de asociación alcanzadas por las combinaciones. Se resalta también que la disparidad entre las combinaciones en las que intervienen palabras de baja frecuencia de uso y las que son muy frecuentes se agranda a medida que crece el tamaño del corpus, mostrándose el test de Dunning como el más estable entre todos ellos. Hay que hacer notar la dificultad para encontrar muestras suficientes de una determinada colocación en un corpus por muy extenso que sea, ya que solo aproximadamente la mitad de las combinaciones en el DCECR aparecen en el corpus completo, reduciéndose drásticamente la cantidad de las que cumplen con el requisito del número mínimo de muestras en el de Galdós.
7. Se propone una metodología en la que la extracción se realiza de forma independiente para cada palabra, evitando las comparaciones entre formas canónicas diferentes desde el punto de vista de la frecuencia. Se aborda el problema desde el enfoque de la detección de valores atípicos en las muestras, comparando distintas técnicas utilizadas en estadística para resolver el problema de la detección de outliers. La propuesta que se hace tiene el objetivo de detectar, fijada una palabra, con qué otras se puede establecer que su uso se sale de lo común en su ámbito, por lo que se determinan qué valores

extremos alcanzan las frecuencias relativas de la forma canónica seleccionada. Se han valorado tres posibilidades que se denominan: ZChebyshev —se basa en la media aritmética y en la desviación típica, MAD —se basa en la mediana y Box-Plot —se basa en los cuartiles, se incluyen estos últimos por ser más robustos. El primer caso tiene menor cobertura; sin embargo, el estadístico MAD y Box-Plot se revelan como indicadores fiables para la extracción automática de colocaciones, ya que proporciona como principal ventaja la precisión de los resultados, filtrando una gran cantidad de combinaciones libres que se colaban en las mejores posiciones de los rankings. Además proporcionan un valor objetivo en el que establecer el corte, si bien en la frontera se introducen combinaciones libres, las frecuencias relativas coinciden con valores similares a los que registran combinaciones del DCECR.

8. En la inspección de los datos se dejan entrever grupos de palabras relacionadas gracias a cierta afinidad semántica, como es de esperar según la teoría lingüística. Se indaga sobre fórmulas computacionalmente viables para deducir automáticamente cuáles son los que corresponden a los predicados presentes en el corpus. Se valoran técnicas basadas en matrices de contexto para cada palabra: análisis clúster y semántica latente. En ambos casos, debido a la dimensión del corpus, la construcción de las matrices es muy costosa en tiempo y además requiere la interpretación de los resultados para establecer los grupos. Por estas razones se considera que estas soluciones no son apropiadas para una aplicación software en tiempo real como la que marcan los objetivos de este trabajo.
9. Se propone como solución viable y efectiva para captar los nexos motivados por la selección léxica, utilizar la medida de asociación entre verbos y sustantivos de Resnik (1999), sustituyendo WordNet por los grupos semánticos que aparecen en los diccionarios ideológicos, para los que se determina una puntuación denominada asociación de selección. A partir de este valor se pueden determinar rankings de los grupos semánticos de tales recursos léxicos con los que un predicado establece un nexo semántico. Se encuentra que la asociación de selección tiene capacidad de discriminar los grupos semánticos que combinan con un determinado predicado: la mayor parte de los grupos que contienen alguna de las palabras que aparecen junto al colocativo obtienen una puntuación casi nula y siempre se observa un conjunto de grupos del DIV que destaca por su alto valor de asociación de selección. Incluso cuando ninguno de los elementos del grupo tenga una marcada preferencia individual, el conjunto de combinaciones aporta la suficiente fuerza de selección para hacer destacar al grupo. También se valoró sobre los grupos semánticos en el DCECR, se observa que los que se presentan en dicha obra para el predicado en cuestión se posicionan entre las de más alta puntuación. En general, otros grupos con puntuaciones elevadas hacen referencias a

argumentos que comparten rasgos semánticos con los que se combinan con él según este diccionario aunque se presenten en relación con otras entradas.

10. Todo el conocimiento adquirido en el desarrollo de este trabajo se pone a disposición de la comunidad a través de la web mediante la herramienta ColexWeb para la consulta de las capacidades combinatorias de las palabras del español basada en la información extraída del corpus. Mediante una interfaz amigable el usuario puede diseñar la consulta, escogiendo la palabra, alguna de las estructuras admisibles para ella y qué medidas de asociación se quiere utilizar para construir el ranking. Los resultados se muestran por extensión y gráficamente, lo que facilita su comprensión de forma rápida y efectiva. Los indicadores disponibles son: *frecuencias relativas, información mutua, z-score, t-score, Box-Plot* y *MAD*. Del mismo modo se admiten peticiones respecto a los grupos semánticos que selecciona un predicado. También se incorpora en ColexWeb una opción para el procesamiento de un fichero de texto plano siguiendo la metodología llevada a cabo para el corpus del GEDLC, mostrándose los resultados de forma similar a como lo hace la opción de consultas.

6. Bibliografía

AGIRRE, E., ARREGI, X., DÍAZ DE HARRAZA, A. AND SARASOLA, K. (1994). "Conceptual Distance and Automatic Spelling Correction" in *Workshop on Speech recognition and handwriting*. Leeds, England.

AGIRRE, E., RIGAU, G. (1995). "A Proposal for Word Sense Disambiguation using conceptual Distance", *International Conference on Recent Advances in Natural Language Processing*, Tzigov Chark, Bulgaria.

AGIRRE, E. AND MARTÍNEZ, D. (2001). "Learning class-to-class selectional preferences". In *Proceedings of the 2001 workshop on Computational Natural Language Learning,* Vol. 7. Association for Computational Linguistics, Stroudsburg, PA, USA, Article 3, 8 pages.

ALMELA, M. (2011). "The case for verb-adjective collocations: corpus-based analysis and lexicographical treatment". *Revista de Lingüística y Lenguas Aplicadas*, Vol. 6, págs. 39-52.

ALONSO RAMOS, M. (1994-1995) "Hacia una definición del concepto de colocación: De J. R. Firth a I. A. Mel'čuk", en *Revista de Lexicografía* Vol. I, págs. 9-28.

ALONSO RAMOS, M. (2002). "Colocaciones y contorno de la definición lexicográfica", en *Lingüística Española Actual* XXIV/1, págs. 63-96.

ALONSO RAMOS, M. (2002). "Diccionario de Colocaciones del Español, DICE". [Internet] [Consultado en: 2-12-2013]. Disponible en: http://www.dicesp.com/paginas ISBN: 978-84-693-9869-2

AGGARWAL, C. (2013), *Outlier Analysis*, Springer Science & Business Media.

BARGALLÓ ESCRIVÁ, M., CARAMÉS DÍAZ, J., FERRANDO ARAMO, V., MORENO VILLANUEVA, J.A., (1997-1998) "El tratamiento de los elementos lexicalizados en la lexicografía española monolingüe", en *Revista de Lexicografía*, Vol. IV, págs. 49-65.

BLASCO MATEO, E. (2002) "La lexicalización y las colocaciones", en *Lingüística Española Actual* XXIV/1, págs. 63-96.

BOSQUE, I. (2001) "Sobre el concepto de colocación y sus límites", en *Lingüística Española Actual* XXIII/1, págs. 9-40.

BOSQUE, I. (2 004). *REDES Diccionario combinatorio del español contemporáneo*. Ediciones SM.

BRAUN, E. (1 996) "Caos, Fractales y Cosas Raras". *Fondo de Cultura Económica*. México.

BUDANISTSKY, A. (1 999), "Lexical semantic relatedness and its application in Natural Language Processing". *Technical Report CSRG-390*, Computer Systems Research Group, University of Toronto.

CASARES, J., (1 999) *Diccionario Ideológico de la Lengua Española*. Editorial Gustavo Gili, SA.

CASTILLO CARBALLO, M.A. (1 997-1 998), "El concepto de Unidad Fraseológica" en *Revista de Lexicografía*, Vol. IV, págs. 67-79.

CASTILLO CARBALLO, M. A. (1 998), "El término 'colocación' en la lingüística actual" en *Lingüística Española Actual* XX/1, págs. 41-54.

CASTILLO CARBALLO, M. A. (2 001), "Colocaciones léxicas y variación lingüística: implicaciones didácticas", en *Lingüística Española Actual* XXIII/1, págs. 133-143.

CHURCH, K. W. AND HANKS, P. (1990), "Word association norms, mutual information, and lexicography". *Computational. Linguistic*. 16, 1 págs. 22-29.

CORPAS PASTOR, G. (1 996). *Manual de Fraseología española*, Madrid, Gredos.

CORPAS PASTOR, G. (2 001), "Apuntes para el estudio de la colocación" en *Lingüística Española Actual* XXIII/1, págs. 41-56

CORPAS PASTOR, G. (2 003), "Diez años de investigación en fraseología: análisis sintáctico-semánticos, contrastivos y traductológicos". *Lingüística Iberoamericana*, Vol. 20, Iberoamericana – Vervuert.

Diccionario de sinónimos y antónimos, (2 005) Espasa-Calpe. [Internet] [Consultado en: 2-12-2011] en: http://www.wordreference.com/sinonimos/

ALVAR EZQUERRA, M.(DIR.) (1 998) *Diccionario Ideológico de la Lengua Española,* 2ª ed. VOX-Bibliograf.

EVERT, S. (2 005). "The statistics of word coocurrences. Word pairs and collocations".

Dissertation, Stuttgart University.

GARCÍA PLATERO, J. M. (2002) "Aspectos semánticos de las colocaciones", en *Lingüística Española Actual* XXIV/1 2002, págs. 25-34.

GARCÍA-PAGE, M. (2001), "El adverbio colocacional", en *Lingüística Española Actual* XXIII/1 2001, págs. 89-106.

JORGE-BOTANA, G. (2010). "La técnica del Análisis de la Semántica Latente (LSA/LSI) como modelo informático de la comprensión del texto y el discurso. Una aproximación distribuida al análisis semántico". Tesis doctoral. Universidad Autónoma de Madrid.

JORGE-BOTANA, G., LEÓN, J. A., OLMOS, R., HASSAN-MONTERO, Y. (2010). "Visualizing polysemy using LSA and the predication algorithm". *Journal of the American Society for Information Science and Technology*. Vol. 61(8) , pages1706-1724

LIU, H., SHAH, S. & JIANG, W. (2004). "On-line outlier detection and data cleaning". *Computers & chemical engineering*, Vol. 28(9), pages 1635-1647.

HIGUERAS GARCÍA, M. (2006), "Estudio de las colocaciones léxicas y su enseñanza en español como lengua extranjera" en *ASELE, Colección monografías*, nº 9, Málaga

KOIKE, K. (2001), *Colocaciones léxicas en español*, Universidad de Alcalá, Takushoku University.

KOZIMA, H. AND FURUGORI, T. (1993). "Similarity between words computed by spreading activation on an English dictionary". In *Proceedings of the sixth conference on European chapter of the Association for Computational Linguistics,* pages 232-239.

LANCIA, F. (2007). "Word co-occurrence and similarity in meaning*"*. In Salvatore, S., Valsiner, J. (Eds.) *Mind as infinite dimensionality.* Edizioni Carlo Amore, Rome.

LANDAUER, T. K., FOLTZ, P.W., LAHAM, D. (1998) "An introduction to latent semantic analysis". *Discourse Processes*, 25(2-3), pages 259-284.

LIN, D. (1998) "An information-theoretic definition of similarity". In *Proceedings of the 15th international conference on Machine Learning*, Vol. 1, pages 296-304.

LEYS, C., LEY, C., KLEIN, O., VERNARD, P., LICATA, L. (2013) "Detecting outliers: Do not use standard deviation around the mean, use deviation around the median". *Journal of Experimental Social Pyschology*, 49, pages. 764-766.

LIN, F. AND COHEN, W. W. (2010). A very fast method for clustering Big Text Datasets. In Helder Coelho, Rudi Studer, and Michael Wooldridge (Eds.) *Proceedings of the 2010 conference on ECAI 2010: 19th European Conference on Artificial Intelligence*. IOS Press, pages 303-308.

MANNING, C. D. AND SCHÜTZE, H. (Eds.) (1999) *Foundations of Statistical Natural Language Processing*, MIT press, pages 294-300.

MANOJ, K., SENTHAMARAI KANNAN, K. (2013) "Comparison of methods for detecting outliers". *International Journal of Scientific & Engineering Research*, Vol. 4, (9),

MCCALLUM, A., NIGAM, K. AND UNGAR L. H. (2013). "Efficient clustering of high-dimensional data sets with application to reference matching". In *Proceedings of the sixth ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, New York, NY, USA, pages 169-178.

MORENO SANDOVAL, A. (1998) *Lingüística computacional*. Colección "Textos de Apoyo", Editorial Síntesis.

MORRIS, J. AND HIRST, G. (1998). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational linguistics, 17(1), pages 21-48*.

OLMOS, R., JORGE-BOTANA, G., LEÓN, J.A, AND ESCUDERO, I. (2013). "Giving an interpretation for the semantic dimensions in Latent Semantic Analysis". In *Proceedings of the twenty-third Annual Meeting of the Society for Text and Discourse* Valencia, pages 16-18.

OLMOS, R., JORGE-BOTANA, G., LEÓN, J.A, AND ESCUDERO, I. (2014) "Transforming Selected Concepts Into Dimensions in Latent Semantic Analysis". *Discourse Processes,* 51(5-6), pages 494-510.

PEARCE, D. (2002). "A comparative evaluation of collocation extraction techniques". In *Proceedings of The Third International Conference on Language Resources and Evaluation*, Las Palmas de GC, Canary Islands, pages 1530-1536.

PENADÉS MARTÍNEZ, I. (2001), "¿Colocaciones o locuciones verbales?", en *Lingüística*

Española Actual XXIII/1, págs. 57-88.

QUASTHOFF, U. AND WOLFF, C. (2002) "The Poisson Collocation Measure and its Applications". In *Second International Workshop on Computational Approaches to Collocations*, 22/23 7 2002, Wien.

R version 2.15.0 (2012) Copyright © 2012. The R foundation for Statistical Computing.

REAL ACADEMIA ESPAÑOLA, (1992) Diccionario de la Lengua Española, Vigésima primera edición, Espasa Calpe.

RESNIK, P. (1993), "Selection and information a class-base approach to lexical relationships", *University of Pennsylvania Institute for Research in Cognitive Science Technical Report No. IRCS-93-42*.

RESNIK, P. (1999), "Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language". In *Journal of Artificial Intelligence Research* 11, pages 95-130.

RODRÍGUEZ HONTORIA, H. (2003), "Similitud Semántica" en Martí Antonín, M.A., Fernández Montraveta, A., Vázquez García, G.(Editoras) *Lexicografía Computacional y Semántica*. Colección UB 64, Ediciones de la Universidad de Barcelona.

SECO, M., ANDRÉS, O., RAMOS, G., (1999). *Diccionario del Español Actual*, Aguilar.

SERRA SEPÚLVEDA, S., (2009). "Las restricciones de selección en los diccionarios generales de lengua española". *Boletín de Filología de la Universidad de Chile*, tomo XLIV/ 2, págs.. 187-213.

SCHÜTZE, H. (1993). "Part-of-speech induction from scratch". In *Proceedings of the 31st annual meeting on Association for Computational Linguistics (ACL '93)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 251-258.

PHAM-GIA T. AND HUNG T.L., (2001) "The mean and median absolute deviations", *Mathematical and Computer Modelling*, 34(7), pages 921-936.

REAL ACADEMIA ESPAÑOLA: Banco de datos (CREA) [en línea]. *Corpus de referencia del español actual.* <http://www.rae.es> [7/06/2011]

VINCZE O., ALONSO RAMOS M. (2013). "Incorporating Frequency Information in a Collocation Dictionary: Establishing a Methodology". *Procedia-Social and Behavioral Sciences*, Vol. 95, pages 241-248.

VAREAL, F., KUBARTH, H. (1994), *Diccionario Fraseológico del Español Moderno*. Gredos, Madrid.

ZULUAGA, A. (2002) "Los «enlaces frecuentes » de María Moliner. Observaciones sobre las llamadas colocaciones". *Lingüística Española Actual* XXIV/1, págs. 97-114.

Anexo

Forma Canónica	**Cabecera**	***A(*palpitar*, G)***
agallas, arrestos, ***corazón****, hígados, pecho, redaños*	*valentía*	*0,014154*
aleteo, ***contracción****, diástole,* ***golpe****,* ***latido****, palpitación, pulsación,* ***pulso, salto****, sístole*	*agitación*	*0,01233356*
benevolencia*,* ***corazón, entraña****, filantropía, humanidad, ternura*	*benevolencia*	*0,01185943*
anular, ***auricular****,* ***corazón****, cordial, dátil,* ***dedo, índice, medio****, meñique,* ***pulgar***	*dedo*	*0,01134058*
albedrío, arbitrio, ***corazón****,* ***elección, facultad****,* ***intención, libertad,*** *libre albedrío,* ***merced****, opción,* ***potestad, sentido, voluntad***	*voluntad*	*0,01117997*
afección, ***corazón****, entretelas, podredumbre, psicología,* ***sentimiento, sentir***	*sentimiento*	*0,01093138*
corazón, entraña*, ombligo, riñón,* ***vientre, yema***	*centro*	*0,01039362*
alborozo, ***alegría****, alegrón, buen humor,* ***dicha****, euforia,* ***felicidad****, flash, gozo,* ***humor****, júbilo,* ***placer****, regocijo*	*alegría*	*0,01022647*
canción*,* ***canto****, leitmotiv,* ***matarile, motivo****, ritornelo, sintonía, solfa, unísono*	*canción*	*0,00904735*
buche, cogollo, coleto, ***cuerpo, entraña, medula****, médula, meollo, molla, núcleo,* ***pecho****, tuétano*	*esencia*	*0,00901596*
ánimo*, designio,* ***disposición, idea, intención, intento****, lineamiento,* ***mente, plan****, predestinación, predisposición,* ***propósito, proyecto****, talante,* ***vena, vista***	*intención*	*0,00854951*
entrecejo, frente, sien*, sobreceja, testera, testuz*	*cabeza*	*0,00843233*
batolito, brecha, conglomerado, crestón, eluvión, estrato, filón, saca, salbanda, ***vena****, veta*	*suelo*	*0,00811938*
arranque, arrechucho, basca, capricho, corazonada, ***ida, impulso****, llamada, llamarada,* ***pronto, rapto, vena,*** *venida, ventolera, viaraza*	*determinismo*	*0,00806875*
caída, cuesta, declive, dirección, grada, gradiente, medianil, pecho, pechuga, pendiente, rampa, ranfla, rasante, repecho, reventón, ribazo, salmonera, subida, talud, varga, vertiente	*inclinación*	*0,00766294*
banda*, cadeneta,* ***cinta****, faja, franja, fres, friso,* ***goma, lista, pieza****, serpentina,* ***tira, vena****, veta*	*anchura*	*0,00760636*

altura, apotema, base, calibre, cateto, cosecante, coseno, cuerda, diagonal, diámetro, eje, flecha, hipotenusa, lado, luz, mediana, mediatriz, módulo, radio, sagita, secante, seno	*línea*	*0,00757381*
latido, tictac	*sonido*	*0,00753423*
agrado, amenidad, delectación, deleite, delicia, dulzura, fruición, gloria, goce, gozo, gusto, placer, regalo, sabrosura, satisfacción	*placer*	*0,00704114*
arteria**, capilar, hacecillo, **vaso**, **vena	*vena*	*0,00659664*

Tabla 6-1 Los 25 grupos con mayor Asociación de Selección para el verbo palpitar.

Forma Canónica	**Cabecera**	***A(brusco, G)***
arribismo, electoralismo, maquiavelismo, maroma, oportunismo, secesión, secesionismo, transfuguismo, viraje, virazón	*deshonestidad*	0,0438858
ahogadilla, chapuzón, inmersión, sumersión, viraje, zambullida	*dentro*	0,0380421
bolina, bordada, ceñida, guiñada, huida, revuelta, viraje	*desviación*	0,0339251
insolación, polavisión, revelado, veladura, viraje	*cinematografía*	0,0316622
deflación, depreciación, devaluación	*rebaja*	0,0165424
ademanes, cortesía, formas, maneras, modales, modos, perfiles	*cortesía*	0,0149323
abducción, acarreo, porte, transbordo, transferencia, transfusión, transición, tránsito, transporte, transposición, traspaso, trasplante	*transporte*	0,0139917
arreada, atraco, golpe, mangoneo, sirla, tirón	*robo*	*0,0127632*
conversión, evolución, mutación, paso, salto, transformación, transición, tránsito	*cambio*	*0,0113854*
caracol, corvadura, curva, dobladura, entrada, meandro, mocheta, recodo, recoveco, rincón, sucucho	*ángulo*	*0,0112814*
concusión, sacudida, vaivén	*agitación*	*0,0111638*
ademán, aspaviento, batimán, coco, coquito, deíxis, esguince, gesto, guiño, guiñada, jeribeque, llamada, mohín, morra, mueca, muestra, musaraña, rictus	*gesto*	*0,0111396*
acometida, agresión, apretón, arrancada, asalto, ataque, atentado, carga, contraataque, embate, irrupción, viaje	*ataque*	*0,0104063*
agarrón, jalón, orejón, repelón, tirón	*atracción*	*0,0100942*
acometida, bisagra, costura, empalme, engranaje, enlace, entronque, escarpe, inglete, juntura, traba	*unión*	*0,0093489*
alteración, cambio, evolución, innovación, modificación, muda, mudanza, reformismo, revés, sensibilidad, variación, variante, variedad, vuelta	*cambio*	*0,0093489*
circulación, desplazamiento, jubileo, locomoción, movimiento, paso, telequinesia, tráfico, tránsito, traslación, traslado, trayecto	*tránsito*	*0,0089822*
alboroto, alzamiento, asonada, cataclismo, comunidades, conmoción, contrarrevolución, convulsión, cuartelazo, golpe de estado, inquietud, insurrección, levantamiento, militarada, motín, movimiento, pronunciamiento, rebelión, revolución, revuelta, secesión, sedición, terrorismo, traición, tumulto	*desobediencia*	*0,0087821*
deducción, ilación, inducción, transición	*razonamiento*	*0,0087006*

acción, anagnórisis, anticlímax, argumento, clímax, desenlace, enredo, episodio, golpe, intriga, jornada, máquina, movimiento, nudo, peripecia, perístasis, prótasis, solución, trama	*mensaje*	*0,0086801*
balotada, batido, batuda, botepronto, brinco, cabriola, corcovo, jumping, pedico, rebote, saltadero, salto, salto trenzado	*salto*	*0,0082866*
aerobic, atletismo, calistenia, cros, culturismo, decatlón, footing, gimnasia, gimnasia rítmica, halterofilia, joggin, lanzamiento, olimpismo, pedestrismo, pentatlón, salto, tiro, yoga	*deporte*	*0,008161*
evolución, marcha, meneo, moción, motilidad, móvil, movilidad, movilización, movimiento, transmisión	*movimiento*	*0,0071675*
atolondramiento, aturdimiento, azoro, confusión, conmoción, corte, movimiento, tupición, turbación	*intranquilidad*	*0,0069941*
alocución, arenga, declamación, manifestación, mitin, movimiento, proclama	*difusión*	*0,0066185*

Tabla 6-2 Grupos de sustantivos con los valores máximos de Asociación de Selección para el adjetivo brusco.

Forma Canónica	**Cabecera**	***A(doloroso, G)***
aguijonada, alfilerazo, clavadura, cornada, espoleadura, estacada, filazo, mojada, picadura, pinchazo, piquete, puntada, puntazo, puntura, punzada, repique	*herida*	*0,0195348*
ay, cojijo, gemido, guaya, lamentación, lamento, lástima, queja, quejido, querella, suspiro	*lamento*	*0,0182051*
acíbar, aflicción, agonía, agrá, agraz, ahogo, amargura, angustia, apuro, congoja, desconsuelo, dolor, duelo, hiel, pasión, pena, pesar, presura, quebranto, sentimiento, sufrimiento	*sufrimiento*	*0,0161039*
acritud, analgesia, clavo, dolor, puntada, punzada, ramalazo, rayo, sinalgia, tormento, yaya	*dolor*	*0,0131621*
arestín, desazón, desplacer, disgusto, martelo, patatús, pensión, pesadumbre, punzada, puñalada, sinsabor, tribulación, varapalo	*sufrimiento*	*0,0130223*
aleteo, contracción, diástole, golpe, latido, palpitación, pulsación, pulso, salto, sístole	*agitación*	*0,0092431*
aféresis, apócope, contracción, epéntesis, metaplasmo, metátesis, paragoge, prótesis, síncopa	*representación*	*0,0092057*
afán, agonía, anhelo, ansia, apetito, concupiscencia, deseo, desiderátum, empeño, emulación, envidia, gana, golondro, hambre, hipo, pelota, pelusa, pío, pique, prurito, pujo, rabanillo, reconcomio, sed, sueño dorado, voto, yesca	*deseo*	*0,0090338*
cortadura, corte, fraile, gargantil, incisión, marceo, mediacaña, picadura, rafa, raja, rumbo, sajadura, siete, tajo	*corte*	*0,0083892*
adormecimiento, anquilosis, artritis, artritismo, artrosis, calambre, contractura, convulsión, espasmo, gota, miopatía, osteopatía, raquitismo, reuma, reúma, reumatismo, rigor, risa sardesca, risa sardonia, risa sardónica	*reumatismo*	*0,0080548*
cucaracha, macuba, manila, palillo, picadura, rapé, tabaco, tirulo, tripa	*veneno*	*0,00789266*
boquerón, boquete, brecha, calada, gotera, magaña, picadura, pinchazo, punto, rasa, silbato, tomate	*agujero*	*0,00773525*
astringencia, contracción, contractilidad	*disminución*	*0,00727213*
cariadura, caries, neguijón, picadura	*degeneración*	*0,00701377*
aclamación, alarido, baladro, berrido, bramido, chillido,	*voz*	*0,00684484*

clamor, exclamación, grito, ladrido, rugido, voz		
arañazo, carnicería, chispazo, colisión, corte, dehiscencia, fístula, grieta, herida, perforación, pupa, quemadura, rasponazo, rozadura, tocadura, transfixión, transverberación, yaya,	*herida*	*0,00677393*
admiración, asombro, consternación, escándalo, espanto, estupor, extrañamiento, extrañeza, flash, golpe, marasmo, novedad, pasmo, repullo, sorpresa, suspensión	*asombro*	*0,00663455*
arcada, contorsión, convulsión, espasmo, jeribeque, socollón, tic, torozón	*agitación*	*0,00641441*
agitación, agobio, alarma, ansia, ansiedad, batalla, combate, comecome, comezón, conflicto, desabrimiento, desasosiego, estrés, guerra, hormigueo, inquietud, intranquilidad, malestar, mar de fondo, mareta, pesadilla, polvareda, preocupación, rebato, repunta, revuelo, susidio, suspense, taco, tártago, torozón, zozobra	*intranquilidad*	*0,00627659*

Tabla 6-3 Grupos de sustantivos con los valores máximos de Asociación de Selección para el adjetivo doloroso.

Forma Canónica	Cabecera	*A(abundantemente, G)*
alargar, dar, dotar, proporcionar, proveer, servir, suministrar, surtir	*suministro*	*0,1102675*
brotar, destilar, efluir, emanar, exudar, fluir, instilar, manar, resurgir, revenir, rezumar, saltar, sudar, surgir, surtir, trasvinar	*fuera*	*0,0998385*
aparejar, apercibir, aprestar, aprevenir, aprontar, armar, aviar, avisar, conrear, disponer, estructurar, formatear, intentar, mullir, organizar, parar, perdigar, pertrechar, poner, preconcebir, preparar, prevenir, prevenir, proveer, reservar, sembrar	*preparación*	*0,0961376*
asolear, azacanar, azacanear, batallar, bregar, insudar, luchar, navegar, pegar, pernear, pringar, remar, sudar, trabajar, velar, zarandearse	*actividad*	*0,0500253*
llorar, resinar, sangrar, sudar	*savia*	*0,0466727*
agarrar, atrapar, cachar, cazar, empuñar, ligar, pescar, sacar, sudar	*éxito*	*0,0396564*
asperjar, espurrear, espurriar, irrigar, nebulizar, perlar, petrolear, pulverizar, regar, rociar, salpicar, vaporizar	*dispersión*	*0,0390331*
resudar, sudar, transpirar, trasudar	*sudor*	*0,0383897*
ensalivar, sudar, trasudar	*humedad*	*0,0383897*
enristrar, entramar, interceptar, regar, trabar, tramar, tranzar, trenzar	*intersección*	*0,0334439*

Tabla 6-4 Grupos con mayor asociación de Selección para el adverbio abundantemente.

Forma Canónica	Cabecera	*A(enérgicamente, G)*
acariciar, amasar, arrastrar, barrer, cardar, emborrizar, emprimar, estregar, fregar, fricar, friccionar, frisar, frotar, ludir, malaxar, masajear, mesar, peinar, rasar, rascar, raspar, rastrear, refregar, restregar, rozar, sobar, tazar, tropezarse	*tangencia*	*0,0388546*
aventurar, hacer muestras, motivar, parir, protestar	*comunicación*	*0,0356015*
chillar, encerrarse, manifestar, piar, protestar	*oposición*	*0,0342641*
acusarse, condenarse, confesar, protestar, reconciliarse, reconocerse	*culpabilidad*	*0,0275147*
agitar, cimbrar, menear, sacudir, traquear, traquetear, ventilar, vibrar, zamarrear, zangolotear, zarandar, zarandear	*agitación*	*0,026274*

ahuyentar, deponer, descartar, desechar, despedir, desterrar, negar, rechazar, recusar, reherir, repeler, repudiar, repugnar, resistir	*rechazo*	*0,0257924*
confutar, contradecir, desmentir, negar, rebatir, rechazar, rectificar, refutar, repeler, resistir, triturar	*refutación*	*0,0247705*
amordazar, atajar, atarse, atascar, contrariar, embarazar, embargar, embazar, empachar, empatar, emperezar, entorpecer, estorbar, estrangular, impedir, implicar, imposibilitar, incapacitar, obstruir, oponer, perturbar, reaccionar, trabar, tronchar, tropezar, vedar, yugular	*dificultad*	*0,0201045*
afrontar, arrostrar, contrarrestar, contrastar, desafiar, enfrentarse, rechazar, recibir, reparar, resistir, revolver	*ataque*	*0,018998*
acaudillar, acuadrillar, adiestrar, administrar, conducir, controlar, dirigir, encabezar, encaminar, gobernar, guiar, intervenir, maestrear, manejar, menear, normar, orquestar, patronear, presidir, regentar, regir, seguir	*control*	*0,0189946*

Tabla 6-5 Grupos con mayor Asociación de Selección para el adverbio enérgicamente.