Procesamiento del Lenguaje Natural, Revista nº 47 septiembre de 2011, pp 145-152 recibido 29-04-2011 aceptado 24-05-2011

Extracción automática de colocaciones terminológicas en un corpus extenso de lengua general

Automatic terminological collocations extraction from large corpus

Octavio Santana Suárez
Universidad de Las Palmas de Gran Canaria
Edificio Departamental de Informática y Matemáticas
Campus Universitario de Tafira
35017 Las Palmas de Gran Canaria
osantana@dis.ulpgc.es

Isabel Sánchez Berriel
Universidad de La Laguna
Edificio de Física y Matemáticas
Campus Universitario Anchieta
C/Astrofísico Francisco Sánchez s/n
38271 La Laguna
isanchez@ull.es

José Pérez Aguiar
Universidad de Las Palmas de Gran Canaria
Edificio Departamental de Informática y Matemáticas
Campus Universitario de Tafira
35017 Las Palmas de Gran Canaria
jperez@dis.ulpgc.es

Virginia Gutiérrez Rodríguez
Universidad de La Laguna
Edificio de Física y Matemáticas
Campus Universitario Anchieta
C/Astrofísico Francisco Sánchez s/n
38271 La Laguna
vgutier@ull.es

Resumen: Los sistemas automáticos de extracción de términos constituyen una herramienta fundamental cuando se afronta la tarea de compilación del léxico restringido a un campo de especialidad. Los análisis textuales llevados a cabo por este tipo de software deben incorporar estrategias que permitan detectar las colocaciones en la especialidad que se trabaje. En este trabajo se estudia la viabilidad del uso de corpus textuales extensos, sin información lingüística, como sucede con los que se pueden compilar a través de Internet, como fuente de información para la recopilación de colocaciones terminológicas. Con este propósito se analiza el comportamiento de distintos indicadores basados en las frecuencias registradas para una colección de términos económicos en un corpus del español de 300 000 000 palabras.

Palabras clave: extracción automática de colocaciones, terminología, lingüística computacional, minería de textos.

Abstract: The automatic systems which deal with term's extractions constitute an important tool when they make reference to the labor of compilation of lexemes, which is restricted to a specific field or specialty. The textual analysis that are realized for this type of software must include strategies that could detect collocations in the field in which is done. In this topic is studied the viability of the use from extensive textual's corpus, that have not contain linguistic information, as happen with those textual's corpus that could be compiled from internet. The internet is used like a source of information for the recompilation of terminology's collocations. With that purpose is analyzed the behavior of different indicators based on the frequencies registered for a collection of economic terms in a Spanish corpus of 300.000 words.

Keywords: automatic extraction of collocations, terminology, computational linguistics, text mining

ISSN 1135-5948

1 Extracción automática de términos y colocaciones

La compilación de bases de datos terminológicas se fundamenta en la extracción automática o asistida de términos a partir de colecciones de textos en el dominio tratado. Los términos se refieren a elementos del léxico de una lengua de especialidad que representan conceptos propios de su ámbito, hay que destacar que pueden estar formados por una o varias palabras. La extracción automática de los mismos a partir de corpus textuales especializados constituye una herramienta fundamental en el análisis del vocabulario correspondiente a un campo específico del conocimiento.

Las técnicas que se emplean en la extracción de candidatos a términos se fundamentan en información lingüística, o en valores de estadísticos basados en las frecuencias con que son utilizados en corpus textuales específicos del campo de especialidad, o bien combinan ambos métodos en los denominados sistemas híbridos.

1.1 Colocaciones

Las unidades terminológicas complejas de dos elementos son denominadas en este trabajo colocaciones terminológicas, y serán consideradas un caso particular de colocaciones léxicas de la lengua general. Una colocación constituye una combinación de palabras usadas con relativa frecuencia en una determinada lengua que, sin llegar a ser combinaciones libres, tampoco constituyen locuciones o expresiones idiomáticas. Los elementos que forman una colocación reciben el nombre de **colocados**, uno de ellos goza de autonomía semántica, llamado **base**, y selecciona una acepción especial en el **colocativo**, del que se reconoce su significado gracias a la base con la que aparece. Este fenómeno lingüístico se caracteriza entre otros aspectos por el de la precisión semántica que aportan (Koike, 2001), constituyendo un elemento de comunicación que permite denotar conceptos inconfundibles de la lengua.

1.2 La web como corpus textual

Una metodología habitual para la extracción automática de términos consiste en la explotación de corpus textuales de la lengua de especialidad. De la misma forma se explotan corpus de lengua general para la obtención de colocaciones. Las frecuencias relativas en las que se basan los cálculos estadísticos quedan determinadas por el contexto, de forma que el uso en la lengua general difumina la concentración de muestras que se espera de ellas en los textos de especialidad. Por esta razón, los valores que recogen la condición de uso preferente de una colocación frente a una combinación libre podrían alterarse. Sin embargo,

> *Determinadas colocaciones aparecen casi exclusivamente en un determinado registro*
>
> *Registro informático: implementar un programa, crear un directorio, etc.*
>
> *Registro científico: espacio sideral, estructura molecular, cólico nefrítico, etc.*
>
> (Corpas, 2001)

en clara alusión a lo que en este trabajo se identifica con las colocaciones terminológicas.

Es por ello que se espera que si se encuentran muestras de las mismas en un corpus de la lengua general, la información que se recopile estará fuertemente marcada por su condición de colocaciones terminológicas. Esta característica hace pensar que si éste es lo suficientemente extenso, aunque no esté conformado exclusivamente por textos especializados, aportará indicios respecto a su comportamiento como colocaciones terminológicas.

El uso de Internet como fuente de información respecto al uso de la lengua proporciona un corpus textual accesible universalmente, lo que la convierte en un medio para vencer la dificultad de la compilación de corpus específicos. Sin embargo, se debe garantizar que la extensión sea tal que pueda ser considerado como una muestra representativa de la lengua. Por otra parte, los avances en la potencia de procesamiento de equipos al alcance de un usuario medio hace que se pueda tratar una gran cantidad de textos de forma relativamente sencilla. Este binomio se traduce en que sea factible recopilar grandes corpus textuales desde la web para la caracterización del léxico. La propuesta que aquí se recoge analiza el comportamiento de indicadores de colocabilidad para términos económicos cuando son evaluados sobre un corpus extenso representativo de la lengua general.

2 Técnicas cuantitativas para la detección de colocaciones

Aunque teóricamente se podría escoger cualquier combinación posible, en muchas ocasiones algunas no son usuales o de uso preferente, el objetivo es detectar las que han sido sancionadas por la comunidad.

Así, en lanzarse al ataque, lanzarse ha sido preferido a arrojarse, y ataque ha sido preferido a lucha, o en guerra mundial, mundial ha sido preferido a internacional (Zuluaga, 2 002).

Esta propiedad justifica el uso de técnicas estadísticas para la explotación de textos en busca de combinaciones de palabras cuya frecuencia de aparición puede considerarse que no se debe al azar. Se determinan medidas de lo estrechamente relacionadas que estén 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 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. De esta forma las combinaciones libres deben cumplir que las variables aleatorias que representan la aparición en el corpus de los elementos que la constituyen sean independientes estadísticamente hablando. En otras palabras, la probabilidad de ocurrencia conjunta, *P(x,y)*, y las probabilidades individuales de la base y de los colocativos, *P(x)* y *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 el **z-score**, el **t-score** y la **fórmula de Dunning**, pasando por la **información mutua** –todos ellos han sido ensayados en este trabajo.

2.1 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.

$$frecRelat(x \text{ con } y) = \frac{frec(x, y)}{frec(x)}$$

2.2 Información Mutua

La medida de **información mutua** puntúa la coocurrencia de dos palabras mediante la expresión:

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

Tal valor mide la cantidad de información (en bits) que la coocurrencia aporta sobre la información de las apariciones individuales de las dos palabras. Las probabilidades se calculan directamente usando frecuencias relativas. Esta medida es inestable para cantidades pequeñas –se calcula sólo si la frecuencia absoluta *f(x,y) > 5*.

2.3 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. Se basa en la aproximación normal de la variable aleatoria binomial que considera como éxito el que la palabra *x* ocurra con *y*:

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

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

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

Siendo *p(y)* la probabilidad de que ocurra *y*, en una posición distinta en el corpus, *f(x)* la frecuencia de *x*, *N* 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.

2.4 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 t-score y no de un test de hipótesis (Evert, 2004).

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

2.5 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:

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

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

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

(Manning, Schütze, 1999)

3 La Base de Datos de combinaciones

Se dispone de una Base de Datos (B. DD.) en la que se recogen combinaciones de formas canónicas, frecuencia de uso de la combinación y frecuencia de uso de la forma evaluada sobre 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. En definitiva, una amplia muestra del español con un número total de palabras que está en torno a los 300000000. Los textos no incorporan ninguna información lingüística, es decir, se trata de ficheros de texto plano.

Los datos se recopilaron a partir de la ejecución de un programa implementado con este propósito, que simula la obtención de líneas de concordancias para cada una de las palabras en el corpus. El ámbito de una línea de concordancia se restringe a una frase y los valores de las frecuencias se recopilan para distancias entre palabras que van desde 1 hasta 10, tanto a la derecha como a la izquierda. La ejecución del programa sobre el corpus de trabajo produjo 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 han despreciado debido a la explosión de combinaciones irrelevantes que producen.

Se incorporaron ciertas restricciones determinadas por algunas características fundamentales de las colocaciones 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 descarta del proceso 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. (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 entran en el recuento sólo aquellas combinaciones que se correspondan con alguno de estos patrones, en orden indistinto. En concreto, este estudio explota las correspondientes a los que pertenecen con mayor frecuencia los términos en español:

a) Sustantivo – Adjetivo

b) *Sustantivo – Sustantivo* (*que puede usarse como adjetivo en el corpus*)

4 Del contraste de las colocaciones léxicas frente a las terminológicas.

La evaluación de los resultados se ha llevado a cabo considerando distintos conjuntos de ensayo que permiten contrastar el comportamiento de las medidas de asociación léxica en colocaciones terminológicas frente a colocaciones de la lengua general. En el grupo "*Colocaciones Recopiladas*", se incluyen los ejemplos aportados por lingüistas en trabajos sobre colocaciones léxicas del español consultados. En el de "*Colocaciones Terminológicas*", todos los elementos del corpus de términos económicos formados por dos elementos, y en el "*Corpus*", el total de combinaciones del tipo *Sustantivo – Adjetivo* o *Sustantivo – Sustantivo* (que puede usarse como adjetivo en el corpus).

El corpus de términos económicos analizado contiene 3630 palabras, todas ellas registradas en la base de datos de combinaciones. Se distribuye en 2115 términos simples o formando parte de 2699 complejos, de los cuales 2570 son colocaciones, constituyendo el objeto del estudio realizado. Estos datos reflejan la importancia de este tipo de unidades fraseológicas en la lengua de especialidad (Figura 1).

El estudio de los indicadores realizado sobre el fenómeno de las colocaciones léxicas condujo a la definición de una serie de restricciones que se deben imponer a los indicadores evaluados. Éstas garantizan una amplia cobertura y excluyen la gran cantidad de registros irrelevantes originados por la metodología, que si bien incorpora cierta información lingüística aportada por el lematizador, no utiliza técnicas de desambiguación.

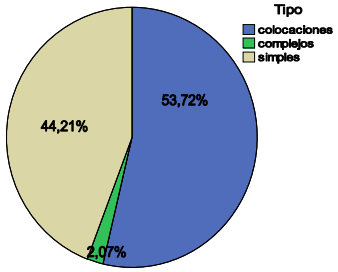

Figura 1 Distribución de los términos según la cantidad de elementos que lo forman.

4.1 Resultados en los distintos conjuntos de ensayo

Las combinaciones registradas en la B. DD. se ordenan según las puntuaciones que hayan obtenido con el propósito de discriminar las combinaciones libres de las colocaciones. El análisis comparativo de los valores alcanzados en los distintos grupos de ensayo llevó a establecer algunas restricciones que deben cumplir los elementos que se incorporen en tales rankings. El propósito de tales restricciones es excluir de los listados la gran cantidad de ruido que se entremezcla en los mismos. También se trasluce del análisis la necesidad de establecer los conjuntos de colocaciones para un término fijado, dando fiabilidad a la ordenación. En caso contrario se incurre en catalogar las combinaciones utilizando una metodología que trata todas las palabras por igual, independientemente de cuál es su frecuencia de uso en el corpus. 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).

4.1.1 Frecuencia relativa

Se observó una gran cantidad de combinaciones libres que quedan en las mejores posiciones del ranking, puesto que alcanzan la máxima frecuencia relativa, es decir, 1. Esta situación se da en los casos con frecuencia de la combinación pequeña, lo que llevó a considerar relevante la información obtenida cuando se dispone de al menos 10 muestras de la misma. En todos los grupos de ensayo se aprecia que los máximos valores cuando se evalúan respecto al adjetivo son mayores que cuando se hace respecto al sustantivo. Las restricciones que se impondrán a este indicador se traducen en exigir un número mínimo de muestras y que el máximo de las dos frecuencias relativas asociadas a una combinación supere un umbral:

$$Frec.\ (x,y) >= 10$$
$$Y$$
$$max(Frec.\ Relat._x Frec.Relat._y) >= 0.0005$$

Utilizando este criterio se reconocen las 727 colocaciones terminológicas que se encuentran en el corpus con el número de muestras exigido. Si se impone un valor de corte para la frecuencia relativa de 0,005 este número bajaría a 503.

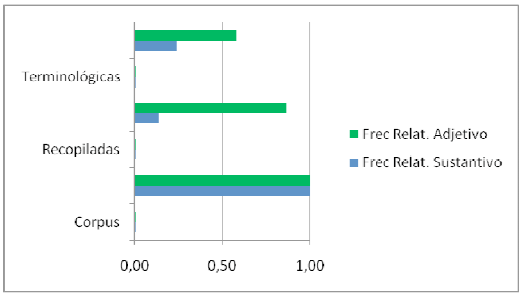

Figura 2: Frecuencias relativas en los distintos grupos de ensayo.

Se muestran como ejemplos las combinaciones mejor posicionadas cuando el ranking se establece según los valores de las frecuencias relativas para los siguientes términos: *crédito, auditoría y comisión (1)(2)(3)*. En la Tabla 1 se aporta un resumen respecto a la cobertura que se alcanza.

(1) crédito: hipotecario, bancario, cooperativo, suplementario, inmobiliario, granjeado, territorial, agrícola,…
(2) auditoría: general, ambiental, sometido, anual, fiscal, externo, independiente, especializado,…
(3) comisión: dictaminador, paritario, trilateral, gestor, sectorial, bicameral, codificador, interamericano, mixto, permanente, plenario, demarcador, interestatal, organizador, integrado, inculpado, presidido,…

	crédito	auditoría	comisión
Colocaciones Terminológicas	10	8	6
Corpus general	7	2	4
Combinaciones	943	8	1066
Frec. Total>=10 y Frec. Relat.>0,0005	28	8	124

Tabla 1: Resumen de los resultados para: crédito, auditoría y comisión.

4.1.2 Información Mutua

Los valores para las colocaciones en los conjuntos de ensayo oscilan entre 0 y 15,5 tanto referidos a la lengua general como a las terminológicas, si bien en el corpus el rango de este indicador es [-3,4 25,8]. Cabe destacar que los valores más altos se registran en palabras en el mismo campo semántico, pero no necesariamente son colocaciones (1). Esto provoca que se introduzcan en los puestos más altos del ranking combinaciones libres que deben ser evitadas.

(1) adarguero algueador, teobromina acético, metacarpo falángico, sortero suertero, aldehído fórmico, dióptrico catóptrico, denuncia reconciliación, largo puntiagudo, santería vudú, deshonrabuenos bellaco, efímera marchitable, gamba paté,…

Por otra parte, también refleja mayor fiabilidad cuando se dispone de un número mínimo de muestras –se fija en 10.
Las restricciones impuestas en este caso se traducen en:

Inf. Muta >0 and
Inf. Mutua <15 and
Frec. (x, y) >10

Aún así, quedan bien posicionadas en el ranking combinaciones pertenecientes al mimo campo semántico, independientemente que sean colocaciones o no (2), (3).

(2) garitero tahúr, breviario misal, bejín musgo,…
(3) acidez estomacal, machete platanero, lóbulo olfatorio, paralelogramo diagonal, isótopo radiactivo, efectismo folletinesco, cubertería inoxidable,…

4.1.3 Z-Score

Los valores alcanzados por las colocaciones recopiladas y terminológicas varían en el intervalo [-50 200] (Figura 3). En ambos conjuntos se observa que valores del z-score negativos y mínimos corresponden a colocaciones en las que el adjetivo tiene un valor funcional que actúa como intensificador de la base:

(4) señor problema, día oscuro, tiempo horrible, historia vivo, razón profundo,…
(5) bien libre, libro blanco, carta verde, valor seguro,…

frente a los máximos valores del z-score que revelan colocaciones léxicas:

(6) pecado venial, testigo ocular, prisión preventiva, malformación congénita, delito flagrante, vegetación exuberante,…
(7) ordenamiento jurídico, tarjeta postal, cabina telefónica, materia prima,…

En este caso, se propone desechar los valores negativos cuando se trata de detectar colocaciones terminológicas. Siguiendo con los mismos ejemplos, se listan en (8), (9), (10) las

mejores puntuadas en el ranking basado en este indicador:

(8) crédito: cooperativo, hipotecario, bancario, público, agrícola, territorial, presupuesto, digno, financiero, pago, suplementario, adicional, ejercido, nacional, inmobiliario,…

(9) auditoría: general, ambiental,…

(10) comisión: organizador, diputado, designado, parlamentario, municipal, interamericano, proyecto, bicameral, compuesto, senado, respectivo, desempeñado, delegado, federal, representante, editor, propuesto,…

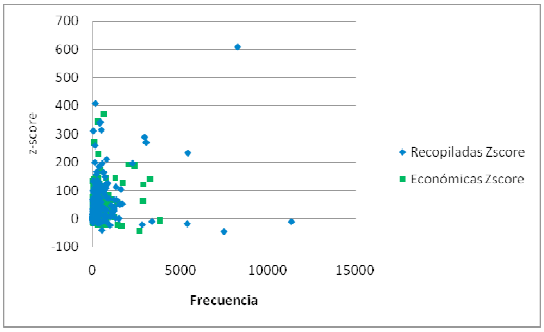

Figura 3 Z-Score en las colocaciones recopiladas y las terminológicas

4.1.4 T-Score

Se detectó una relación directa entre la puntuación y la cantidad de muestras encontradas de una combinación dada. En la Figura 4 se han representado las frecuencias frente al t-score para mostrar este comportamiento. Por esta razón se considera que no capta ningún fenómeno de asociación que permita establecer rankings globales con el fin de discriminar las colocaciones respecto a las combinaciones libres.

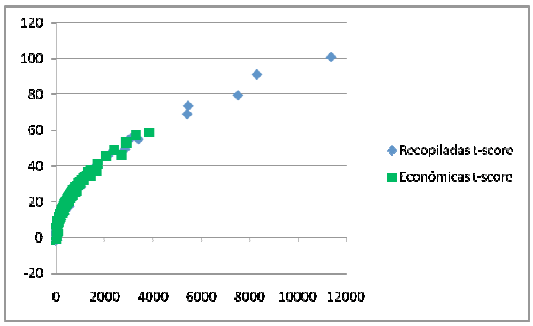

Figura 4: Relación entre el número de muestras y el t-score.

4.1.5 Test de Dunning:

Se produce una concentración de los valores para las colocaciones en el intervalo $[-20*10^3\ 20*10^3]$. Este test es más estable que los restantes, en la Figura 5 se observa que no está influenciado por la frecuencia de la combinación.

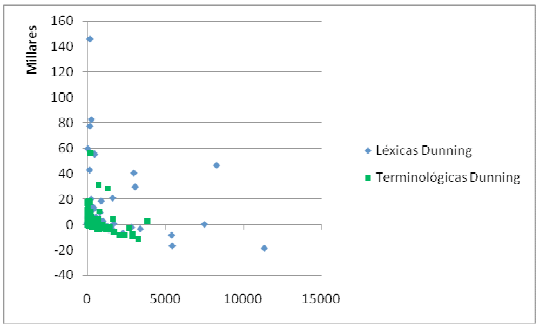

Figura 5 Test de Dunning en las colocaciones léxicas y terminológicas.

Como en el caso del z-score se apreció una polarización entre las colocaciones léxicas estrechas en los valores máximos y las colocaciones funcionales en los valores negativos mínimos. De igual forma, este fenómeno se produce tanto en las colocaciones recopiladas como en las terminológicas.

	Recopiladas	Terminológicas
>=0	poder adquisitivo, vida azaroso, trabajo ímprobo, pan ácimo, …	derechos político, administración público, materia prima, materia primo,
<0	dolor profundo, vida santo, silencio absoluto, color amarillo, casa grande,…	dinero contante, bienes mueble, bien mueble, fuerza productivo,…

Tabla 2 Colocaciones recopiladas y terminológicas con valores del test de Dunning extremo.

Utilizando el test de Dunning como valor para catalogar las combinaciones cuya base es crédito, auditoría y comisión se obtienen los colocativos:

(11) crédito: suplemental, intersindical, insoluto, incobrable, hipotecario, reemplazable, lionés, presupuestado, público, devengado, bancario, inmobiliario,

(12) auditoría: general,…

(13) comisión: suelto, dictaminador, paritario, trilateral, bicameral, gestor, sectorial, rogatorio, improbado, codificador, iliberal, lucífugo, secretarial, interinsular, interamericano, depurador, bipartito,…

5 Conclusiones

Se han contrastado los resultados obtenidos sobre una variedad de indicadores utilizados para la detección de colocaciones en un corpus textual extenso; se confrontan los resultados entre un conjunto de colocaciones léxicas y otro terminológicas. Para un término económico fijado, se comprueba que los rankings basados en frecuencias relativas, z-score y test de Dunning permiten ampliar el catálogo de partida con nuevas colocaciones, la mayoría de ellas también económicas. En estos tres casos, los resultados no presentan grandes diferencias, salvo que se pone de manifiesto la capacidad del z-score para separar colocaciones estrechas frente a colocaciones funcionales. Por último, la información mutua mezcla colocaciones con combinaciones de palabras en el mismo campo semántico y el t-score adolece de una gran dependencia de los valores registrados respecto a la cantidad de muestras de la combinación, por lo que no se aconseja su utilización para la construcción de tales rankings con este tipo de corpus.

Bibliografía

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

Bosque, I. 2001. "Sobre el concepto de colocación y sus límites", *Lingüística Española Actual* XXIII/1: 9-40.

Corpas Pastor, G. 1996. *Manual de Fraseología española*, Madrid, Gredos.

Corpas Pastor, G. 2001. "Apuntes para el estudio de la colocación". *Lingüística Española Actual* XXIII/1: 41-56.

Church, K. W.; Hanks, P. 1990. "Word association norms, mutual information, and lexicography". *Comput. Linguist.* 16, 1: 22-29.

Dunning, T. 1 993. "Accurate Methods for the Statistics of Surprise and Coincidence". *Comput. Linguist.* 19: 61-74.

Evert, S.; Krenn, B. 2001. "Methods for the Qualitative Evaluation of Lexical Association Measures". *Proceedings of the 39th Annual Meeting on Association For Computational Linguistics*: 188-195.

Koike, K. 2001. *Colocaciones léxicas en español*. Universidad de Alcalá, Takushoku University.

Manning, C.; Schütze, D. 1 999. "Foundations of Statistical Natural Language Processing". *MIT Press*: 141-177.

Pamies Bertrán, A; Pazos Bretaña, J. M. 2003. "Acceso automatizado a fraseologismos y colocaciones en corpus no etiquetado". *Language Desing:* 39-50.

Santana, O.; Pérez, J. y otros. 2007. "Development of Support Services for Linguistic Research over the Internet TIN2004-03988". *Jornadas de Seguimiento de Proyectos en Tecnologías Informáticas*: 167-174.

Santana, O.; Pérez, J. y otros. 1999. "FLANOM: Flexionador y lematizador automático de formas nominales". *Lingüística Española Actual* XXI, 2, Ed. Arco/Libros, S.L.: 253/297.

Santana, O.; Pérez, J. y otros. 1997. "FLAVER: Flexionador y lematizador automático de formas verbales". *Lingüística Española Actual* XIX, 2, Ed. Arco/Libros, S.L.: 229/282.

Seco, M.; Andrés, O.; Ramos, G. 1999. *Diccionario del Español Actual* (DEA), Aguilar.

Varela, F.; Kubarth, H. 1994. *Diccionario Fraseológico del Español Moderno* (DFEM) Gredos, Madrid.

Zuluaga, A. 2002. "Los «enlaces frecuentes » de María Moliner. Observaciones sobre las llamadas colocaciones". *Lingüística Española Actual* XXIV/1: 97-114.