Detección automatizada de colocaciones y otras unidades fraseológicas en un corpus electrónico*

José Manuel Pazos Bretaña**
Antonio Pamies Bertrán***
Universidad de Granada

1 Introducción

Desde un punto de vista formal las colocaciones no son identificables a priori en una búsqueda automática, de hecho, parecen incluso plantear problemas para la detección "manual", a juzgar por las unidades ausentes (o sobrantes) en los escasos repertorios especializados, tales como – p. ej. – el diccionario español *Redes* (Bosque 2004), dentro de una rama lexicográfica que no por ser todavía incipiente deja de ser absolutamente prioritaria.[1] Desde el punto de vista computacional, la definición formalista "clásica" de la colocación sería, en la práctica, aplicable a cualquier tipo de UF en una rutina de búsqueda, ya que dicha definición se basa exclusivamente en la frecuencia de coaparición de los componentes: *co-ocurrencia estadísticamente superior a lo esperable* (Halliday 1961: 276).

Sin embargo, la relación verificable entre frecuencia y coaparición no está exenta de problemas que los propios criterios de búsqueda ponen de manifiesto. El primero de ellos, tan evidente

* Este trabajo es parte de un proyecto colectivo, Estudio translingüístico y cognitivo de las estructuras analíticas léxico-sintácticas verbo+nombre o colocaciones verbo-nominales, programa español de I+D dirigido por J.d.D. Luque Durán y financiado por el Ministerio de Educación y Ciencia (BFF2003-0038).
** jmpazos@supercable.es
*** apamies@supercable.es
1 cf. Luque 1995, Haussmann 1979, 1981, 1984, 1985.

que ni se suele mencionar, es, por supuesto, que este criterio conlleva necesariamente que sólo las colocaciones que aparecen más de una vez en el corpus son detectables como tales. El segundo problema es que esta definición "puramente" estadística de las colocaciones no permite distinguirlas de otro tipo de unidades fraseológicas (locuciones idiomáticas, locuciones gramaticales, paremias...), distinción de subclases fraseológicas que difícilmente se puede automatizar desde tal enfoque.

Para averiguar, mediante un análisis cuantitativo, hasta qué punto la frecuencia de un determinado patrón de concurrencia léxica difiere de lo que podría esperarse estadísticamente, es preciso comparar dicha frecuencia con otra magnitud: la *frecuencia esperada* (es decir, la frecuencia de coaparición en ausencia de cualquier factor no aleatorio que tienda a asociar dos unidades). Para calcularla, el modelo distribucional más sencillo es el modelo *aleatorio* o *normal*. Éste presupone que – en ausencia de ninguna regla – la probabilidad de una determinada combinación en un determinado fragmento del texto debería ser la misma que en la totalidad del texto. Pruebas estadísticas tales como *z-score*: y *t-score* se basan directamente en estos datos. Esta ausencia de otros factores constituye un apriorismo no comprobable como tal, ya criticado por Dunning (1993) que considerara poco realista la distribución aleatoria, dadas las considerables variaciones comprobadas según el tipo y tamaño del corpus. Este autor propuso el uso de un coeficiente con mayor grado de independencia del tamaño del corpus y que no presuponga una distribución normal del léxico –considerando distribuciones binomiales y multinomiales. La información necesaria para el cálculo del coeficiente que propone (una variación del marcador *log-likelihood*) es un listado de todos los bigramas recurrentes, tomando como contexto relevante un intervalo de *x* palabras a derecha e izquierda del primer componente.

2 Experimentos

Aunque todo aparato estadístico necesita obviamente ser aplicado a un corpus de grandes dimensiones para conseguir el rendimiento esperado, es necesario experimentar primero en uno reducido para poder verificar "manualmente" los resultados y comprobar cuáles son los problemas y fallos inherentes a la propia metodología experimental. Por ello hemos realizado la mayor parte de nuestros experimentos con textos breves, que nos han permitido ir refinando la metodología experimental, antes de que sea aplicable con eficiencia a *corpora* de gran tamaño.

2.1 *Primer experimento*

Elegimos como campo de pruebas el *Quijote* de Miguel de Cervantes, cuyo tamaño es, por un lado, lo suficientemente grande para obtener resultados estadísticos relevantes, y, por otro, lo suficientemente reducido como para poder verificar inmediatamente sobre el texto la pertinencia de la información obtenida. El "ruido" amenaza con saturar la información recuperada a causa de elementos indeseables como la coaparición del tipo palabra plena + clítico cuya ocurrencia es muy alta (p.ej. *Art+N*). La manera más simple y "primitiva" de filtrar el corpus de este efecto parasitario, es aplicar un filtro léxico (o *stop word list*), que elimina los morfemas gramaticales gráficamente autónomos (determinantes, conjunciones, preposiciones, pronombres personales, auxiliares, etc.), con ello perdemos, naturalmente, muchas locuciones conjuntivas, prepositivas y adverbiales, así como ciertas oposiciones relevantes del tipo *dar de mano* ≠ *dar la mano*, limitación que, aun así, debería quedar ampliamente compensada por las ventajas de tamaña "limpieza". El listado obtenido (con el programa *TACT*) a partir del corpus "purgado" de *palabras herramienta* y de nombre propios de los protagonistas nombrados más veces, incluye aun así muchas combinaciones recurrentes que tampoco deberían aparecer en una búsqueda eficaz.

Así, abundaban las combinaciones textuales, que podemos llamar "aleatorias" (p. ej. *ama* & *sobrina,* que co-ocurren 24 veces), abundan incluso más que las UF, pues éstas sólo representan 723 bigramas sobre un total de 10.716 combinaciones detectadas, o sea un 93% del total de bigramas con recurrencia igual o superior a dos, NO son fraseológicos. Entiéndase unidades distintas (los *tokens* serían muchos más). En cuanto a la distribución del 7% restante (los únicos datos que resutklaron relevantes), nuestra hipótesis inicial era que, al ordenar de forma descendente el listado obtenido, según una de las magnitudes estadísticas, las UF deberían concentrarse en una zona en particular del mismo, distinguiéndose así de las combinaciones "aleatorias", por los mencionados criterios estadísticos.

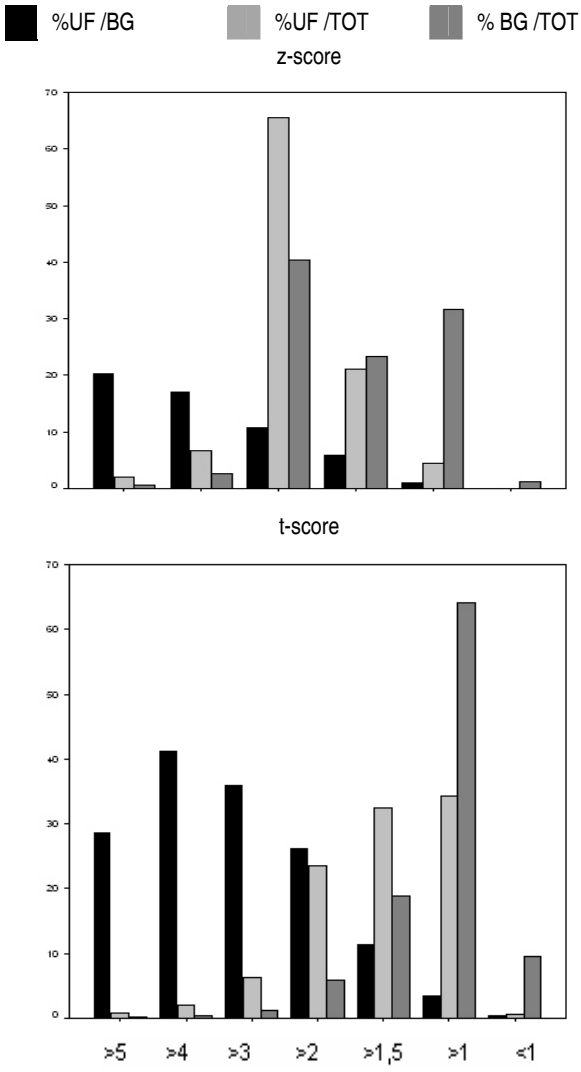
Mostramos a continuación (ver **figura 1**), para cada valor estadístico (tanto con *z-score*,como con *t-score* y *f-D*), el número de unidades fraseológicas (*UF*) comparado con el número total de bigramas recurrentes (*BG*), identificando tres valores relevantes:

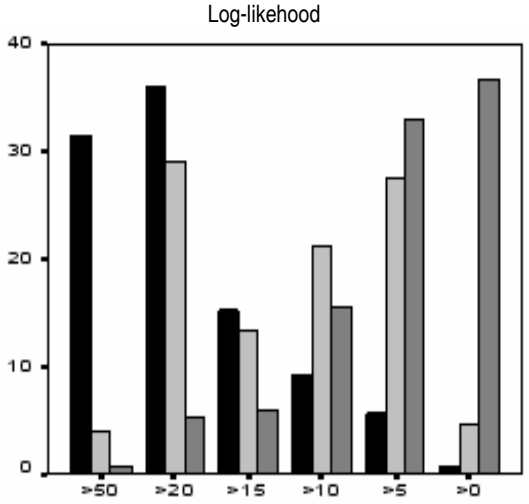
a) la *densidad fraseológica* (proporción entre el número de unidades fraseológicas y de bigramas recurrentes para una zona determinada de listado: *%UF/BG*),

b) el *volumen fraseológico* (proporción entre el número de UF de una zona del listado y el total de UF del corpus: *%UF /TOTAL*),

c) el *volumen de co-ocurrencia* (proporción entre el número de bigramas de la zona del listado y el total de bigramas recurrentes del corpus: *%BG /TOTAL*).

Figura 1
(gráficos para z-score, t-score y f-D [log-likelihood])





Cuantitativamente, los tres parámetros sólo validan parcialmente la hipótesis de partida: en todos ellos podemos ver que, grosso modo, a mayor valor de *z-score*, de *t-score* o de la fórmula de Dunning, mayor densidad fraseológica. Pero, al mismo tiempo, esta correlación tiene poca utilidad práctica, debido a que las zonas de alta *densidad fraseológica*, delimitadas gracias a los valores superiores de *z-score* (propuesto por Berry-Roghe en 1973) y *t-score* (propuesta en Church et al. en 1991), coinciden con zonas de bajo *volumen fraseológico.* La fórmula *t-score* resulta más eficaz que *z-score,* pero menos que el *log-likelihood* propuesto por Dunning en 1993 (*fD*), que es la que mejor consigue delimitar zonas con mayor densidad fraseológica. Aún así, dos tercios de las UF quedan en zonas de baja densidad, con lo cual tampoco podemos esperar de la mera aplicación de esta fórmula una eficaz decantación automática de las UF de un corpus.

Por otra parte, esta forma de recuperación no logra eliminar un tipo de combinaciones sintagmáticas que no son UF, y consiguen aun así los mismos valores estadísticos que las UF, sobre todo en las zonas de mayor volumen fraseológico, llegando incluso a superarlas en número**:** nombres propios y topónimos (*Alejandro+Magno*), combinaciones contextuales (*ingenioso+hidalgo; barbero+cura*), combinaciones conceptuales debidas a nexos lógicos u ontológicos entre dos referentes (*come+bebe; hambre+sed*), y reduplicaciones retóricas (*¡ladrones, ladrones!; de mesón en mesón y de venta en venta*), etc. Ni siquiera en la zona con mayor densidad fraseológica, dicha correlación permite discriminar por sí sola las UF de las combinaciones.

2.2 *Segundo experimento*

Antes de plantearnos mejorar los resultados debíamos modificar la metodología para que éstos sean comparables deun experimento a otro, y de un coprus a otro, por lo que cambiamos las subdivisiones del eje horizontal del gráfico (segmentación del listado de bigramas basada en valores absolutos de *fD* (>50, >20 etc.), por fragmentos del listado iguales entre sí (particiones del 10% del listado de bigramas).

Reanalizamos así el texto del *Quijote*, y el resultado confirma la correlación planteada por la hipótesis teórica de partida: a mayor valor del marcador estadístico, mayor densidad fraseológica (ver **figura 2**).

Para comparar estos resultados con los de otro texto, hemos empleado un corpus pequeño, en el que resulta menos costoso evaluar "manualmente" la exactitud del análisis automático, elegimos *La familia de Pascual Duarte* de Camilo José Cela, texto breve y proporcionalmente abundante en locuciones idiomáticas, proverbios y colocaciones. Una vez eliminados del texto los clíticos, auxiliares, etc., extraemos todos los bigramas recurrentes y los ordenamos decrecientemente según su valor de *fD* (ver **figura 3**).

Figura 2

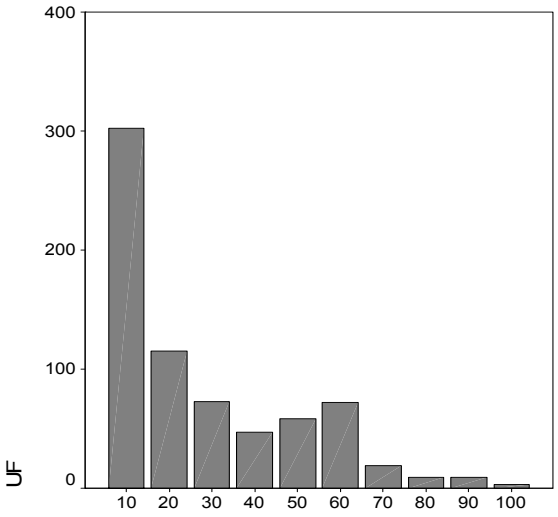
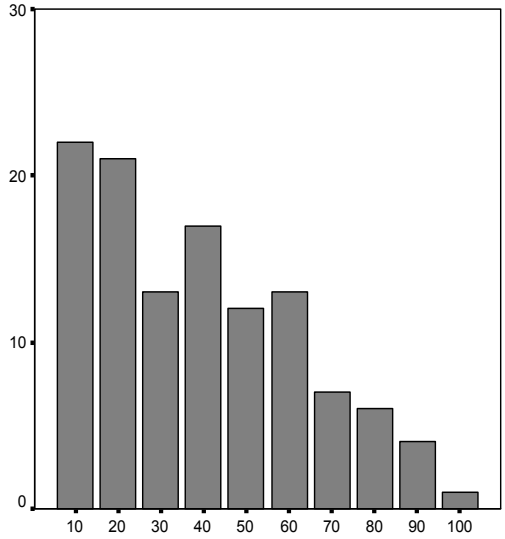


Figura 3



El corpus queda reducido a 14.261 palabras con un total de 556 bigramas recurrentes. La densidad fraseológica total del corpus es muy alta (a pesar de que únicamente son detectables las UF que aparecen más de una vez. La proporción de "ruido" sigue siendo bastante mayor de lo deseable: de los 556 bigramas recurrentes de *Pascual Duarte*, la mitad (282) son combinaciones no fraseológicas (ontológicas, contextuales o aleatorias (*media+docena, banco+iglesia, domingos+misa, litro+vino*).

Sospechamos que el resultado estadístico "puro" podía verse alterado por la lematización del corpus (si las variantes gramaticales se unifican reduciéndose a sus formas canónicas, se suman entre sí en el cómputo: las posibilidades de detección aumentan, al incorporarse variantes que, por aparecer una sola vez, quedaban antes excluidas antes). Aunque la lematización sea una operación lingüística, repercute sobre la estadística, por ello no supone un cambio radical de criterios. Pero la eficacia global no mejoró, pues aunque la lematización logró multiplicar por dos el número de UF detectadas, también favoreció las combinaciones irrelevantes: la densidad fraseológica es la misma que en el texto no lematizado: en el mejor de los casos [1er tramo] sigue siendo del 49%.

2.3 *Tercer experimento*

Un inconveniente de nuestra metodología es que excluye los pares de palabras que co-ocurren una sola vez. Podríamos esquivar parcialmente este obstáculo duplicando el texto, sistema muy tosco pero que, de forma rápida y sencilla, permite obtener datos estadísticos en los que participan todas las combinaciones.

En *La familia de Pascual Duarte*, se obtienen diez veces más datos con un corpus dos veces mayor, lo cual da una idea de la proporción de unidades "marginadas" por el umbral mínimo de recurrencia, aunque la rentabilidad no mejora (ni puede mejorar) la ratio entre información y ruido (puesto que ambos resultan estadísticamente favorecidos en la misma proporción).

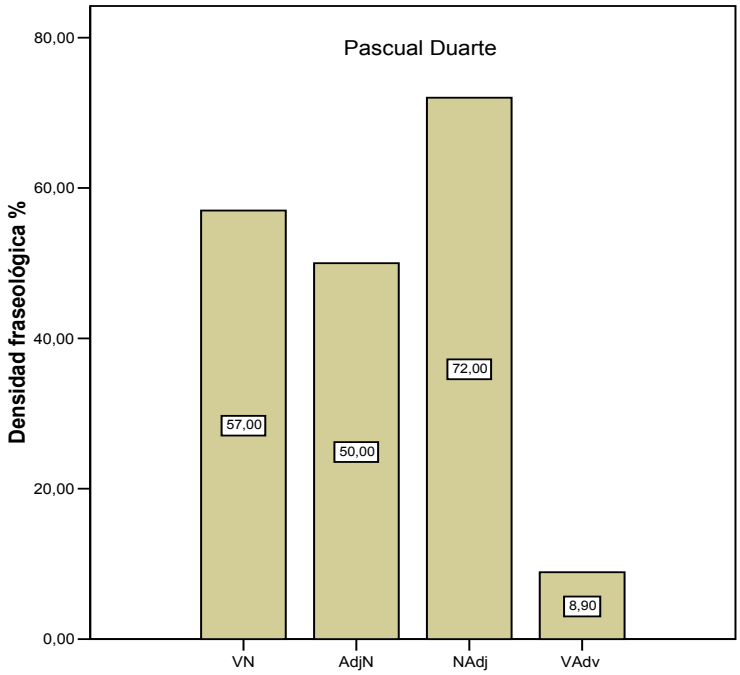
2.4 *Cuarto experimento*

El hecho de haber alcanzado el límite de efectividad de los métodos estadísticos "puros" nos lleva a plantearnos la necesidad de combinarlos con criterios lingüísticos, los mínimos, dado que la automatización requiere por definición limitar la intervención humana y el almacenamiento de conocimiento lingüístico previo. Otro mecanismo de gran aplicabilidad es la etiquetación (morfosintáctica) automática del corpus. Siguiendo el procedimiento empleado hasta ahora, aplicamos primero las nuevas formas de búsqueda al texto de *La familia de Pascal Duarte* y luego al *Quijote*, de modo que el resultado sea comparable a los anteriores,

En primer lugar hemos lematizado y etiquetado la totalidad del texto original (esta vez sin aplicar ningún filtro. El lematizador utilizado es la aplicación *Freeling 1.2* (Carreras, X. et al., 2004). El fichero de salida que se obtiene proporciona el lema de cada elemento y les asigna etiquetas, de esta manera las diferentes variantes de los lexemas no se computan como si fuesen elementos distintos, sino que se procesan como de un mismo elemento (p. ej. *pide perdón* y *pidió perdón* se computan estadísticamente como ocurrencias de *pedir+perdón*). El fichero obtenido se convierte en base de datos relacional en formato *Access 2003*, ordenable por cualquiera de sus campos (1er componente, 2ndo componente, frecuencia del 1er componente, frecuencia del 2ndo componente, frecuencia de coaparición, valor de *log-likelihood*...). Se crean varias bases de datos, según la categoría gramatical de los componentes de los bigramas (verbo + nombre, nombre + adjetivo, etc.).Por coherencia con nuestro planteamiento estadístico, y también para que los resultados sean comensurables con los de anteriores experimentos, el umbral de una co-ocurrencia sigue siendo igual o mayor que dos.

En el caso de *Pascual Duarte*, el número de unidades recuperadas sigue siendo reducido en cifras absolutas, dado el escaso tamaño del corpus, pero las proporciones entre el número de UF y de bigramas son interesantes, especialmente en algunas categorías gramaticales (ver **figura 4**). No fue necesario el análisis por tramos del listado dada su reducida extensión.

Figura 4

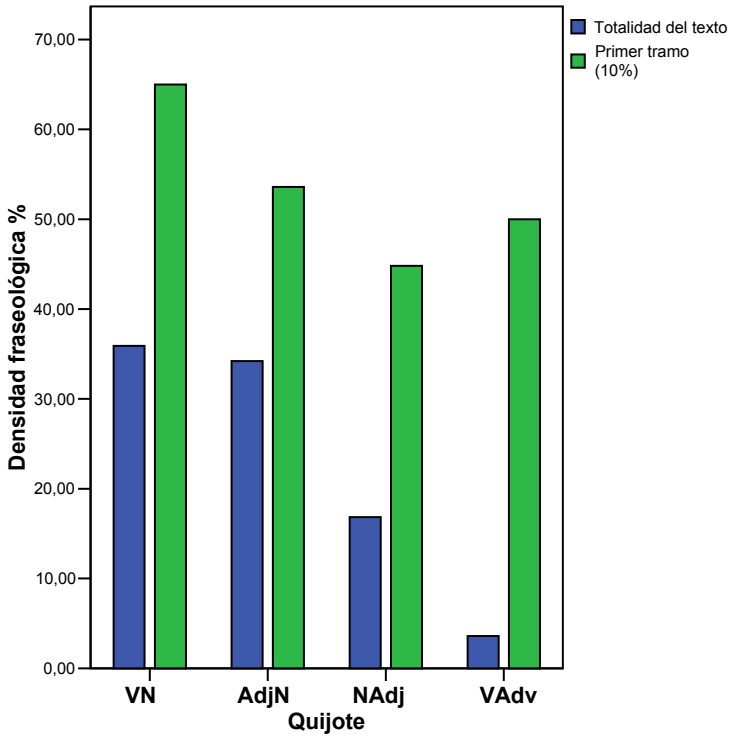


En la categoría Verbo + Nombre obtenemos 42 bigramas, de los cuales 24 son fraseológicos (57%), esencialmente colocaciones de verbo soporte y locuciones verbales. El número es pequeño pero la proporción es bastante alta, logra ser superior – para el texto entero – a la que habíamos obtenido en el mejor tramo con la estadística "pura" (49%). En la categoría Adjetivo + Nombre, obtenemos 32 bigramas de los que 16 son idiomáticos (50%), proporción no desdeñable pese al escaso número absoluto en que se basa. Para la categoría Nombre + Adjetivo, el resultado es de 11 bigramas co-ocurrentes dos o más veces, de los que 8 son de naturaleza fraseológica, número ínfimo pero proporción elevadísima (72%).

Para la categoría Verbo + Adverbio, obtenemos 79 bigramas de los que sólo 7 (8,9%) son unidades fraseológicas, esta vez el número es alto pero la densidad fraseológica es muy reducida.

Estas proporciones hacían sospechar que íbamos por buen camino, y que podíamos aplicar este sistema a un corpus mayor con fundadas esperanzas de mejorar sus resultados. Para comprobarlo, hemos aplicado exactamente el mismo método al texto del *Quijote*, que aunque también es un corpus pequeño, su tamaño – muy superior al de *Pascual Duarte* – sí permite observar ua curva descendente si se divide el output en tramos del 10% del total de bigramas recuperados, de forma que se pueda comparar el resultado en igualdad de condiciones con el obtenido con el mismo corpus no etiquetado (ver **figura 5**).

Figura 5



En la categoría Verbo + Nombre obtenemos 713 bigramas con recurrencia igual o superior a 2, de los que 256 son fraseológicos (35,9%). La densidad fraseológica del texto entero ya no es tan alta como en *Pascual Duarte*, pero vemos que., por ejemplo, para V+N, el primer tramo, es decir el que corresponde a los máximos valores de *fD*, contiene 72 bigramas de los que 47 son UF, es decir un 65%, densidad fraseológica que supera con creces el 28% global obtenido con la estadística "pura" en el mejor tramo del mismo texto. El mal resultado que nuevamente obtiene la categoría V+Adv pone de manifiesto que el marcador *fD* "privilegia" la recurrencia de una combinación pero "penaliza" la altísima frecuencia de sus dos componentes por separado (cosa que afecta con los adverbios).

En resumen, la detección con técnicas "mixtas" (lematización + etiquetacion + selección categorial + filtro estadístico mediante *Log-likelihood*) ha resultado superior a una aproximación exclusivamente estadística, aplicada a las mismas novelas, aunque resulta más eficaz para ciertas categorías formales (p.ej. V+N) que para otras (p.ej. V+Adv).

2.5 *Quinto experimento*

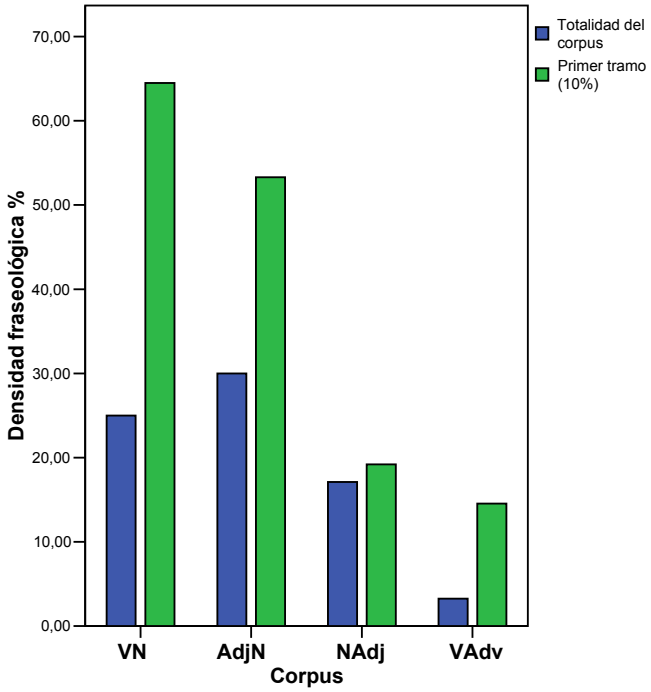
Llegado el momento de ampliar el corpus, hemos procurado que éste tenga una aceptable homogeneidad y representatividad, basándonos en un lugar dado (solamente España), una época dada (el siglo XX), una norma dada (la lengua escrita), y un género discursivo determinado (narrativa). Hemos buscado algo de variación que compense dicha homogeneidad fusionando en un único documento textos y autores distintos entre sí, dentro de los rasgos comunes señalados. El corpus, que agrupa 14 conocidas novelas españolas modernas, consta de **1.137.323** palabras, dimensiones razonablemente representativas para un proyecto como el que aquí se describe.

La búsqueda de co-apariciones con recurrencia superior a 2 en el texto (previamente lematizado) nos da **355.503** bigramas diferentes[2]. Nuevamente, la búsqueda en la categoría V+N resulta mucho más fructífera aún en un corpus grande: de 1375 bigramas, 347 son fraseológicos, o sea un **25%** (mayormente colocaciones). Lo más interesante es que el valor de *fD* logra, en el primer tramo del listado, una densidad fraseológica del **64,5%,** o sea, prácticamente dos tercios. La categoría Adj.+N obtiene 1060 bigramas, de los cuales 321 son fraseológicos (**30%**) y *fD* agrupa una densidad fraseológica bastante elevada en el tramo inicial (**53,3%**) .La categoría N+Adj. obtie-

[2] O sea, *types*, los *tokens* serían mucho más.

ne **1976** bigramas, de los cuales **338** son fraseológicos (**17,1%**) lo que constituye una cifra bastante menos satisfactoria, por otra *fD* apenas logra que la densidad fraseológica se eleve más en el tramo inicial (**19,2%**) (ver **figura 6**). En la categoría V+Adv. el resultado sigue siendo muy bajo: hay 1721 bigramas de este tipo de los que sólo 56 son fraseológicos (**3,25%**). Los adverbios de significado muy general (*bien, mal, mucho, poco, también*, *tampoco*) saturan el listado al acoplarse a todo tipo de verbos: al pertenecer a un paradigma pequeño, se combinan más de una vez con muchos de ellos. Esto conlleva – para esta subcategoría – una total incompatibilidad con la definición "estadística" de la colocación establecida por Halliday (1966) y Sinclair (1970 y 1991). La función magnificadora (Mel'čuk et. al. 1984) justifica perfectamente que las combinaciones del tipo *amar+locamente, llover+torrencialmente* etc., sean colocaciones, sin embargo estadísiticamente quedan muy por detrás de muchas otras que son meramente sintácticas, del tipo *dormir+bien, cantar+mal*, etc.

Figura 6



3 Conclusión

Aunque la técnica de recuperación sigue tropezando con algunos de los obstáculos detectados en los experimentos de calibración, los resultados mejoran claramente con el aumento del corpus, especialmente las combinaciones de tpo V+N y N+Adj. Para estas dos importantes y productivas categorías colocacionales, se puede decir que el resultado es lo suficientemente alentador como para que esta metodología, que combina de forma complementaria herramientas estadísticas y lingüísticas de máxima sencillez, puede aplicarse con rentabilidad a la a la detección de UF con fines lexicográficos en un corpus de gran magnitud.

4 Referencias

BERRY-ROGHE, G. L. (1973) "The computation of collocations and their relevance in lexical studies". en Aitken, A. J., Bailey, R., & Hamilton-Smith, N. (eds.), *The computer and literary studies.* Edimburgo: Edinburgh University Press.

BOSQUE, I. (dir.) (2004) *Redes: Diccionario combinatorio del español contemporáneo.* Madrid: SM editores.

BRADLEY, J et al. (1996) *TACT 2.1*
http://www.chass.utoronto.ca/cch/tact.html (Text Analysis Computer Tools).

CARRERAS, X. CHAO I., PADRÓ L. & PADRÓ M. (2004) "FreeLing: An Open-Source Suite of Language Analyzers". En *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*. Lisboa, Portugal.

CHURCH, K., GALE W, HANKS P, & HINDLE, D (1990): "Using statistics in lexical analysis". En Zernik, U. (ed.) *Lexical Acquisition: Exploiting On-Line Resources*. Hillsdale, NJ: Lawrence Erlbaum Associates, 1990.

CHURCH, K., GALE W, HANKS P, & HINDLE, D. (1989): "Parsing, word associations and typical predicate-argument relations". *Proceedings of the International Workshop on Parsing Technology '89,* pp. 389-398.

COWIE, A. P. (1981) "The Treatment of Collocations and Idioms in Learner´s Dictionaries". *Applied Linguistics,* 2/3, 223-235.

DAILLE, B. (1995) "Combined Approach for Terminology Extraction: lexical statistics and linguistic filtering". *UCREL*, 5 (Univ. of Lancaster), reported by Adam Kilgarriff in http://helmer.aksis.uib.no/corpora/1995-4/0119.html

DUNNING, T (1993): "Accurate Methods for the Statistics of Surprise and Coincidence". *Computational Linguistics*, 19/1.

HALLIDAY, M. A. K. (1961) "Categories of the theory of grammar". *Word* 17 (3): 241-292.

HAUSSMANN, F. J.(1979) "Un dictionnaire des collocations est-il possible?". *Travaux de Linguistique et Litterature* 17: 187-195

HAUSSMANN, F. J. (1981) "Wörterbücher und Wortschatzlernen Spanisch". *Linguistik und Didaktik* 45/46: 71-80.

HAUSSMANN, F. J. (1984) "Wortschatzlernen ist Kollokationslernen". *Praxis des Neusprachlichen Unterrichts* 31: 395-406

HAUSSMANN, F. J. (1985) "Le dictionnaire de collocations". En F. J. Haussmann et al. (eds.) *Wörterbücher, Dictionaries, Dictionnaires : Ein internationales Handbuch zur Lexikographie* (1010-1019). Berlin : Walter de Gruyter

LUQUE DURÁN, J DE D. (1995) "Tipos de diccionario y el diccionario del futuro", en Luque & Pamies (eds.), *Segundas Jornadas sobre Estudio y Enseñanza del Léxico*, Granada: Método, pp. 93-102.

PAMIES, A. & PAZOS, J. M. (2003) "Acceso automatizado a fraseologismos y colocaciones en corpus no etiquetado". *Language Design*, 5: 39-50.

PAMIES, A & PAZOS, J. M. (2004a) "El método estadístico en la detección automatizada de colocaciones y fraseologismos" *VI Congreso Nacional de Lingüística*. Santiago de Compostela 2004 [en prensa].

PAMIES, A & PAZOS, J. M. (2004b) "On automatic retrieval of collocations and idioms in written corpora", *EUROPHRAS 2004* (unpublished). Traducción española en Luque, J.D. & Pamies, A. (eds.), *La creatividad en el lenguaje: colocaciones idiomáticas y fraseología*. Granada: Método 2005.

PAMIES, A & PAZOS, J. M. (2004c) "Extracción automática de colocacións e modismos". *Cadernos de fraseoloxía Galega* 6: 191-203.