Acceso automatizado a fraseologismos y colocaciones en corpus no etiquetado

ANTONIO PAMIES BERTRÁN *(Universidad de Granada)*
JOSÉ MANUEL PAZOS BRETAÑA *(Universidad de Granada)*

Introducción

Sinclair define la colocación léxica como la [reiterada] *ocurrencia de dos o más palabras a corta distancia una de otra en un texto* (1991: 170). Esta definición, independiente de unos presupuestos lingüísticos particulares, ofrece un punto de partida idóneo para investigar la extracción automatizada en un corpus electrónico no anotado. Tiene la ventaja de prestarse a comprobaciones, y de que permite desarrollar distintas hipótesis para el análisis colocacional, tal como han propuesto, entre otros, Church et al. (1989, 1990), Clear (1993), Dunning (1993), o Stubbs (1995). Por otra parte, con este tipo de aproximación es posible trabajar sin necesidad de un gran despliegue de medios técnicos y no se requiere ninguna información lingüística adicional (p.ej., puede aplicarse a un corpus que no requiera previamente ser etiquetado y aprovechar herramientas de análisis ya conocidas)[1]. De este modo, un primer abordaje con métodos meramente cuantitativos compensa por su economía el carácter relativamente "primitivo" y aproximado de lo que, en principio, puede esperarse de ellos.

Intentamos en este trabajo acceder con rapidez a las posibles combinaciones colocacionales de un corpus textual, basándonos en el mero hecho de que la frecuencia de coaparición es el elemento clave, como parte inherente de la propia definición de las colocaciones léxicas. De ser eficaz, esta técnica podría ayudar a construir diccionarios de colocaciones[2]. Naturalmente, las expresiones que sólo aparecen una vez quedan fuera del experimento. El corpus elegido como campo de pruebas ha sido el *Quijote* de Cervantes, cuyo tamaño es, por un lado, lo suficientemente grande para obtener resultados estadísticamente relevantes y, por otro, lo suficientemente reducido como para poder

[1] En este sentido, véase también Aguilar (1994), Kraïf (1997); Pamies, Guirao & Bolívar (1998) o Guirao & Pamies (en prensa).
[2] Véase al respecto, Haussmann 1979, Cowie 1981, Luque 1995 y Koike 2001.

verificar "manualmente" la calidad de la información obtenida. No distinguiremos aquí entre colocación y **unidad fraseológica en general** (**UF**), siendo -por otra parte- las colocaciones una subclase de las UF[3], que también cumplen con la definición de partida (coaparición recurrente de palabras en un texto).

Metodología

Un enfoque meramente cuantitativo de las UF debe comprobar hasta qué punto la frecuencia de un determinado patrón de concurrencia léxica difiere de lo que podría esperarse estadísticamente. Una desviación significativa puede considerarse como un indicio de que la presencia de una determinada palabra influye de alguna manera en la aparición de otras. Podemos considerar dos tipos de frecuencia: una **frecuencia total** (llamamos f_n a la frecuencia total de la base y f_c a la del colocado), que nos indica el número de veces que una palabra aparece en el corpus considerado, y una **frecuencia de coaparición** de dos palabras a una determinada distancia. Para comprobar la relevancia de esta frecuencia, es preciso compararla con otra magnitud: la **frecuencia esperada** (es decir, la frecuencia de coaparición "casual"). Para calcularla, el modelo distribucional más sencillo es el modelo *aleatorio* o *normal*, que 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. La probabilidad aleatoria de coaparición de un bigrama sería igual al producto de las frecuencias totales de sus componentes dividido por el producto de las probabilidades de aparición individual de los mismos (f_n x f_c/ p_n x p_c), siendo éstas, a su vez, iguales a la frecuencia total dividida por el número de palabras del texto. Tests estadísticos como *z-score*[4]: y *t-score*[5] se basan en estos datos.

[3] cf. Zuluaga 1980; Mendívil 1991; Corpas 1996; Ruiz Gurillo 1998; Iñesta & Pamies 2002.

[4] El *z-score* nos permite analizar datos obtenidos de distribuciones que poseen diferentes medias y/o desviaciones estándar y que no podrían ser comparados de otra manera. Viene dado por la fórmula: $z = (O - E)/\sigma$ donde: **O** es la frecuencia en colocación, **E** es la frecuencia de ocurrencia esperada (o normal) del colocado, **σ** es la desviación estándar y equivale a la raíz de la varianza o (desviación cuadrática media). Ésta viene dada por la expresión $\sqrt{N(p(1-p))}$ donde: **N** es la suma de las palabras dentro de los intervalos, **p** es la probabilidad de que esa palabra aparezca en el texto (número total de palabras del texto partido por el número de ocurrencias de dicha palabra).

Dunning (1993) criticó la presunción de distribución aleatoria, poco realista dadas las considerables variaciones comprobadas según cada tipo y tamaño de corpus. Propone el uso de un coeficiente más independiente del tamaño del corpus y que no presupone una distribución normal del léxico, considerando distribuciones binomiales y multinomiales[6].

La información necesaria es un listado de todos los bigramas recurrentes, tomando como contexto relevante un intervalo de x palabras a derecha e izquierda del primer componente, junto con los datos numéricos siguientes para cada pareja:

a) **frecuencia de coaparición (k)**: número de veces que aparecen juntas dos palabras dentro del intervalo elegido;
b) **frecuencias totales (f_n y f_c):** número de veces que aparecen los componentes por separado en la totalidad del texto (en ausencia de conocimientos cualitativos, se presupone convencionalmente que el primero es el nodo y el segundo es el colocando);
c) ***z-score*** magnitud estadística normalizadora que nos indica, en distribuciones normales, la desviación de la media expresada en fracciones o múltiplos de la desviación estándar [índice de variabilidad que nos muestra la dispersión de un conjunto de datos];

---

[5] En cambio, en que el ***t-score*** se considera la desviación estándar **σ** de la base en combinación con el colocado (no únicamente de la del colocado como era el caso en el *z-score*):

$$t = (O - E)/\sigma$$

$$\sigma = \sqrt{N(q(1-q))}$$

donde *q* viene dado por la probabilidad de que la base y el colocado concurran dentro del rango seleccionado.

[6] La fórmula *log likehood,* tal como la reformula Daille (1995), sería la siguiente:
2 [a·loga + b·logb +c·logc +d·logd - (a+b)·log(a+b) - (a+c)·log(a+c) -
- (b+d) ·log(b+d) - (c+d) ·log(c+d) + (a+b+c+d) ·log(a+b+c+d)],
donde **a** es el número de binomios (bigramas recurrentes) en los que X aparece junto a Y; **b** es el número de binomios en los que X **no** aparece junto a Y; **c** es el número de binomios donde Y aparece pero no X; **d** es el número de binomios donde no aparecen ni X ni Y. Así, *a+b* equivale al número de ocurrencias de X; *a+c* al al número de ocurrencias de Y; y *a+b+c+d* al número total de palabras del corpus considerado.

d) ***t-score***: otra magnitud estadística que considera la desviación estándar [varianza] de la supuesta base en combinación con el supuesto colocado (no únicamente de la del colocado como ocurre con *z-score*).
e) **fórmula de Dunning**: magnitud que compara la distribución entre binomios donde ocurren X con Y, X sin Y, Y sin X, ni X ni Y (siendo X e Y componentes de cualquier bigrama recurrente del corpus).

Previamente, elaboramos una tabla de "concordancias" (todas las apariciones en contexto de cada palabra en el texto), realizada con el programa *Concordance v.3.0.* (Watt 2002), que nos servirá de referencia y comprobación, además del propio corpus. Con ayuda del programa *TACT* v.2.1. (Bradley 1996)[7], obtenemos el listado de los bigramas recurrentes y sus tres frecuencias. Las tablas resultantes se procesan con el programa de estadísticas *SPSS* v.11. para calcular *z-score*, *t-score* y la fórmula de Dunning. Se trata como vemos de unos medios computacionales muy "ligeros", con los que, en principio, deberíamos poder extraer una información cuantitativa que identifica las colocaciones léxicas como tales, por ser el correlato numérico de la definición de partida.

Resultados

En un primer intento, claramente fallido, aplicamos -como microcontexto- una ventana de cuatro palabras al texto íntegro del *Quijote*, sin más, con lo cual, se garantiza la recuperacción de la información, pero las unidades detectadas se pierden en el inmenso fichero resultante (un primer intento nos dio 27.250 entradas, equivalentes a unos 500 folios). Resulta imposible comprobar la información dispersada entre miles de combinaciones puramente gramaticales (p.ej. Pron.+V., Aux+ Part, Art+ N, etc).

Para filtrar estos datos "parasitarios", se repite el experimento aplicando un filtro léxico (*stop word list*)[8], que elimina todos los morfemas gráficamente autónomos (determinantes, conjunciones, preposiciones, pronombres, verbos auxiliares, verbos modales y adverbios que no acaben en -mente). Con ello perdemos naturalmente las locuciones conjuntivas, prepositivas y adverbiales (p.ej., *siempre y cuando*, *en contra de, a causa de, de cuando en cuando*), y también colocaciones que se distinguen por su falta de artículo interior (del tipo *abrir fuego*), que se

[7] véase también Pérez Guerra 1998.
[8] cf. Guirao & Pamies (en prensa)

unirán a sus correlatos no colocacionales, o incluso a otra colocación diferente (p.ej. *dar **de** mano ≠ dar **la** mano*). Ello queda ampliamente compensado por las ventajas de la reducción del fichero, ya que, en la práctica, hemos comprobado que la inmensa mayoría de las coapariciones en que participan *ser* y *haber* son meras combinaciones sintácticas con participios para formar tiempos compuestos y/o pasivas. También eliminamos algunos nombres propios de protagonistas principales, porque dan lugar a multitud de co-ocurrencias irrelevantes con palabras de alta frecuencia (*Don Quijote [de la Mancha], Sancho [Panza], Dulcinea [del Toboso]*). Esto nos permite usar como contexto de coaparición una ventana más reducida, puesto que, una vez eliminadas las palabras gramaticales, basta con un intervalo de dos palabras para detectar una colocación. En cuanto a las UF mayores, no quedan eliminadas por dicho binarismo, en el peor de los casos quedan truncadas y solapadas en bigramas diferentes (*pedir cotufas en el golfo> pedir+cotufas & cotufas+golfo*).

Este segundo experimento permite así obtener un fichero que no contiene toda la información deseada, pero que, en cambio, permite examinar y verificar los datos, gracias a sus dimensiones manejables y su mayor adecuación al objeto investigado. El listado obtenido con *TACT* incluye aun así muchas combinaciones recurrentes que cumplen con la citada definición de Sinclair pero que no son UF pese a ello. Estas 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 6,75%[9]. Nuestra hipótesis inicial era que, al ordenar de forma descendente el listado obtenido, según una de las magnitudes estadísticas, las UF se deberían concentrar en una zona en particular, distinguiéndose así de las combinaciones "aleatorias", por criterios exclusivamente estadísticos.

Mostramos a continuación, para cada zona delimitada por los resultados estadísticos, el número de unidades fraseológicas (UF), el número de bigramas recurrentes (BG); la **densidad fraseológica**

[9] Entiéndase unidades distintas (los *tokens* serían muchos más); por otra parte, no olvidemos que sólo se detectan las co-ocurrencias que se dan -como mínimo- dos veces en el corpus. Los datos absolutos también variarían si hubiésemos lematizado el corpus, y si no hubiese solapamiento entre fragmentos de UF mayores truncadas en varios bigramas (*pedir+cotufas* y *cotufas+golfo*). Las únicas cifras realmente representativas son por tanto las proporciones.

(proporción entre el número de unidades fraseológicas y de bigramas para una zona: %UF/ BG), el **volumen fraseológico** (proporción entre el número de UF de una zona y el total de UF del corpus: %UF /TOT), así como el **volumen de co-ocurrencia** (proporción entre el número de bigramas de la zona y el total de bigramas recurrentes del corpus: %BG /TOT).

Tab.1 z-score

z-score	UF	BG	%UF /BG	%UF /TOT	%BG /TOT
>200	14	69	20,29	1,98	0,64
>150	21	83	25,30	2,97	0,77
>100	27	199	13,57	3,82	1,86
>50	81	534	15,17	11,46	4,98
>25	154	1184	13,01	21,78	11,05
>10	229	2614	8,76	32,39	24,39
>9	29	365	7,95	4,10	3,41
>8	40	423	9,46	5,66	3,95
>7	33	481	6,86	4,67	4,49
>6	29	552	5,25	4,10	5,15
>5	18	671	2,68	2,55	6,26
>4	10	689	1,45	1,41	6,43
>3	14	826	1,69	1,98	7,71
>2	4	826	0,48	0,57	7,71
>1	3	656	0,46	0,42	6,12
>0	1	404	0,25	0,14	3,77
<0	0	140	0,00	0,00	1,31

Tab.2 t-score

t-score	UF	BG	%UF /BG	%UF /TOT	%BG /TOT
>5	6	21	28,57	0,85	0,20
>4	14	34	41,18	1,98	0,32
>3	45	125	36,00	6,36	1,17
>2	167	638	26,18	23,62	5,95
>1,5	229	2019	11,34	32,39	18,84
>1	242	6862	3,53	34,23	64,04
<1	4	1017	0,39	0,57	9,49

Tab.3 fórmula de Dunning

F.Dunn.	UF	BG	%UF /BG	%UF /TOT	%BG /TOT
>100	6	28	21,43	0,83	0,26
>50	23	64	35,94	3,18	0,60
>40	29	66	43,94	4,01	0,62
>30	38	122	31,15	5,26	1,14
>20	134	394	36,29	19,78	3,68
>15	95	637	15,23	13,42	5,94
>10	151	1668	9,17	21,16	15,57
>9	23	520	4,42	3,18	4,85
>8	33	689	4,79	4,56	6,43
>7	38	792	4,80	5,26	7,39
>6	62	831	7,82	8,99	7,75
>5	41	964	5,91	5,67	6,48
>4	16	1002	1,60	2,21	9,35
>3	8	977	0,82	1,11	9,12
>2	7	887	0,79	0,97	8,28
>1	2	634	0,32	0,28	5,92
>0	1	441	0,23	0,14	4,12

Discusión

Cuantitativamente, los tres parámetros sólo coinciden parcialmente con 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, porque resulta que las zonas de alta *densidad fraseológica*, delimitadas gracias a los valores superiores de *z-score* y *t-score,* coinciden con zonas de bajo *volumen fraseológico* porque hay su vez poco *volumen de co-ocurrencia* (fig. 1 & fig 2). Este hecho es especialmente llamativo en el caso de *t-score*, ya que las zonas que contienen la gran mayoría del volumen fraseológico están prácticamente en correlación inversa con las de alta densidad fraseológica.

En este sentido, la fórmula de Dunning resulta mejor parada (fig. 3), ya que sus valores superiores delimitan zonas en las que coinciden densidad y volumen fraseológicos. 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 decantación automatizada de las UF del corpus.

El problema reside por tanto en que, a pesar de que a mayor marcador estadístico mayor densidad fraseológica, hay otro tipo de combinaciones, siempre mayoritarias, que consiguen los mismos valores estadísticos que las UF, sobre todo en las zonas de mayor volumen fraseológico.

Cualitativamente, este "ruido" corresponde a varios tipos de bigramas:

a) **nombres propios y topónimos** (*alejandro+magno, julio+césar, amadis+gaula, miguel+cervantes, san+pedro; vélez+málaga, ínsula+barataria*);
b) **combinaciones contextuales** cuya recurrencia es característica del corpus elegido (p.ej. *ingenioso+hidalgo/, yelmo+mambrino/ barbero+cura/ bachiller+cura/ gobierno+ínsula/ bacia+barbero/ ensartar+refranes*);
c) **combinaciones conceptuales** debidas a fuertes nexos lógicos u ontológicos entre dos referentes, pero que siguen siendo combinaciones libres (*come+bebe/ hambre+sed/ silla+sentado/ leer+libros/ mano+derecha/ lado+izquierdo/ rayos+sol/ escudos+oro/ pan+queso/, manadas+ovejas/ jarro+agua/ tronco+árbol/ hojas+árboles/ alfanje+morisco/ pequeña+aldea/ dormir+colchones/ lengua+castellana/ emperador+romano/ mandamiento+cumplir/ torre+alta*);
d) **combinaciones artificiales** producidas por nuestra propia metodología, como, p.ej., la eliminación de los clíticos que separaban dos palabras (*dice+merced <dice* [~~vuesa~~] *merced/*)[10].
e) **reduplicaciones retóricas,** ya sea por **énfasis** (*¡Ladrones, ladrones!/ Cásese, cásese/ ...refranes y más refranes*), **paralelismo** (*de mesón en mesón y de venta en venta*) **antítesis** (*porque yo me case o no me case con aquella señora),* **pleonasmo** (*Carlomagno es Carlomagno*), **anadiplosis** (*no todos los caballeros pueden ser cortesanos ni todos los cortesanos pueden ni deben ser caballeros andantes / la primavera sigue al verano, el verano al estío*...), o **calambur** (*no osa la voz entrar por tan estrecho estrecho*)...

[10] Algunas veces las palabras que se juntaron por esta razón ya formaban de todas maneras una expresión fija, p.ej. *honrada+pata <...~~la~~ mujer honrada, ~~la~~ pata quebrada ~~y~~....*

fig.1 z-score

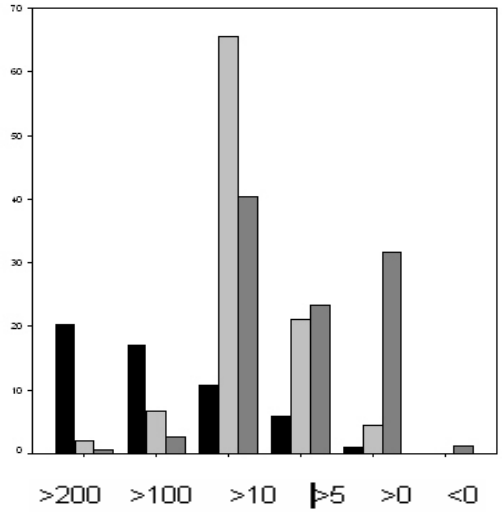


fig.2 t-score

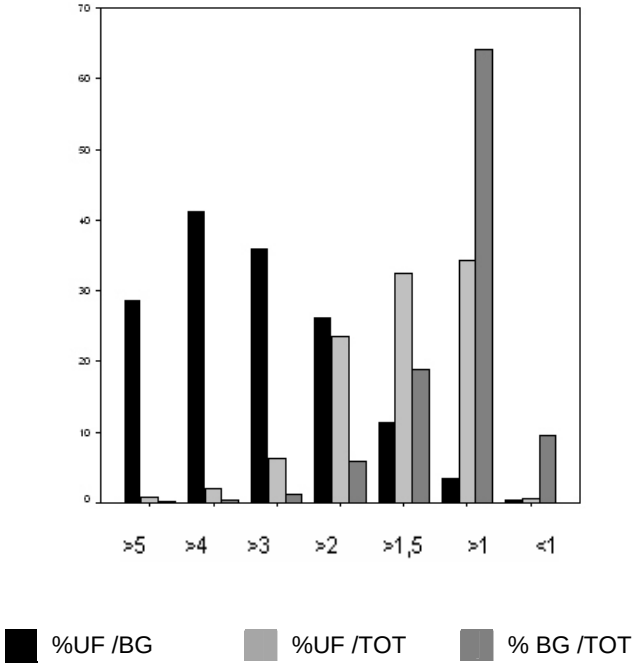
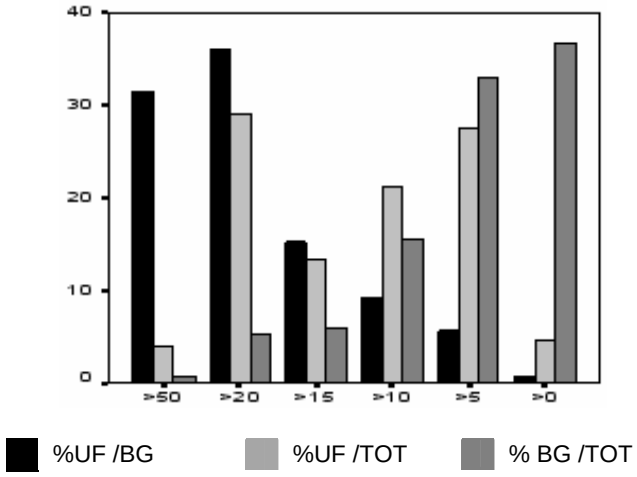


fig.3 fórmula de Dunning



Conclusiones

-La idea de que la colocación consiste en una coaparición léxica recurrente cuya probabilidad es mayor que el producto de las probabilidades de sus componentes por separado (Church et al. 1989) queda en parte confirmada y en parte cuestionada. Es cierta, en la medida en que los valores más altos de fórmulas estadísticas basadas en las frecuencias de aparición y coaparición permiten delimitar zonas de mayor densidad fraseológica. Pero, al mismo tiempo, dicha correlación no permite discriminar mecánicamente las UF de las combinaciones aleatorias que también logran cumplir este requisito en las zonas con mayor volumen fraseológico.

-La fórmula de Dunning es, sin embargo, notablemente más eficaz que *z-score* y *t-score*, ya que logra delimitar zonas con mayor densidad fraseológica que corresponden a un volumen fraseológico más elevado.

-Existe por tanto cierta correlación entre los valores estadísticos basados en las probabilidades de (co)aparición y la naturaleza fraseológica de los bigramas recurrentes, pero ésta es insuficiente para detectar UF basándonos exclusivamente en las frecuencias, porque siempre hay una importante proporción de bigramas no fraseológicos que obtienen los mismos resultados.

Bibliografía

AGUILAR AMAT, A (1994) "Colocaciones en un corpus: detección y aplicaciones", en Martín Vide, C. (ed.) *Actas del X congreso de lenguajes naturales y formales.* Barcelona: PPU.

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

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.

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

CLEAR, J. (1993) "From Firth principles: computational tools for the study of collocation", en Baker et al. (eds.) *Text and Technology*. Amsterdam: John Benjamins

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

CORPAS, G. (ED.) (2000): *Las lenguas de Europa: estudios de fraseología, fraseografía y traducción*. Granada: Comares.

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

DAILLE, BÉATRICE (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.

FELLBAUM, C. (1998): *WordNet. An Electronic Lexical Database*. Massachusetts: MIT Press.

GUIRAO, J.M.; PAMIES, A. (en prensa): "Transfrags, an automatic extraction tool of translation equivalents fragments".

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

HAUSSMANN, F. J. (1985) "Kollokationen im deutschen Wörterbuch: ein Beitrag zur Theorie des lexikographischen Beispiels", en Bergenholtz H. & Mugdan, J. (eds.) *Lexikographie und Grammatik,* Tübingen: Max Niemeyer.

IÑESTA, E.M.; PAMIES, A. (2002): *Fraseología y metáfora: aspectos tipológicos y cognitivos*. Granada: Método/Granada Lingvistica

IRSULA PEÑA, J. (1992) *Substantiv-Verb-Kollokationen*. Frankfurt am Main: Peter Lang

KJELLMER, G. (1990) "Patterns of Collocability", in Aarts, J. & Meijs, W. (eds.) *Theory and practice in corpus linguistics, Language and Computers: Studies in practical linguistics 4*, 163-178. Amsterdam: Rodopi

KOIKE, K (2001) *Colocaciones léxicas en el español actual: estudio formal y léxico semántico.* Madrid: Universidad de Alcalá.

KRAIF, O. (1997): "Modèles probabilistes pour le traitement automatique de corpus textuel", *Travaux du LILLA*, 2. pp. 81-100.

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

LUQUE DURÁN, J. DE D. (2001), Aspectos universales y particulares del léxico de las lenguas del mundo, Granada Lingvistica, Granada.

MENDÍVIL, J. L. (1991) "Consideraciones sobre el carácter no discreto de las expresiones idiomáticas" en Martín Vide (ed.) *Actas del VI congreso de lenguajes naturales y formales.* Barcelona: PPU.

MORENO SANDOVAL, A (1998) *Lingüística computacional.* Madrid: Síntesis.

OAKES, M. P. (1998) *Statistics for Corpus Linguistics.* . Edinburgh: Edinburgh University Press.

PAMIES, A.; GUIRAO, J.M.; BOLIVAR, J. (1998): "Critères pour la détection automatisée de phraséologismes en corpus réel". *Travaux du L.I.L.L.A* (Nice), 3.

PÉREZ GUERRA, J. (1998) *Análisis computerizado de textos. Una introducción a TACT.* Vigo: Universidad.

ROTHKEGEL, A. (1973) *Feste Syntagmen. Grundlagen, Strukturbeschreibung und automatische Analyse*. Tübingen: Niemeyer

RUIZ GURILLO, L. (1998): "Una clasificación no discreta de las unidades fraseológicas del español", en WOTJAK, G. (ed.), *Fraseología y fraseografía del español actual*. Frankfurt/Madrid: Vervuert-Iberoamericana. pp.13-37.

SINCLAIR, J. M. (1991) *Corpus, Concordance, Collocation*. Oxford : Oxford University Press.

SINCLAIR, J.M. (2000): "The search for units of meaning", in Corpas Pastor (ed.): *Las lenguas de Europa: estudios de fraseología, fraseografía y traducción.* Granada: Comares. pp. 7-38.

SINCLAIR, J.M.; JONES, S. (1974) "English Lexical Collocations. A Study in Computational Linguistics" en *Cahiers de Lexicologie* 24, pp. 15-61

SPSS Inc.(2001) *SPSS.v.11*.

STUBBS, M. (1995) "Collocations and semantic profiles: on the cause of trouble with quantitative studies" *Functions of Language 2* pp. 23-55.

WATT, R.J.C. (2002) *Concordance 3.0* (http://www.rjcw.freeserve.co.uk).

WOODS, A.; FLETCHER, P. & HUGHES, A. (1986) *Statistics in language studies*. Cambridge: CUP.

ZULUAGA, A. (1980) *Introducción al estudio de las expresiones fijas*. Frankfurt am Main: Peter Lang.