Cadernos de Fraseoloxía Galega 6,2004,191-203

Extracción automática de colocacións e modismos

Antonio Pamies Bertrán[1]
Universidad de Granada
José Manuel Pazos Bretaña
Universidad de Granada

Nas definicións estatísticas das colocacións estas son descritas como combinacións de palabras que coaparecen con máis frecuencia do que se prediciría a partir das súas frecuencias respectivas e a lonxitude do texto. Dende que Sinclair (1970) propuxo este suposto nos seus traballos leváronse a cabo, con diferentes criterios e métodos, múltiples estudos experimentais con córpora electrónicos dos que se obtiveron resultados diversos (p.ex. Berry-Roghe 1973; Church e Hanks 1989; Clear 1993; Dunning 1993). No noso traballo aplícanse métodos diferentes a un pequeno corpus literario da lingua española co fin de avaliar, co mesmo texto e os mesmos criterios, cada unha das ferramentas metodolóxicas que poderían ser empregadas na detección automática de colocacións en base a datos estritamente cuantitativos, as cales poderían tamén manexar locucións e mesmo refráns.

Statistical definitions of collocations describe them as combinations of words which co-occur more often than their respective frequencies and the length of the text would predict. Since Sinclair's works (1970) proposed this assumption, many experimental works, with different methods and criteria, have been carried out with electronic corpora obtaining different results (e.g. Berry-Roghe 1973; Church & Hanks 1989; Clear 1993; Dunning 1993). Our work applies different methods to a small literary corpus in Spanish language, in order to evaluate, with the same text and the same criteria, each methodological tool that could be involved in automatic detection of collocations on the basis of strictly quantitative data, which should deal also with idioms and even proverbs.

[1] Traducido por Cristina Veiga Novoa.

1. Introdución e metodoloxía

Nun traballo anterior (Pamies e Pazos 2003) investigámo-la extracción automatizada de unidades fraseolóxicas nun corpus (non etiquetado) mediante criterios puramente cuantitativos, presupoñendo que a fraseoloxía, no senso máis amplo do termo, cumpre os requisitos estatísticos que segundo Halliday (1961; 1966: 158) definen como tales as colocacións: a aparición repetida de varias palabras a curta distancia unhas doutras nun corpus textual[2] (véxase tamén Sinclair 1970: 150; 1991: 170; Jones e Sinclair 1974: 19; Berry-Roghe 1973; Church e Hanks 1989; Church et alii 1989; 1991; Clear 1993; Oakes 1998; Manning e Schütze 1999). Para iso aplicamos ó texto completo de *El Quijote* unha técnica de extracción baseada no tratamento automático das frecuencias de aparición e de coaparición das palabras, coa idea de elaborarmos unha metodoloxía susceptible de ser aplicada posteriormente a un corpus de grandes dimensións para a elaboración de grandes dicionarios de colocacións (cf. Haussmann 1979; 1985; Cowie 1981; Luque 1995).

As unidades pluriverbais que só aparecen unha vez quedan naturalmente fóra do alcance do noso experimento, pois este parte exclusivamente da frecuencia e a probabilidade (cf. Church e Hanks 1989), e o corpus foi filtrado previamente para eliminar secuencias recorrentes de carácter meramente gramatical (artigos, pronomes, auxiliares, etc.), para evitar que o ruído[3] que xeran distorsione os resultados. O texto resultante consta de 145.261 palabras. O número de bigramas recorrentes detectados foi de 10.716, dos cales só 707 (6,6%) son colocacións, modismos ou paremias, mentres que o resto (93,4%) eran combinacións lexicamente irrelevantes: nomes propios (*amadis+gaula; vélez+málaga*), figuras retóricas (*reduplicación, anadiplosis, pleonasmo,* etc.), combinacións conceptualmente motivadas (*silla+sentado, jarro+agua*) ou unións contextuais cuxa frecuencia é máis ou menos específica do corpus elixido (*ingenioso+hidalgo, barbero+cura, gobierno+ínsula*)[4]. O listado de bigramas e de frecuencias obtívose co programa Tact 2.1 (Bradley 1996, Pérez Guerra 1998) e o procesamento do mesmo realizouse con SPSS 11, de acordo con tres criterios estatísticos: *z-score*, *t-score* e a fórmula de Dunning (de agora en diante *fD*) co obxectivo de que as unidades fraseolóxicas quedasen automaticamente *decantadas*, separándose do resto ó reordenar decrecentemente o listado a partir dalgún destes tres marcadores.

Estudámo-lo número de bigramas recorrentes (BG), o número de unidades fraseolóxicas (UF), a *densidade fraseolóxica* (proporción entre o número de UF e de BG), o *volume fraseolóxico* (proporción entre o número de UF dun fragmento determinado e o total de UF do corpus) e o *volume de coaparición* (proporción entre o número de BG nun fragmento determinado e o total de BG no corpus). O experimento co texto de *El Quijote*

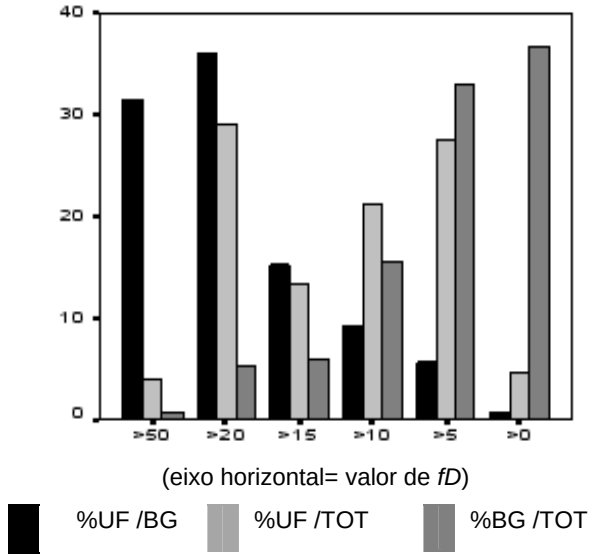
---

[2] *The syntagmatic association of lexical items, quantifiable, textually, as the probability that they will occur at n removes (a distance of* n *lexical items) from an item* x*, the items* a,b,c.... (Halliday 1961).

[3] cf. Kilgarriff e Rundell 2002: 810.

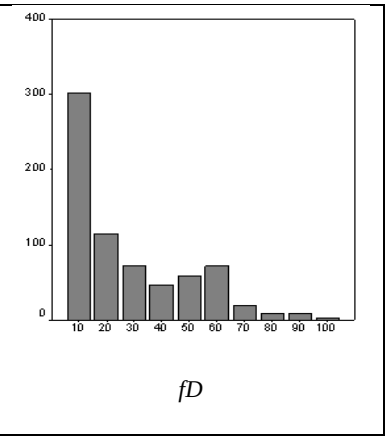
[4] Nos listados obtidos por Berry-Roghe (1973) predominaban as combinacións gramaticais, mais tamén encontramos combinacións "conceptuais" como *house+family*, *house+decorate*, *house+buying* cun *z-score* superior ó de unidades léxicas coma *ghost+house*.

demostrou que existe correlación entre as tres fórmulas baseadas en frecuencias de coaparición e o carácter fraseolóxico ou colocacional dun bigrama recorrente, aínda que non tódolos parámetros o reflictan no mesmo grao. A información mutua (*MI-score*), proposta por Church e Hanks en 1989, resultou moi pouco eficaz, polo que abandonamos incluso o seu emprego, malia que dá bos resultados para as unidades frecuentes cuxos compoñentes, por separado, teñen unha frecuencia relativamente baixa (cf. Kilgarriff e Rundell 2002, Čermak 2004). *T-score* (proposto en Church et alii 1991) resulta máis eficaz ca *z-score* (proposto por Berry-Roghe 1973), pero menos có *log-likelihood* proposto por Dunning en 1993 (*fD*), que é o que obtén a maior densidade fraseolóxica na parte inicial do listado: a maior valor de *fD*, maior proporción de UF nun fragmento e á inversa (Pamies e Pazos 2003).



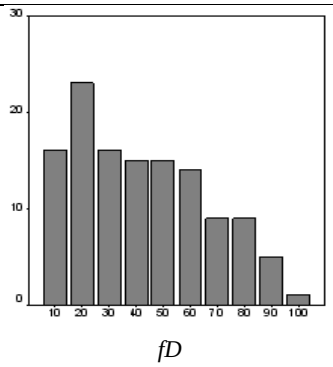
Nun segundo experimento modificámo-lo método para que os resultados fosen extrapolables a outro corpus. As subdivisións do eixo horizontal do gráfico xa non debían corresponder a unha segmentación do listado en valores absolutos de *fD* (>50, >20 etc.), senón a anacos do listado iguais entre si (p.ex. 10%). Reanalizamos así o texto de *El Quijote* e verifícase aínda mellor a correlación: a maior valor do marcador estatístico, maior densidade fraseolóxica, así como a superioridade da fórmula de Dunning sobre *z-score* e *t-score*: 302 UF concéntranse no primeiro 10% do listado de bigramas recorrentes ordenado de forma decrecente segundo o valor de *fD* (42,7% do total do volume fraseolóxico). A densidade fraseolóxica (%UF/BG) do tramo inicial é de 28%.

BG	UF		
	fD (dens. %)	*t-score*	*z-score*
0 a 10%	302 (28%)	288	164
10 a 20%	115 (11%)	128	139
20 a 30%	73 (7%)	71	115
30 a 40%	47 (4%)	65	81
40 a 50%	58 (5%)	41	87
50 a 60%	72 (7%)	57	69
60 a 70%	19 (2%)	37	27
70 a 80%	9 (1%)	7	15
80 a 90%	9 (1%)	9	6
90 a 100%	3 (0,3%)	4	4



Para comproba-la extrapolabilidade destes criterios aplicámolos a outro corpus de fácil verificación polo seu reducido tamaño (non hai que esquecer que a discriminación das UF se comproba manualmente), neste caso a novela *La familia de Pascual Duarte* de C.J. Cela, texto breve mais abundante en modismos e paremias. Unha vez eliminados do texto os artigos, pronomes, preposicións, auxiliares, etc. extraémo-los bigramas recorrentes e ordenámolos decrecentemente segundo o valor de *fD*. O corpus queda en 14.261 palabras, das que se extraen automaticamente 556 bigramas recorrentes. A densidade fraseolóxica total do corpus é moi alta, a pesar de que só sexan detectables as UF que aparecen máis dunha vez: 123 bigramas corresponden a UF (unha proporción tres veces máis elevada ca en *El Quijote*). A análise estatística obtén unha pendente decrecente, agás o primeiro tramo, un imprevisto que requirirá unha comprobación cualitativa.

BG	UF (dens.%)
0 a 10%	16 (29%)
10 a 20%	23(41%)
20 a 30%	16 (29%)
30 a 40%	15 (27%)
40 a 50%	15 (27%)
50 a 60%	14 (25%)
60 a 70%	9 (16%)
70 a 80%	9 (16%)
80 a 90%	5 (9%)
90 a 100%	1 (2%)

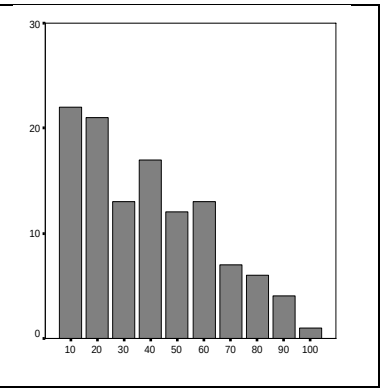


2. Discusión

Nunha análise cualitativa comprobamos que o primeiro tramo do listado contén máis nomes propios e duplicacións ca UF. As duplicacións son especialmente "nocivas", xa

que resultan privilexiadas no cálculo de *fD* porque *TACT-2.1*. computa erroneamente a coaparición dun termo consigo mesmo[5]. Isto lévanos a efectuar un terceiro experimento: eliminar tanto os nomes propios coma as duplicacións, de xeito que só se procesen bigramas realmente susceptibles, a priori, de seren colocacións ou modismos. Os 556 bigramas recorrentes de *La familia de Pascual Duarte* corresponden en realidade a 63 nomes propios (ou combinación recorrente dun nome propio con outra palabra), 88 duplicacións (das cales unha tamén é un nome propio), 114 UF, máis 9 duplicacións que tamén son UF, e 282 combinacións aleatorias (conceptuais, contextuais ou casuais, do tipo *domingos+misa*, *litro+vino*). Se eliminamos nomes propios e duplicacións[6], os bigramas recorrentes (451)[7] reordenaríanse dentro do listado do seguinte modo:

BG	**UF** (dens.%)
0 a 10%	22 (49%)
10 a 20%	21 (47%)
20 a 30%	13 (29%)
30 a 40%	17 (38%)
40 a 50%	12 (27%)
50 a 60%	13 (29%)
60 a 70%	7 (16%)
70 a 80%	6 (13%)
80 a 90%	4 (9%)
90 a 100%	1 (2%)



A liña volve ser descendente e a densidade fraseolóxica do primeiro tramo é do 49%. Isto supón unha melloría considerable con respecto ó método anterior pero, a pesar de todo, non se pode dicir que se conseguise separar as combinacións fraseolóxicas das aleatorias (estas mesmo representan un 51% no primeiro tramo)[8].

Outro factor susceptible de influencia-lo resultado estatístico é a lematización do corpus (Kilgarriff e Rundell 2002: 811): se as variantes gramaticais se unifican, sumaranse entre si no cómputo, polo que non só se incrementa o seu valor de *fD*, senón que as posibilidades de detección aumentan ó agrupárense entre si unhas variantes que, por apareceren unha única vez, quedaban excluídas nunha busca sen lematización. Por iso, realizamos un cuarto experimento aplicando un lematizador (Moreno e Guirao 2003) ó mesmo corpus, eliminando de principio nomes propios e duplicacións, obtendo un total de 1.054 bigramas recorrentes, dos que 265 son UF, coa distribución seguinte:

---

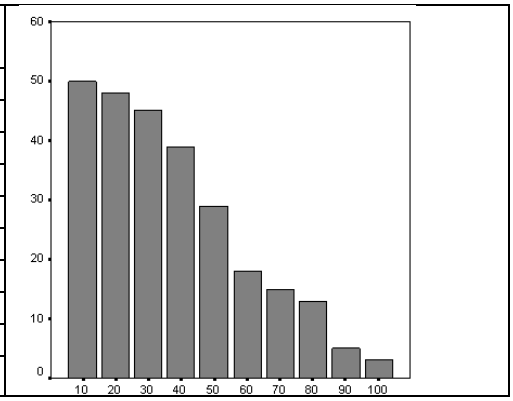
[5] Por exemplo, para o bigrama <*yegua+yegua*>sinala 4 coaparicións cando na verdade só hai dúas: <*¡To, yegua! ~~La~~ yegua se arrimó...*>, <-*¡To, yegua! ~~La~~ yegua se movía...*>, ou ben en <*enflaqueciendo +enflaqueciendo*> a coincidencia dunha palabra consigo mesma é computada erroneamente como dúas coincidencias <*vamos enflaqueciendo, enflaqueciendo*>, o que eleva artificialmente o valor de *fD*.

[6] Sacrificamos ó facelo as 9 UF que, ademais, eran duplicacións (*gota* <*~~a~~*> *gota, costase* <*~~lo que~~*> *costase*).

[7] Ó non incluírmos xa as duplicacións, queda corrixido o cómputo dos bigramas.

[8] A inevitable recorrencia deste atranco xa foi sinalada por Church e Hanks en 1989. Fontenelle (2001: 82) tamén afirma: "les données extraites par ces outils, aussi précieuses soient-elles, sont sémantiquement hétérogènes…"

BG	UF (dens.%)
0 a 10%	50 (47%)
10 a 20%	48 (46%)
20 a 30%	45 (43%)
30 a 40%	39 (37%)
40 a 50%	29 (28%)
50 a 60%	18 (17%)
60 a 70%	15 (14%)
70 a 80%	13 (12%)
80 a 90%	5 (5%)
90 a 100%	3 (3%)



Como vemos, a lematización permitiu duplica-lo número de UF detectadas, tanto na totalidade do texto coma nas que se acumulan no primeiro tramo, manténdose –con maior regularidade- a liña descendente da densidade fraseolóxica entre cada tramo, polo que resulta metodoloxicamente máis eficaz.

Con todo, proporcionalmente, a densidade fraseolóxica é a mesma ca no texto non lematizado: no mellor dos casos (primeiro tramo) seguimos arredor do 49% de UF. Isto demostra que os bigramas recorrentes aleatorios tamén se viron beneficiados pola lematización na mesma proporción. As combinacións casuais (*sangrar+pensar, orar+clavo*) ou temáticas (*fumar+pitillo, encender+luz*) demostran ser tan frecuentes coma as combinacións fraseolóxicas no primeiro tramo, e moito máis frecuentes se consideramos a totalidade dos bigramas extraídos con métodos estatísticos. No mellor dos casos, o "ruído" representa aquí o 51% malia os filtros empregados.

Outra posibilidade sería amplia-lo tamaño da fiestra de busca, aínda que é de supoñer que, ó igual que acontece coa lematización, o ruído tamén se beneficiaría na mesma proporción, pois como observan Kilgarriff e Rundell:

> *different windows show different kinds of information* [...] *lists of this type are usefully suggestive but contain too much noise, require too much interpretation and are too arbitrary in how they are specified, to be an indispensable lexicographic tool* (Kilgarriff e Rundell 2002: 811)[9].

3. Aspectos cualitativos

A metodoloxía aquí descrita exclúe totalmente os pares de palabras que coaparecen unha única vez, o que podería eventualmente influír no resultado. Para esquivar este obstáculo o sistema máis simple é duplica-lo texto[10], estatisticamente non mellora a recuperación automática, posto que o ruído tamén será favorecido na mesma proporción; así e todo,

[9] Por outra parte, Berry-Roghe (1973) xa sinalaba que a fiestra óptima depende da natureza estilística do corpus elixido.

[10] Só leva uns minutos, posto que o texto xa está filtrado e lematizado.

permite obter datos estatísticos sobre calquera combinación independentemente da súa frecuencia.

Aplicamos este método nun quinto experimento e obtemos en *La familia de Pascual Duarte* un total de 11.921 bigramas, dos cales 10.871 corresponden a pares que coaparecen unha única vez, é dicir, obtéñense dez veces máis datos cun corpus dúas veces maior. Naturalmente, as combinacións irrelevantes resultaron igualmente favorecidas por esta operación, pero isto permítenos investigar unhas colocacións en particular con respecto ó resto das combinacións.

Investigando as colocacións verbo-nominais cuxo soporte son os verbos *dar, hacer, tener* e *tomar*, obtemos en *Pascual Duarte* 230 bigramas dos que só 53 foran detectados no texto "simple", ou sexa, detectando 177 bigramas que só coaparecen unha vez, entre os cales 33 son UF (*dar+consejo, dar+corazonada, dar+dolor, dar+abrazo,* etc). Estes resultados demostran que a capacidade de detección de colocacións (a partir dun compoñente) se incrementa nunha razón moi superior a dous se duplicámo-lo corpus:

dar: 230 BG [53 UF] (só 53 BG [20 UF] coaparecen dúas ou máis veces)

hacer: 217 BG [48 UF] (só 52 BG [16 UF] coaparecen dúas ou máis veces)

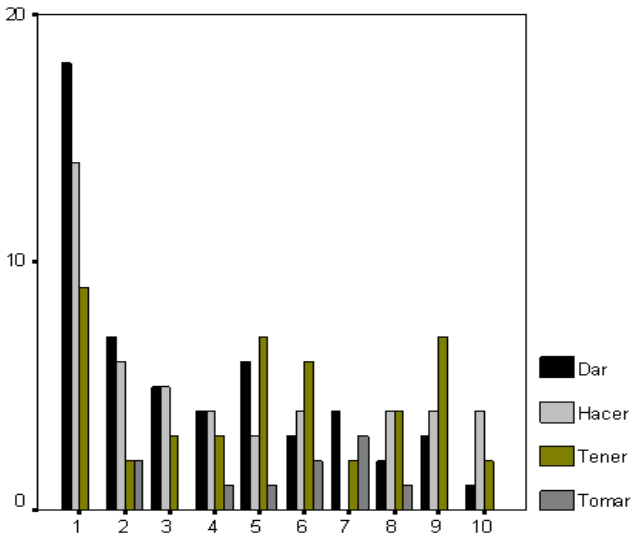
tener: 272 BG [45 UF] (só 72 BG [13 UF] coaparecen dúas ou máis veces)

tomar: 28 BG [10 UF] (só un bigrama coaparecía máis dunha vez, *tomar+tren*)

Dentro do ficheiro que contén só as combinacións do verbo seleccionado, verifícase que, ó seren ordenadas mediante o valor *fD* de cada combinación, a maior densidade colocacional tende unha vez máis a concentrarse na zona inicial do listado[11].

BG	UF							
	dar		*hacer*		*tener*		*tomar*	
tramo	nº	%	nº	%	nº	%	nº	%
0 a 10%	18	78	14	64	9	33	0	0
10 a 20%	7	30	6	27	2	7	2	67
20 a 30%	5	22	5	23	3	11	0	0
30 a 40%	4	17	4	18	3	11	1	33
40 a 50%	6	26	3	14	7	26	1	33
50 a 60%	3	13	4	18	6	22	2	67
60 a 70%	4	17	0	0	2	7	3	100
70 a 80%	2	9	4	18	4	15	1	33
80 a 90%	3	13	4	18	7	26	0	0
90 a 100%	1	4	4	18	2	7	0	0

[11] O resultado é negativo para o verbo *tomar*, pero o número total de datos para esta palabra é tan escaso neste texto (10% de 30 aparicións) que a estatística non pode funcionar, como parece confirmalo o feito de que *fD* e *z-score* dean exactamente o mesmo resultado nesta columna.



Distribución do nº de UF no listado ordenado segundo *fD*.

Así mesmo, compróbase que, tamén en buscas lingüisticamente orientadas, a fórmula de Dunning é moito máis eficaz ca *z-score*, cuxo resultado é moito máis confuso para a mesma operación[12]:

BG	UF							
	dar		*Hacer*		*tener*		*tomar*	
Tramo	nº	%	nº	%	nº	%	nº	%
0 a 10%	12	52	15	68	4	15	0	0
10 a 20%	4	17	1	5	3	11	2	67
20 a 30%	9	39	8	36	5	19	0	0
30 a 40%	7	30	4	18	6	22	1	33
40 a 50%	8	35	5	23	7	26	1	33
50 a 60%	4	17	1	5	4	15	2	67
60 a 70%	3	13	2	9	4	15	3	100
70 a 80%	2	9	4	18	3	11	1	33
80 a 90%	2	9	5	23	7	26	0	0
90 a 100%	2	9	3	14	2	7	0	0

Para comproba-la extrapolabilidade desta hipótese, aplicámo-lo mesmo test a *El Quijote*, filtrado, lematizado e duplicado, cun resultado menos vistoso pero máis homoxéneo.

[12] Verlinde et alii (2003) aplican un método deste tipo para a creación de dicionarios: extracción orientada de bigramas mediante concordancias (*Wordcruncher*) nun gran corpus periodístico francés (non-etiquetado), empregando *z-score* como criterio de relevancia estatística. Os nosos datos permítennos sospeitar que a clasificación obtida sería máis produtiva coa fórmula de Dunning.

dar: 1157 BG [103 UF]: só 491BG [42 UF] coaparecen dúas ou máis veces

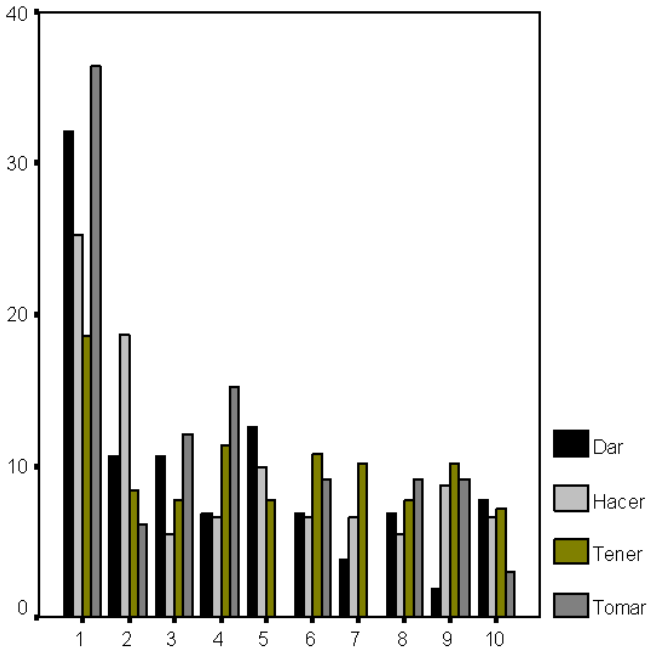
hacer: 1258 BG [91 UF]: só 514 BG [39 UF] coaparecen dúas ou máis veces

tener: 1617 BG [167 UF]: só 633 BG [79 UF] coaparecen dúas ou máis veces

tomar: 507 BG [33 UF]: só 123 BG [15 UF] coaparecen dúas ou máis veces

Ordenado o listado obtido polo valor de *fD*, a maior densidade colocacional concéntrase no tramo inicial, o que confirma a existencia dunha correlación entre o devandito marcador estatístico e o carácter fraseolóxico-colocacional dunha combinación de palabras.

BG	UF							
Tramo	*dar*		*hacer*		*tener*		*tomar*	
	nº	%	nº	%	nº	%	nº	%
0 a 10%	33	32,04	23	25,27	31	18,56	12	36,36
10 a 20%	11	10,68	17	18,68	14	8,38	2	6,06
20 a 30%	11	10,68	5	5,49	13	7,78	4	12,12
30 a 40%	7	6,80	6	6,59	19	11,38	5	15,15
40 a 50%	13	12,62	9	9,89	13	7,78	0	,00
50 a 60%	7	6,80	6	6,59	18	10,78	3	9,09
60 a 70%	4	3,88	6	6,59	17	10,18	0	,00
70 a 80%	7	6,80	5	5,49	13	7,78	3	9,09
80 a 90%	2	1,94	8	8,79	17	10,18	3	9,09
90 a 100%	8	7,77	6	6,59	12	7,19	1	3,03



O marcador *z-score* para a mesma proba volve ser menos fiable ca *fD*:

BG	UF							
	dar		*hacer*		*tener*		*tomar*	
tramo	nº	%	nº	%	nº	%	nº	%
0 a 10%	27	26,21	22	24,18	17	10,18	3	9,09
10 a 20%	12	11,65	10	10,99	18	10,78	9	27,27
20 a 30%	8	7,77	10	10,99	23	13,77	6	18,18
30 a 40%	9	8,74	8	8,79	15	8,98	2	6,06
40 a 50%	15	14,56	6	6,59	22	13,17	3	9,09
50 a 60%	9	8,74	7	7,69	17	10,18	2	6,06
60 a 70%	8	7,77	3	3,30	11	6,59	1	3,03
70 a 80%	5	4,85	12	13,19	17	10,18	3	9,09
80 a 90%	4	3,88	6	6,59	16	9,58	3	9,09
90 a 100%	6	5,83	7	7,69	11	6,59	1	3,03

4. Conclusións

a) A definición das unidades fraseolóxicas baseada nun criterio estatístico é "útil" na medida en que permite extraer automaticamente un listado de bigramas recorrentes que, ó seren ordenados segundo o valor decrecente de *fD*, concentra no seu primeiro 10% unha alta proporción de UF.

b) Con todo, esta definición non se axusta á realidade empírica, na medida en que a frecuencia de coaparición é un trazo compartido por outras combinacións aleatorias (conceptuais ou casuais) nunha proporción bastante superior á das UF, constituíndo, mesmo no mellor dos casos, a maioría do listado extraído (o tramo con maior valor de *fD*).

c) A lematización do corpus aumenta considerablemente a capacidade de detección de combinacións fraseolóxicas en termos absolutos, e presenta unha agrupación máis coherente dos valores de *fD*. Pero tamén aumenta a detección de bigramas non-fraseolóxicos, de xeito que a proporción entre ámbolos dous tipos de combinación no listado extraído non varía moito con respecto ó obtido nun corpus non-lematizado.

d) A duplicación dun mesmo corpus lematizado permite recupera-las combinacións que, por apareceren unha soa vez, quedarían excluídas da detección. Dende o punto de vista cuantitativo o resultado non mellora porque o ruído aumenta na mesma proporción mais, cualitativamente, permite multiplicar considerablemente a capacidade de detectar combinacións fraseoloxicamente relevantes, por exemplo, obter bases a partir do colocativo ou viceversa.

De todo isto dedúcese que os métodos matemáticos chocan unha e outra vez cos mesmos obstáculos. Non sorprende que a maioría das investigacións actuais se oriente cara a métodos "híbridos" que combinan a estatística con criterios de carácter lingüístico, como por exemplo a categoría sintáctica dos compoñentes nun corpus etiquetado (Daille 1994, Heid et alii 2000, Daille e Williams 2001, Evert e Krenn 2001, Zinsmeister e Heid 2002, Kilgarriff e Rundell 2002, Pearce 2002, Čermák 2004, Granjer et alii 2004). Con todo, convén tamén salientar que as devanditas limitacións dos modelos estatísticos "puros" derivan á súa vez da mesma definición "clásica" de colocación, xa que foron os lingüistas quen as describiron exclusivamente en termos de frecuencia de coaparición (Halliday 1961, 1966; Sinclair 1970: 150, 1991: 170; Jones e Sinclair 1974:19), o que resulta ser unha condición necesaria pero non suficiente.

5. Bibliografía

BERRY-ROGHE, G.L.M. (1973): "The computation of collocations and their relevance in lexical studies" en AITKEN, A. J. et alii (eds.): *The computer and literary studies.* University Press, Edinburgh.

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

ČERMÁK. F. (2004): "Statistical Methods for searching Idioms in Text Corpora". [Comunicación presentada en *Europhras 2004 (Basilea, 2004)*].

CHURCH, K. e HANKS P. (1989): "Word association norms, mutual information and leicography" en *Computational Linguistics* 16/1,1989,22-29.

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

(1991): "*Using statistics in lexical analysis*" en ZERNIK, U. (ed.): *Lexical Acquisition: Exploiting On-Line Resources*. Lawrence Erlbaum Associates, Hillsdale.

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

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

DAILLE, B. (1994): *Combined approach for terminology extraction: lexical statistics and linguistic filtering*. Université de Paris, Paris. (http://www.comp.lancs.ac.uk/ucrel/tech_papers.html).

DAILLE, B. e WILLIAMS, G. (EDS.) (2001): *Proceedings of ACL Workshop on Collocations: Computational Extraction, Analisis ad Exploitation*. Université de Toulousse, Toulousse.

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

EVERT, S. e KRENN, B. (2001): "Methods for the qualitative evaluation of lexical association measures" en *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*. Université de Toulousse, Toulousse.

FONTENELLE, T. (2003): "Collocations et traitement automatique du langage naturel" en GROSSMANN, F. e TUTIN, A. (eds.): *Les Collocations. Analyse et traitement.* De Werelt, Amsterdam, 75-88.

GRANJER, S., PAQUOT, M., RAYSON, P. e FAIRON, C. (2004): "Extraction of multi-word units from EFL and native English corpora: the phraseology of the verb *make*". [Comunicación presentada en *Europhras 2004 (Basilea, 2004)*].

GROSSMANN, F. e TUTIN, A. (eds.) (2003): *Les Collocations: Analyse et traitement*. De Werelt, Amsterdam.

GUIRAO, J.M., PAMIES, A. (no prelo): "Transfrags, an automatic extraction tool of translation equivalents fragments".

HALLIDAY, M. A. K. (1961): "Categories of the Theory of Grammar" en *Word* 17,1961,241-292.

(1966): "Lexis as a linguistic level" en BAZELL et alii (eds.): *In Memory of John Firth*. Longman, London, 148-162.

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

(1985): "Kollokationen im deutschen Wörterbuch: ein Beitrag zur Theorie des lexikographischen Beispiels" en BERGENHOLTZ H. e MUGDAN, J. (eds.): *Lexikographie und Grammatik.* Max Neumeyer, Tübingen.

HEID, U., EVERT, S., DOCHERTY, V., WORSCH, W. e WERMKE, M. (2000): "Computational tools for semi-automatic corpus-based updating of Dictionaries" en *EURALEX 2000* [Stuttgart], 183-196.

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

KILGARRIFF, A. e RUNDELL, M. (2002): "Lexical Profiling Software and its Lexicographic Applications – a Case Study" en BRAASCH, A. e POVLSEN,

C. (eds.): *Proceedings of the Tenth EURALEX International Congress (Copenhagent 2002)*. Københavns Universitet, Copenhagen, 807-818.

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

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

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

MANNING, C. e SCHÜTZE, H. (1999): *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.

MERKEL, M., NILSSON, B. e AHRENBER, L. (1994): "A phrase-retrieval system based on recurrence" en *Proceedings of the Second Annual Workshop on Very Large Corpora (WVLC-2)*, Kyoto, 43–56.

MORENO SANDOVAL, A. e GUIRAO MIRAS, J.M. (2003): "Tagging a Spontaneous Speech Corpora of Spanish" en *Proocedings of the International Conference in Recent Advances in Natural Language Processing (Borovets 2003)*, 292-296.

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

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

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

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

PEARCE, D. (2002): "A Comparative Evaluation of Collocation Extraction Techniques" en *Third International Conference on Language Resources and Evaluation (Las Palmas de Gran Canaria, 2002)*.

SINCLAIR, J. et alii (1970): *English lexical studies.* OSTI Report. University of Birmingham, Birmingham.

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

VERLINDE, S., SELVA, T. e BINON, J. (2003): "Les collocations dans les dictionnaires d'apprentissage: repérage, présentation et accès" en GROSSMANN, F. e TUTIN, A. (eds.): *Les Collocations. Analyse et traitement.* De Werelt, Amsterdam,105-116.

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

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

ZINSMEISTER, H. e HEID, U. (2002): "Collocations of complex words: implications for the acquisition with a stochastic grammar" en *Proceedings of the International Workshop on 'Computational Approaches to Collocations' (Vienna, 2002)*.