RECUENTO ESTADISTICO DE PALABRAS, LETRAS, DIGRAMAS Y TRIGRAMAS EN TITULOS DE ARTICULOS CIENTIFICOS Y TECNICOS EN ESPAÑOL

Francisco Gutiérrez Muñoz*, Gloria del Rey Gutiérrez; Alfredo del Rey Guerrero*

Resumen: Se muestran los resultados de una serie de estudios estadísticos realizados con el fin de determinar las frecuencias de aparición de las plabras, letras, digramas y trigramas contenidos en una colección de 12.540 títulos de artículos científicos y técnicos en español obtenidos a partir de la base de datos ICYT. Se discute la metodología seguida, y se sugieren posibles aplicaciones de tales resultados.
Palabras clave: Digramas, Trigramas, títulos, linguística estadística.

Abstracts: A series of statistic studies were carried out to determine the occurrence frecuency of words, characters, digrams and trigrams, in a total of 12.540 titles of scientific and technic papers indexed in the database ICYT. The are discussed the used methods and are exposed the probable uses of the obtained results.
Kewordy: Digrams, trigrams, titles, statistics linguistic.

Introducción

La base de datos del Indice Español de Ciencia y Tecnología desarrollada en el ICYT plantea problemas de ocupación excesiva de memoria en disco, lo que hace conveniente su compactación. También, los errores ortográficos introducidos durante la grabación de los registros dan lugar a una pérdida de calidad en el producto resultante y a un aumento notable del tamaño de los ficheros de descriptores que se han de generar para la recuperación automatizada de información a partir de dicha base de datos. En consecuencia, se hace necesario dar solución a tales problemas.

Con tal fin, decidimos que lo más conveniente sería desarrollar, nosotros mismos, los programas de compresión de textos y de detección y corrección de errores ortográficos y tipográficos que se precisan. De entre los métodos existentes orientados a tal fin (1-6), decidimos adoptar, por su sencillez y eficacia, los que se basan en estudios de frecuencias de aparición de digramas y trigramas (combinaciones de dos o tres caracteres, respectivamente).

Por ello, y con el objeto de disponer de una base sólida para poder desarrollar tales programas, utilizando el miniordenador SECOINSA 40 del ICYT y el lenguaje COBOL, llevamos a cabo una serie de recuentos estadísticos a partir de una colección de 12.540 títulos de artículos científico -técnicos escritos en español, que se obtuvieron de la base de datos del Indice Español de Ciencia y Tecnología. Parte de los resultados obtenidos ya fueron publicados (7).

En este trabajo se reflejan, de forma resumida, los resultados más relevantes de todos los recuentos estadísticos que se han realizado en la fase previa al desarrollo de los programas de compresión de textos y de corrección de errores, actualmente en marcha.

* Instituto de Información y Documentación en Ciencia y Tecnología (ICYT), Madrid. CSIC.

Resultados

La colección de partida contiene 12.540 títulos de artículos científicos y técnicos originalmente en español. Dado que dichos títulos habían sido grabados utilizando solamente letras mayúsculas, no se han tenido en cuenta, en este estudio, ni minúsculas, ni acentos ni signos diacríticos.

La distribución temática de los títulos incluídos en la colección de trabajo se muestra en la Tabla 1, y es un reflejo de la cobertura del propio Indice Español de Ciencia y Tecnología, que abarca todos los campos de la ciencia y la tecnología con excepción de la medicina, cubierta, de manera exhaustiva, por otra publicación similar: el Indice Médico Español.

Tabla 1

Distribución temática de títulos

Area Temática	Incidencia (%)
Tecnología	37.66
Ciencias Agrarias	15.12
Ciencias Biológicas	14.72
Ciencias Químicas	9.5
Ciencias de la Tierra	9.38
Ciencias Matemáticas	5.52
Ciencias Físicas	5.32
Farmacia	2.76

A partir de dicha colección de títulos se independizaron en total 142.944 palabras, de las cuales sólo 18.099 eran diferentes. El promedio de palabras extraídas por título fue de 11'3, con valores extremos comprendidos entre 1 y 48 palabras por título. Se consideró como palabra a cualquier secuencia de caracteres alfanuméricos que comenzase por una letra y que estuviese delimitada por espacios en blanco, signos de puntuación o caracteres especiales. Su frecuencia de aparición resultó ser muy variable, así: un 54,44 por ciento de las palabras diferentes aparece una sola vez en la colección; por el contrario, un 0,56 por ciento de tales palabras, 101 palabras, aparece con mucha frecuencia, suponiendo su aportación el 50,04 por ciento del total de ocurrencias. En la Tabla 2 pueden verse las 20 palabras más frecuentes: su aportación supone el 40,17 por ciento del total; son consideradas como palabras vacías.

La longitud media de las palabas contenidas en los títulos resultó ser de 5,81 caracteres: Sin embargo, si sólo se tiene en cuenta la colección de palabras indentificadas como diferentes, la longitud media resultante es de 8,97 caracteres por palabra. En cualquier caso, la gran mayoría de las palabras tiene una longitud menor de 16 caracteres, el 99,28 por ciento de las palabras de los títulos, frente al 96,86 de las palabras diferentes.

Partiendo de la colección de palabras diferentes obtenida a partir de los títulos y teniendo en cuenta sus frecuencias de aparición respectivas, se es-

tudiaron las distribuciones de frecuencias de letras, digramas y trigramas. Todas ellas reflejan la desigual frecuencia de aparición de las palabras en los títulos de la colección de partida y, en particular, la fuerte incidencia que tienen las palabras de alta frecuencia incluidas en la Tabla 2.

Tabla 2

Relación de palabras de alta frecuencia de aparición

Palabra	Frecuencia	Longitud
DE	20.227	2
LA	6.781	2
EN	5.693	2
Y	4.407	1
DEL	3.353	3
EL	2.362	2
LOS	2.304	3
LAS	1.755	3
SOBRE	1.314	5
PARA	1.263	4
ESTUDIO	1.260	7
A	1.142	1
CON	1.015	3
POR	944	3
UN	821	2
I	611	1
AL	607	2
UNA	572	3
II	563	2
SU	431	2

Las frecuencias de aparición de las letras simples en los títulos de la colección se muestran en la Tabla 3. Puede apreciarse en ella la preponderancia de las vocales, cuya presencia supone el 45,23 por ciento de los caracteres más frecuentes, destacan las letras; S, R, N, C, L, D y T.

Para determinar la ocurrencia de digramas y trigramas se consideró que a cada palabra de la colección de palabras diferentes le precede un espacio en blanco y le sigue otro blanco. Los digramas y trigramas podrían estar formados por combinaciones de caracteres que estuviesen contenidos en una lista de 38, integrada por las 27 letras simples del alfabeto español, los diez números dígitos y el espacio en blanco. En las tablas de digramas y trigramas (tabla 4 y tabla 5, respectivamente) el espacio en blanco se ha representado mediante un guión. Por razones de simplificación, en esta primera aproximación al tema no se han tenido en cuenta como posibles integrantes de digramas y trigramas a los signos ortográficos más frecuentes, como son: el punto, la coma, el punto y coma, etc., conscientes de que aunque ello no era interesante desde el punto de vista de la corrección de errores sí lo era desde el de la compresión de textos.

Tabla 3

Frecuencia de aparición de letras

Letra	Frecuencia	Letra	Frecuencia
E	99.789	G	9.401
A	95.304	B	9.305
O	79.085	F	8.759
I	77.641	V	7.551
S	60.264	Y	5.565
N	56.622	H	4.087
R	50.933	Z	30.033
C	50.696	X	2.146
L	49.133	Q	1.754
D	48.592	J	1.585
T	40.944	Ñ	1.351
U	23.934	K	352
M	21.979	W	216
P	19.974		

Para hacer el recuento de los digramas se creó una tabla de 38 filas por 38 columnas, en ellas, cada uno de sus elementos correspondía al contador de frecuencias de un determinado digrama. Por limitaciones del ordenador con el que se ha trabajado, en el caso de los trigramas hubo que seguir un procedimiento diferente, que consistía en simular el desplazamiento, sobre cada línea de texto, de una ventana de tres celdillas (una por cada letra), con lo que se pudieron independizar, para posterior recuento, todos los trigramas contenidos en la colección de palabras.

Aunque en teoría podrían existir 1.444 digramas diferentes (38 x 38), sólo se encontraron 769, de los cuales 33 aportan el 50 por ciento de los 972.789 digramas obtenidos. Los 200 que se muestran en la Tabla 4 aportan el 96,12 por ciento del total de digramas. En dicha tabla, de cada digrama se indica su frecuencia de aparición y su versatilidad (entendida ésta como el número de palabras diferentes en las que aparece).

El número de trigramas posibles sería en nuestro caso de 54.872 (38 x 38 x 38). Sin embargo, se han encontrado solamente 5.628 trigramas diferentes, de los cuales 200 aportan el 51,87 por ciento del total de los 756.646 trigramas independizados. En cambio, hay otros 4.000 trigramas cuya aportación solamente supone el 4,32 por ciento de todos los trigramas independizados. La Tabla 5 muestra la frecuencia de aparición y versatilidad de los 200 trigramas más frecuentes.

Discusión y Conclusiones

En este estudio se ha partido de una colección de títulos grabados utilizando sólo letras mayúsculas. De haber sido escritos empleando mayúsculas y minúsculas, ¿en qué medida podrían haberse visto afectados los resultados

finales que se han mostrado? En relación con la ocurrencia global de letras, en nada, si bien, con respecto a las mayúsculas, podría comprobarse la incidencia destacada de algunas de ellas, por ser iniciales de palabras que frecuentemente aparecen al comienzo de un título o de nombres propios de uso frecuente.

En relación con digramas y trigramas, supondría la posibilidad de encontrar una mayor variedad de ellos, al poder intervenir mayúsculas y minúsculas en su composición; sin embargo, aunque las ocurrencias de parte de los digramas considerados en nuestro estudio se verían afectadas, no es probable que se introdujeran alteraciones importantes en las listas de digramas y trigramas más frecuentes (Tablas 4 y 5), si bien se podría apreciar en ellas la presencia de digramas y trigramas que comiencen con mayúsculas.

De otra parte, al tener que trabajar con títulos en mayúsculas, tampoco se ha podido obtener datos sobre la repercusión que pudieran tener los acentos y signos diacríticos que, en el caso de los acentos, sí creemos que pudiera ser significativa, ya que van ligados a letras de muy alta frecuencia, como son las vocales, y a gran número de palabras diferentes, muchas de las cuales son de frecuente aparición.

Los resultados obtenidos tienen un valor orientativo acerca de lo que pueda encontrarse al considerar todos los títulos incluidos en los 50.000 registros que en la actualidad contiene la Base de Datos de Información en Ciencia y Tecnología (ICYT). No pretenden ser extrapolables a otros tipos de textos; sin embargo, parte de dichos resultados sí pueden ser útiles para quienes trabajen en el desarrollo o explotación de bases de datos bibliográficas de cualquier tipo, y para quienes hayan de desarrollar índices permutados basados en títulos.

Tabla 4

D	F	V	D	F	V	D	F	V	D	F	V
VA	2.293	462	GU	1.375	213	S-	32.917	6.391	DI	5.793	1.054
-G	2.238	574	EA	1.372	387	-D	28.049	972	LE	5.775	1.400
6P	2.227	306	IB	1.355	243	DE	27.854	940	ID	5.680	1.494
-U	2.161	172	BA	1.339	381	E-	25.966	989	DO	5.464	1.522
BU	2.154	356	PI	1.290	471	A-	25.602	3.648	NI	5.435	1.209
PI	2.151	608	CR	1.284	390	N-	18.081	1.296	SI	5.338	1.016
CU	2.122	536	LI	1.273	439	ON	17.544	2.549	-I	5.206	724
LU	2.102	415	AG	1.243	280	-E	16.727	1.210	ME	5.093	1.044
UL	2.087	626	IP	1.231	375	EN	16.553	2.240	MA	5.015	1.063
-B	2.059	716	NU	1.220	213	ES	16.042	2.662	IS	4.992	1.016
OD	2.032	396	RU	1.217	315	O-	15.936	2.863	OL	4.943	1.248
NS	2.017	435	II	1.192	65	CI	15.833	2.094	DA	4.801	1.353
-O	1.998	448	HI	1.188	481	IO	15.708	2.038	NE	4.665	954
PL	1.968	330	UN	1.148	365	OS	14.584	2.690	EC	4.591	811
VI	1.965	446	EO	1.132	343	LA	13.925	1.532	PA	4.586	617
IV	1.959	432	RC	1.091	254	-L	13.203	632	-Y	4.508	39
ND	1.959	479	EV	1.088	139	-C	12.544	2.155	Y-	4.495	58
RM	1.942	388	FA	1.078	257	AS	12.213	2.455	-T	3.968	950
SA	1.937	586	EG	1.059	211	RA	11.372	2.222	AT	3.938	1.064
I-	1.930	246	UT	1.041	313	IC	10.915	2.775	-R	3.922	831
US	1.911	567	IG	1.025	294	CO	10.891	2.209	NO	3.920	1.072
-H	1.896	668	CH	1.001	403	-A	10.873	1.814	ET	3.898	1.030
DU	1.891	260	IT	996	215	AC	10.820	1.615	SO	3.828	574
RT	1.830	363	EX	986	260	-P	10.293	1.680	CT	3.812	672
OG	1.811	559	FL	977	157	AL	9.961	1.897	EO	3.808	623
RO	1.805	279	BL	969	224	ER	9.899	2.247	NI	3.739	1.136
NP	1.789	322	CL	942	289	L-	9.814	609	PR	3.643	597
GI	1.774	451	AR	932	50	CA	9.315	2.051	-F	3.606	829
ZA	1.760	488	BC	873	346	TE	9.223	1.693	IE	3.502	590
OT	1.735	554	EP	870	155	EL	8.733	746	OM	3.394	802
QU	1.731	373	RO	867	280	NT	8.590	1.571	IM	3.263	1.109
D-	1.723	313	NO	862	289	IN	8.433	2.029	TU	3.205	404
OC	1.715	509	RG	856	169	RE	8.337	1.561	R-	3.135	624
GE	1.694	429	IR	848	309	RI	8.318	2.039	IM	3.133	653
CC	1.691	159	XI	823	283	RO	7.865	1.818	NC	3.015	556
UD	1.642	130	GO	823	233	OR	7.856	1.669	CE	2.940	705
BI	1.574	488	NP	820	116	AR	7.539	1.691	UE	2.913	378
FE	1.556	384	EP	804	278	TO	7.323	1.575	UN	2.894	315
RR	1.521	363	BU	788	169	-S	7.149	1.132	AM	2.821	739
VE	1.512	315	IP	787	271	TA	7.038	1.676	IT	2.820	811
GA	1.509	384	DR	760	285	ST	7.028	1.128	-N	2.726	411
AP	1.489	272	BE	735	302	TI	6.921	1.850	MO	2.649	720
GR	1.480	364	MU	723	220	IA	6.740	1.433	PE	2.630	621
IZ	1.438	381	OX	696	245	AN	6.652	1.859	UR	2.622	620
UI	1.428	338	FR	690	155	NA	6.628	1.582	UC	2.558	335
UA	1.410	223	RO	675	49	LO	6.531	1.086	BR	2.423	329
FO	1.408	420	HE	670	280	AD	6.526	1.774	SE	2.411	585
AB	1.398	399	PU	659	193	TR	6.268	1.335	OB	2.355	262
VO	1.398	253	FU	648	138	LI	6.222	1.548	-V	2.334	435
OP	1.385	506	RA	632	64	-N	6.010	1.280	EM	2.322	440

Tabla 5

T	F	V	T	F	V	T	F	V
STU	1.460	24	TUR	1.058	171	NES	2.287	387
IDO	1.442	341	NAL	1.051	189	TIC	2.265	567
OBR	1.436	35	QUI	1.049	229	-MB	2.227	285
TUD	1.428	17	INT	1.047	212	ONE	2.200	354
-AC	1.425	211	ATO	1.028	308	NCI	2.172	280
UDI	1.414	30	RAS	1.027	252	-SO	2.075	134
MEN	1.411	282	TAC	1.014	171	NTO	2.047	310
ART	1.391	420	OLO	1.005	310	RIA	2.039	399
MIC	1.391	343	-MO	1.004	234	IEN	2.026	342
RIC	1.367	368	LOS	1.001	287	ALI	2.001	397
CTO	1.366	178	EEN	996	177	TOS	1.996	328
TCI	1.349	267	RIO	994	235	ALE	1.989	413
SOB	1.346	27	DA-	992	335	-PO	1.980	256
DES	1.337	287	OMP	981	133	CO-	1.979	544
AD-	1.337	231	UCI	980	69	MA-	1.947	500
MIE	1.335	167	NER	974	132	DOS	1.938	483
ACT	1.330	213	BIO	972	169	DAD	1.927	315
TEB	1.327	294	MAS	959	271	EDI	1.922	371
URA	1.327	219	LAC	957	173	TRI	1.916	465
NTR	1.281	134	CID	955	161	DIO	1.861	119
STE	1.271	228	INE	949	191	RAC	1.849	254
TAL	1.266	220	CAL	947	179	TRA	1.825	405
ANA	1.259	209	ENE	945	156	TE-	1.807	290
IZA	1.254	336	EDI	932	96	OR-	1.797	185
ILI	1.237	323	GEN	930	227	COS	1.784	515
ONT	1.237	162	AIA	929	399	-DI	1.781	398
LIC	1.215	262	MAT	923	186	RE-	1.756	53
STI	1.202	216	SPA	921	50	-SU	1.714	226
IMI	1.201	190	RAL	914	151	ENC	1.694	241
ROS	1.186	280	OGI	906	247	ECT	1.693	293
TIV	1.186	260	TEN	906	168	POR	1.680	125
IS-	1.184	299	OLI	899	333	SIS	1.646	173
STA	1.182	258	TOR	898	215	IST	1.643	242
RAN	1.176	295	ESI	895	156	ERA	1.641	352
NO-	1.176	261	UNA	894	38	CAS	1.638	436
NIC	1.171	311	-AP	893	62	CCI	1.619	129
-AN	1.170	254	ETI	889	282	TRO	1.611	336
-TR	1.168	285	II-	889	35	COM	1.605	205
GRA	1.120	260	ELE	883	209	MIM	1.577	303
ERO	1.108	321	UN-	882	8	ANT	1.576	445
EMA	1.097	100	LO-	881	149	ESP	1.575	157
CAR	1.084	283	ORP	880	171	NTA	1.560	290
ONA	1.082	381	RAT	869	177	-MA	1.551	359
FIC	1.082	274	LEC	867	178	-UN	1.537	48
RO-	1.076	211	-VI	866	222	ANI	1.521	351
TA-	1.071	277	ORE	866	190	BRE	1.495	63
CAC	1.068	102	LID	864	213	MET	1.487	371
PER	1.064	274	NOS	864	88	STR	1.480	262
-SE	1.062	255	RMI	854	68	-TE	1.468	266
ATI	1.058	348	MED	850	67	-BI	1.468	171
			-NU	849	56			

T	F	V	T	F	V
-DE	25.397	377			
DE-	20.323	38			
OS-	12.420	2.068			
ION	11.315	1.270			
ON-	10.443	998			
CIO	10.306	1.072			
AS-	9.925	1.880			
-LA	9.014	175			
ES-	8.271	1.619			
LA-	7.488	218			
ACI	7.349	881			
-EN	6.831	184			
-CO	6.034	725			
EN-	6.024	94			
EL-	5.953	45			
ICA	4.991	1.082			
ENT	4.620	764			
-ES	4.005	364			
IA-	3.654	654			
DEL	3.608	48			
CON	3.542	362			
ICC	3.504	1.063			
EST	3.368	411			
-PR	3.077	407			
LOS	2.987	140			
-EL	2.965	117			
NTE	2.938	607			
RA-	2.915	335			
-CA	2.816	547			
-IN	2.782	472			
TER	2.766	566			
-PI	2.701	299			
LAS	2.633	192			
IDA	2.603	677			
-RE	2.592	460			
ADO	2.575	821			
AL-	2.533	306			
IO-	2.512	227			
-LO	2.512	72			
PAR	2.465	193			
CA-	2.462	511			
TO-	2.396	416			
-AL	2.391	304			
ARA	2.369	210			
PRO	2.335	308			
LES	2.319	468			
INA	2.314	668			
RES	2.309	494			
DO-	2.299	517			
CIA	2.297	321			

Bibliografía

1. HELD, G.; MARSHALL, T.R. Data Compression. Techniques and Applications Hardware and Software Considerations. New York: John Wiley & Sons, 206 p., 1987.
2. CHOROS, K.; MAJEWSKA, P.; SIEMINSKI, A. Bibliographic Database Compression.Int. Forum Inf. Doc., 12 (1), 28-32, 1987.
3. REGHBATO, H.K. An overview of Data Compression Techniques. Computer, 14 (4), 71-75, 1981.
4. PEPTERSON J.L. Computer Programs for Detecting and Correcting Spelling Errors. Communications of the ACM, 23 (12), 676-687, 1980.
5. BELONOGOV, G.G.; DUGANOVA, I.S; KUZNETSOV, B.A; SMTURMAN, Ya. P.; PARTYKO, Z.V.; POZDNYAK, M.V. Automatic Error Detection and Correction in Scientific Texts. Nauch. Tekn, Inform., Ser. 2, 16 (3), 85-89, 1982.
6. ZAMORA, A. Automatic Detection and Correction of Spelling Errors in a Large Data Base. J. Amer. Soc. Inform. Sci., 31 (1), 51-57, 1980.
7. GUTIERREZ MUÑOZ, F.; REY GUERRERO, A. Estudio estadístico de palabras y caracteres en títulos de artículos científico-técnicos en español. Rev. Esp. Doc. Cient., 9 (2), 121-132, 1986.