Anatomía de los antropónimos españoles

Rafael Camps Paré
Universitat Politécnica de Catalunya (Barcelona) RCAMPS@LSI.UPC.ES

Resumen: Se presenta un estudio cuantitativo sobre morfología y léxico de los antropónimos, del que se dan algunos de los resultados obtenidos para los apellidos españoles. Se analiza su estructura, la distribución de las longitudes y frecuencias, la posición de las letras, los n-gramas, la relación entre el vocabulario y la dimensión del corpus (ley de Zipf-Mandelbrot) la entropía y el vocabulario equivalente. Se hacen comparaciones con la lengua castellana general y con los apellidos de los EUA.

1 Introducción

Entre las inmensas cantidades de datos que se manejan, consultan y almacenan en las bases de datos de los Sistemas de Información actuales, unos de los más habituales son los nombres y apellidos de personas; clientes, enfermos, contribuyentes, alumnos, autores, etc. La expansión de Internet conlleva también un uso extensivo de esas identificaciones personales o antropónimos. Sin embargo tenemos muy pocos conocimientos sobre ellos. Para el diseño de procedimientos eficaces y eficientes en el manejo de antropónimos, convendría tener un mejor conocimiento de su estructura y composición.

No tenemos noticia de ningun estudio cuantitativo relativo a los antropónimos españoles. Por ello hemos realizado un estudio experimental de algunos de sus aspectos morfológicos y léxicos. Presentamos aquí algunos de los resultados obtenidos.

En el apartado 2 presentamos el corpus estudiado. En el 3 analizamos la estructura de palabras de los antropónimos, en el 4 estudiamos su longitud, en el 5 la distribución de las letras, en el 6 las frecuencias y el vocabulario, en el 7 los bigramas y trigramas y en el apartado 8 hacemos unos comentarios finales.

2 Presentación del corpus

Para el análisis experimental presentado hemos utilizado un corpus al que denominamos ***APELLIDOS*** que contiene los apellidos de un conjunto de 1,6 millones de ciudadanos españoles. Como cada persona tiene dos apellidos, en el corpus hay 3,2 millones de apellidos, a los que llamaremos *apariciones*, aunque solo contiene 74112 apellidos distintos, a los que llamaremos *formas*. En el estudio completo se han utilizado otros corpus españoles y norteamericanos, tanto con apellidos como con nombres de pila.

El corpus *APELLIDOS* se ha sometido a un proceso de normalización que realiza algunas transformaciones previas (conversión a mayúsculas, eliminación de acentos, etc).

3 Compuestos: palabras y partículas

Los antropónimos españoles constan de tres partes; el nombre (nombre de pila) el primer apellido y el segundo apellido. Cada una de las tres partes es una cadena de caracteres que puede estar formada por varias palabras. A un nombre o apellido que consta de más de una palabra, le calificamos como **compuesto.** Entre las palabras que forman un apellido o un nombre pueden figurar **partículas** (palabras no "significantes"; *DE, LA,* etc).

En *APELLIDOS* tan solo el 0,8% de las apariciones de apellidos son compuestos. La mitad de estos contiene alguna partícula (ejemplo: *RUIZ DE LA PRADA*) y la otra mitad no contiene ninguna (ejemplo: *GOMEZ RICO*). Son poquísimos los casos que contienen mas de dos palabras, aparte de las partículas. Los apellidos con más palabras tienen cinco, de las que dos son partículas.

Las partículas, o grupos de partículas, siguientes son las mas usuales en *APELLIDOS* y representan el 82% de las apariciones de apellidos con partículas: *DE, SAN, EL, LA, DE LA, DEL, DA, DOS, SANTA.* En otros corpus con apellidos españoles hemos encontrado resultados similares a estos.

En el caso de nombres de pila los compuestos son mucho mas frecuentes pero sin partículas.

4 Longitudes

En este apartado presentamos los resultados obtenidos en el análisis de la longitud de los apellidos.

A partir de ahora llamaremos **vocabulario** de un corpus de antropónimos, al conjunto de antropónimos diferentes que en él existen , o sea el conjunto de formas.

La longitud media de las apariciones de apellidos con compuestos, en *APELLIDOS*, es 6,37 y la de las sin compuestos es 6,32. La longitud media de las formas (vocabulario) con compuestos es 7,45 y sin compuestos es 6,93.

El apellido más largo tiene 28 caracteres, pero es un apellido compuesto. El simple más largo tiene 16.

Veamos en la figura 1 la distribución de las longitudes. Analizando otros corpus españoles hemos encontrado frecuencias muy parecidas a estas y curiosamente también muy parecidas a las de corpus con apellidos norteamericanos. En todos los casos las longitudes entre 4 y 9, ambas incluidas, suponen más del 90% del corpus.

La longitud media del vocabulario es algo superior a la del corpus, pero la diferencia no es tan grande como la que se encuentra en el uso del lenguaje natural. Así, en la lengua castellana la longitud media de las palabras es 4,62 mientras que la de su vocabulario es casi el doble, 8,9 [Ala95].

En la figura 2, relacionamos la longitud con la frecuencia media de los apellidos de esa longitud. Así por ejemplo el número medio de personas de los apellidos de longitud 6 es 50. Vemos que para los apellidos de longitud mayor que 3, la frecuencia media disminuye a medida que aumenta la longitud. Eso ocurre a pesar de que la longitud media es 6,37 y los 10 apellidos mas frecuentes tienen longitudes entre los 5 y los 9 caracteres .

Figura 1
Longitudes en *APELLIDOS*

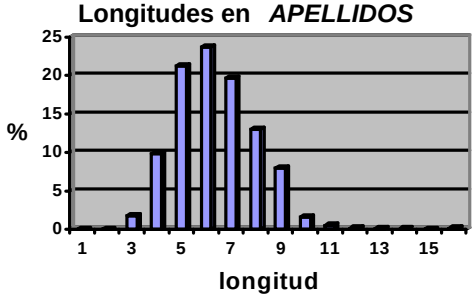
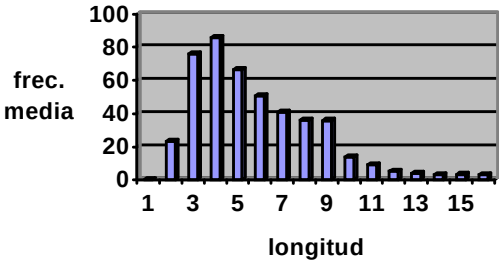


Figura 2
Frecuencia media vs Longitud



5 Letras

5.1 Frecuencia de las letras

Los componentes gráficos más elementales de un antropónimo son los caracteres, símbolos, que lo forman.

En la tabla1 hemos representado las frecuencias relativas, en porcentaje, de cada letra respecto al total de caracteres de las apariciones de apellidos en el corpus.

Para tener un punto de referencia hemos puesto también una columna con las frecuencias en los textos en lengua española [Ala95]. Vemos que en general las letras de *APELLIDOS* tienen frecuencias similares a las del lenguaje natural. Hay que hacer notar que en nuestro corpus, la grafía de los antropónimos vascos está muy castellanizada, por lo que la mayor parte de las *K* aparecen como *C* o *QU*, y la mayor parte de las *TX* aparecen como *CH*.

Si ordenamos las letras en orden decreciente de su frecuencia relativa en las apariciones en *APELLIDOS*, obtenemos la lista siguiente:

A E R O L N I S Z C T M U G D B
P V H F J Ñ Y Q X K W

Las 10 letras más frecuentes suponen casi el 80% de las apariciones de letras. El conjunto de las vocales supone el 43% (el 46,5% en la lengua general).

Podría influir en estos resultados la distribución, tan poco uniforme, de las frecuencias de los apellidos. Por ejemplo ¿ocupa la *R* un lugar destacado porque *RODRIGUEZ* es muy frecuente? Para analizarlo hemos estudiado la frecuencia de las letras en las formas del

vocabulario y vemos que el resultado es practicamente el mismo que con las apariciones

Letra	*APELLIDOS*	*Castellano*
A	14,8	12,8
B	2,0	1,5
C	3,7	4,4
D	2,6	4,9
E	10,6	13,2
F	0,9	0,67
G	3,0	1,1
H	1,0	0,88
I	5,8	6,7
J	0,6	0,43
K	0,04	0,02
L	6,7	5,6
M	3,3	2,9
N	5,9	7
Ñ	0,4	0,19
O	8,3	9,5
P	2,0	2,5
Q	0,3	1,1
R	10,5	6,5
S	5,3	7,5
T	3,4	4,3
U	3,1	1,92
V	1,5	1
W	0,02	0,01
X	0,1	0,15
Y	0,4	0,9
Z	3,8	0,39

Tabla 1: Frecuencias relativas en %

en el corpus. Las únicas excepciones notables son las siguientes:

Letra	**apariciones**	**formas**
Z	3,8%	1,7%
H	1,0%	2,0%
K	0,04%	0,7%
W	0,02%	0,2%

En *APELLIDOS*, las formas que contienen *H,K* o *W* son formas de frecuencias bajas, pero las que contienen *Z* suelen ser de frecuencias altas. Comparando los resultados de varios corpus de antropónimos españoles y norteamericanos, no se encuentran grandes diferencias. Las letras que ocupan los tres primeros lugares en el ranking de frecuencias, son siempre *A E* y *R* (en la lengua castellana la *R* ocupa tan solo el séptimo lugar). Entre el lugar 4 a 9 están siempre las cinco letras siguientes; *I O L N S*. Entre el lugar 8 al 15 aparecen siempre las letras *C D M T*. Las letras *B G U* aparecen entre los lugares 12 a 20. Entre las menos frecuentes (las situadas entre el lugar 16 al 27) aparecen siempre las letras *F K Ñ P Q V W X*. Las restantes letras, *H, Y, J* y *Z,* tienen distinto comportamiento segun los corpus.

5.2 Posiciones y letras

Acabamos de estudiar las frecuencias de las letras en los antropónimos. Pero las letras no se distribuyen uniformemente en las distintas posiciones. Así por ejemplo aunque la letra A es la más frecuente, está muy lejos de ser la inicial más frecuente. En este apartado estudiamos la distinta distribución de las letras segun su posición.

A continuación indicamos las letras más frecuentes en las primeras y últimas posiciones de los apellidos de nuestro corpus. Es notable la discrepancia entre las letras mas frecuentes en la inicial de los antropónimos y las mas frecuentes en general. Obsérvese que las cinco letras mas frecuentes como letra inicial, ocupan en la clasificación general por frecuencias los lugares 12,14,10 y 17.

Inicial	**M**, G, C, P, S, R, B
2ª	**A**, O, E, I, U
3ª	**R**, L, N
penúlt	**E**, A, R
última	**A**, Z, O, S

El 88,5% de los apellidos españoles tienen como inicial una consonante (la M lo es en el 11,2%) y el 83% tiene una vocal como segunda letra (la A lo es en el 31%). Es notable el hecho de que en la tercera posición la letra R sea la más frecuente tanto en España como en EUA. Las letras menos frecuentes como inicial de apellidos españoles son Ñ, W, X, K, Y y U. Las menos frecuentes como final son Ñ, Q, V, W, B y J. Hemos medido también la posición media de cada letra. Las letras con posición media más baja , o sea aquellas que cuando aparecen suelen hacerlo hacia el inicio del antropónimo, son (entre paréntesis aparece la posición media):

F (1,5), P (1,8), M (2,1), B (2,3).

Las letras de posición media mas alta, o sea aquellas que cuando aparecen suelen hacerlo hacia el final del antropónimo, son:

Z (6,1), N (4,6), E (4,6), D (4,5) .

Los resultados vistos hasta aquí en relación a las posiciones de las letras, han sido obtenidos con las apariciones en *APELLIDOS.* Pero si se cuentan solo las formas, o sea el vocabulario, se obtienen resultados parecidos excepto para la letra Z, ya que los apellidos acabados en Z aunque son frecuentísimos son muy pocos.

5.3 Alfabeto equivalente y entropía

El alfabeto que estamos usando está formado por $\alpha = 27$ letras y como acabamos de ver está muy lejos de ser utilizado de forma equiprobable, no lo utilizamos óptimamente. Dicho de otra forma, la **entropía** $H(\alpha)$ de nuestro corpus considerado como fuente de información con un alfabeto de α símbolos, es inferior a la entropía máxima $H_{max}(\alpha)$.

Veámoslo cuantivamente. El valor máximo de la entropía, para una fuente de α letras es $\log_2\alpha$. En nuestro caso será 4,75 bits. La expresión de Shannon para la entropía, medida en bits, es:

$$H(\alpha) = - \sum_{i = 1, \alpha} (p_i \log_2 p_i)$$

en donde p_i es la probabilidad de cada letra, medida como su frecuencia en tanto por uno. El valor de la entropía será

$H(\alpha) = 4{,}06$ bits.

Hemos calculado la entropía en el caso del castellano a partir de las frecuencias dadas en [Ala95] y ha resultado ser 3,93 bits, idéntica a los corpus con nombres de pila, pero menor que en los apellidos. La mayor entropía de los apellidos quizás se pueda explicar por la necesidad de redundancia para compensar la falta de contexto.

Para "visualizar" mejor la pérdida de rendimiento, uso desequilibrado de las letras del alfabeto, podemos imaginar un **alfabeto equivalente** con α_e símbolos, que nos diera la misma entropía siendo utilizado de forma óptima, o sea que sus símbolos se usaran equiprobablemente. El número de símbolos, α_e, de ese alfabeto equivalente, se calculará así:

$$\alpha_e = 2^{H(\alpha)} \text{ letras}$$

Mostramos a continuación ese valor para la letra inicial, la final, y para el caso general. Como vemos, la posición final en los apellidos españoles, tiene un alfabeto equivalente a tan solo 9 letras:

	Alfabeto equivalente α_e		
α	general	inicial	final
27	16,7	16,11	9,06

6 Frecuencia y vocabulario

El espectacular crecimiento del correo electrónico, buscadores de personas, etc, ha disparado el interés por como construir herramientas de búsqueda eficaces y eficientes. Es importante a este respecto conocer las distribuciones de frecuencias de los antropónimos.

El **vocabulario** del corpus *APELLIDOS* solo tiene 74112 formas. La frecuencia media es pues 43, pero su distribución es muy poco uniforme (su desviación tipo es 654). Los apellidos de frecuencia entre 1 y 10 suponen tan solo el 4,6% del corpus, pero sin embargo suponen el 75% del vocabulario. Por otro lado un solo apellido, el más frecuente (*GARCIA*) supone el 2,6% del corpus. En las tablas 2 y 3 mostramos cifras de esa distribución tan poco uniforme.

Conviene advertir que, como veremos mas adelante, alrededor de la mitad de los apellidos que aparecen en los vocabularios obtenidos de ficheros de los Sistemas de Información habituales, suelen ser erróneos. Así es posible que en el conjunto de personas registradas en el fichero, el número de apellidos realmente diferentes no sea superior a 25000. Pero aquí nos interesa precisamente el contenido de los ficheros reales incluyendo las anomalías. En los Sistemas de Información españoles se suele considerar que el número de personas con errores en nombres o apellidos oscila entre el 5% y el 15% en los ficheros habituales.

En la tabla 2 se puede leer que los 10 apellidos mas frecuentes suponen tan solo el 0,013% de los apellidos del vocabulario pero el 13% del corpus. Vemos que con los 34000 apellidos mas frecuentes cubrimos el 98% del corpus. El apellido que ocupa la posición 34000 en un orden descendente de frecuencias, tiene una frecuencia bajísima, 3.

Sin embargo aun quedan más de 40000 apellidos en el vocabulario con frecuencias más bajas. En concreto hay un 36,6% que tiene una frecuencia igual a 1.

N apellidos	%vocab	%corpus
1	0,0013	2,64
10	0,013	13
24	0,032	20,9
100	0,13	33,2
200	0,26	41,1
700	0,94	58
1000	1,35	62,9
2000	2,7	72,4
3500	4,7	80
10000	13,5	90,9
34000	45,8	98,2

Tabla 2: Los N apellidos más frecuentes

Esta proporción tan alta de apellidos únicos es un fenómeno habitual y no exclusivo de los Sistemas de Información españoles. Por ejemplo en el fichero básico de la Seguridad Social de los EUA, son únicos el 33% de los apellidos.

En la tabla 3 damos algunas cifras sobre los apellidos de frecuencias bajas. Hay una cierta coincidencia entre los responsables de los ficheros de población (situados habitualmente entre 10^6 y 10^8 personas) en considerar que la inmensa mayoría de los apellidos con frecuencia igual o inferior a 5 son erróneos. El 65% del diccionario de *APELLIDOS* tiene una frecuencia igual a 5 o inferior. Además puede haber muchos apellidos erróneos que aparezcan con frecuencias muy superiores a 5, ya sea porque el error se ha producido muchas veces, ya sea porque el erróneo coincide con un apellido correcto.

Cuanto mayor sea la calidad de los datos de un fichero, menor será la presencia de apellidos con frecuencias muy bajas. Sin embargo las cifras que hemos encontrado para los diversos corpus españoles, de orígenes muy distintos, son parecidas.

	%voc	%corp
frec < 11	75	1,75
frec = 1	36,6	0,85

frec. med.=43
long. corpus = 3Maps
long .voc.= 74,1Kaps (2,3%)

Tabla 3: Apellidos poco frecuentes

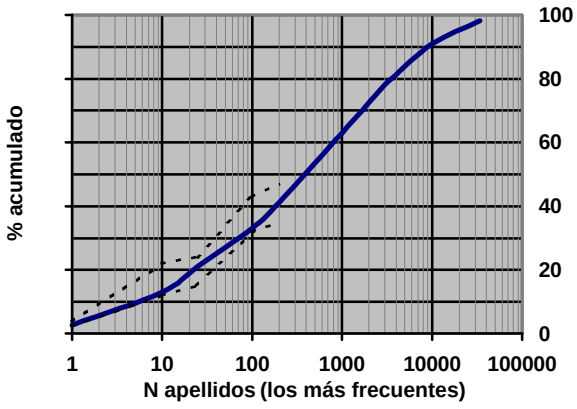
Los 10 apellidos más frecuentes en *APELLIDOS* suponen el 13% de las apariciones. Veamos cuales son y su frecuencia relativa :

	fr%
GARCIA	2,64
MARTINEZ	1,86
LOPEZ	1,79
FERNANDEZ	1,49
PEREZ	1,46
RODRIGUEZ	1,42
GONZALEZ	1,38
SANCHEZ	1,35
GOMEZ	0,85
MARTIN	0,72

En todos los ficheros españoles que hemos estudiado, se cumple lo siguiente: GARCIA es el más frecuente, LOPEZ aparece en los puestos 2º o 3º. MARTINEZ se mueve entre el 2º y el 8º. De FERNANDEZ a SANCHEZ se encuentran entre los puestos 3º a 8º. GOMEZ y MARTIN se disputan el 9º y 10º. Los puestos 11º y 12º suelen estar ocupados por RUIZ y HERNANDEZ, aunque estos dos apellidos se disputan los puestos 11º a 18º con los siguientes apellidos; DIAZ, ALVAREZ, JIMENEZ, NAVARRO, MUÑOZ, MORENO y ROMERO.

Un estudio de los apellidos de los ciudadanos norteamericanos de origen hispano [USC96] obtenidos a partir de ficheros del censo, da casi los mismos resultados en cuanto a los 10 apellidos más frecuentes, con la única diferencia importante que HERNANDEZ ocupa la

Figura 3:
Frecuencia acumulada en función de los N apellidos más frecuentes



posición 5, y entonces FERNANDEZ ocupa la 29. Además el apellido MARTIN no aparece en la lista, porque pronunciado con acento grave es un apellido raro entre hispanos pero muy frecuente entre los no hispanos y como en los ficheros del censo no se tiene en cuenta el acento, decidieron no incluir MARTIN en la lista de apellidos hispanos.

En la figura 3 representamos las frecuencias acumuladas para los N apellidos más frecuentes en *APELLIDOS*. Las lineas de puntos son los limites encontrados en otros corpus españoles. El número N se ha representado en una escala logarítmica. Por ejemplo segun el gráfico, los 200 apellidos más frecuentes suponen el 41%.

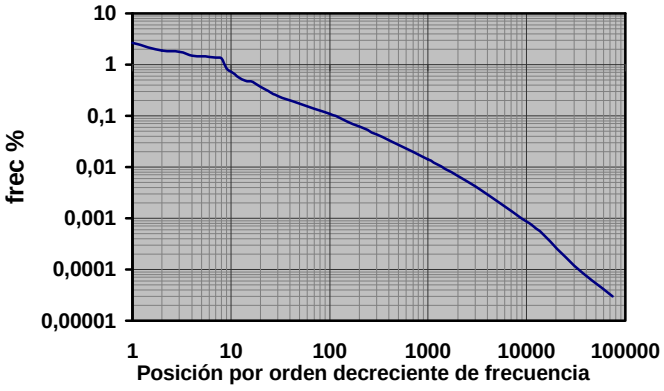
En la figura 4 representamos la frecuencia de cada apellido del corpus en función de su posición en un vocabulario ordenado por frecuencia decreciente. Hemos utilizado escalas logarítmicas en los dos ejes. Así por ejemplo para el apellido que ocupa la posición 10, (que es MARTIN) la figura nos indica una frecuencia del 0,7%.

En el análisis de textos en lenguaje natural se suele aceptar que la frecuencia, ***f***, de una palabra en un texto se puede expresar en función de su posición o rango ***r***, en un diccionario ordenado por ***f*** decreciente, de acuerdo con una ley de Zipf-Mandelbrot [Fair69]:

$$f = K_1 (r + K_2)^{-o}$$

La constante ***o*** suele ser mayor que 1 para los lenguajes naturales, y menor que 1 para lenguajes muy reglados. Hemos ajustado una

Figura 4:
Apellidos: Frecuencia en función de la posición



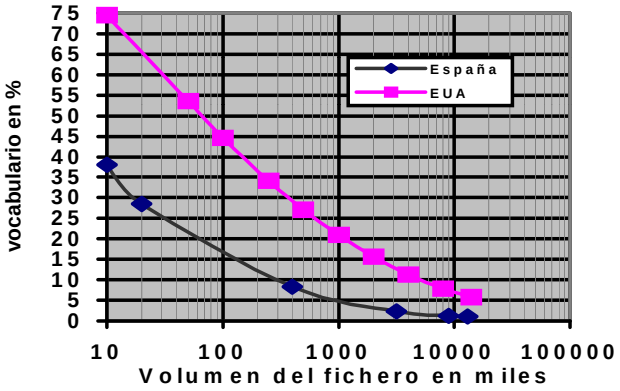
distribución de este tipo a las frecuencias de *APELLIDOS* y hemos obtenido la expresión siguiente:

$$f = N \ 0{,}061 \ (r + 1{,}2)^{-0{,}95}$$

siendo ***N*** = longitud del corpus.

En la figura 5 representamos gráficamente la relación entre el volumen de un fichero y el de su vocabulario. Hemos representado dos curvas; una para apellidos españoles y otra para norteamericanos. Para los apellidos españoles hemos utilizado además de *APELLIDOS* otros corpus de diversos tamaños. Para los apellidos

Figura 5:
Apellidos: Volumen del vocabulario



norteamericanos hemos utilizado los datos que aparecen en [Hil96] y que corresponden a un conjunto de 14 millones de personas, del que hizo subconjuntos de diversos tamaños.

En el eje X, con escala logarítmica, aparece el volumen del corpus medido en miles de apellidos. Por lo tanto, en el caso español será aproximadamente el doble del número de miles de personas. En el eje Y aparece el volumen del vocabulario en % del volumen del corpus. Por ejemplo según el gráfico, un fichero de 500000 españoles tendrá aproximadamente 25000 apellidos diferentes ya que contiene 1 millon de apellidos y para ese volumen le corresponde un vocabulario con un volumen del 5% del volumen del fichero.

Los apellidos en EUA son más diversos, seguramente por el origen inmigratorio reciente y multicultural de su población.

En el mundo de la estadística aplicada a textos en lenguaje natural, se utiliza la siguiente función (a veces considerada una de las leyes de Zipf) para estimar la longitud, ***v***, de un vocabulario (número de palabras distintas) en función del número ***n*** de palabras del texto [Cro99]:

$$v = k \ n^{\beta} \qquad k > 0 \ , \ 0 < \beta < 1$$

Para textos en inglés los parámetros k y β suelen ser $10 < k < 20$, $0{,}4 < \beta < 0{,}6$

Ajustando esta función a los corpus de apellidos españoles hemos obtenido la siguiente expresión:

$$v = 29 \ n^{0{,}532}$$

En el apartado 5.3 hemos calculado la entropía considerando que el alfabeto son las letras. Pero

podemos calcular ahora la entropía de *APELLIDOS* considerando que los símbolos generados por la fuente de información no son letras sino antropónimos, o sea que el alfabeto es el vocabulario. Veamos los resultados:

α	$H_{max}(\alpha)$	$H(\alpha)$	$Hr(\alpha)$	α_e
74112 aps	16,1 bits	11,6 bits	0,72	3104 aps

Ahora el alfabeto equivalente podría ser llamado vocabulario equivalente. Vemos que si utilizáramos equiprobablemente los apellidos, con tan solo 3104 apellidos en lugar de 74112, la fuente de información *APELLIDOS* tendría la misma incertidumbre o entropía.

7 Bigramas y trigramas

En este estudio morfológico de los antropónimos españoles, hemos analizado el nivel *letra* y el nivel *cadena*. Analizaremos ahora estructuras morfológicas intermedias, los pares y trios de letras, ***bigramas*** y ***trigramas***.

Bigramas	%	Trigramas	%
AR	3,2	MAR	1,02
EZ	3,1	ART	0,85
ER	3	ARC	0,77
AN	2,9	GAR	0,77
AL	2,4	RTI	0,76
RA	2	CIA	0,72
RE	1,9	ARR	0,69
RO	1,7	GUE	0,69
LA	1,6	SAN	0,67
AS	1,6	AND	0,67

Tabla 4: Los 10 bigramas y trigramas más frecuentes

Una cadena de *L* caracteres de longitud, tendrá *L*-1 bigramas y *L*-2 trigramas. El corpus *APELLIDOS* contiene 3191039 apariciones de apellidos, con una longitud media de 6,37 caracteres. Por lo tanto contiene aproximadamente 17 millones de bigramas y 14 millones de trigramas.

La longitud máxima del vocabulario de bigramas, número máximo de diferentes parejas de símbolos, con un alfabeto de α símbolos, será α^2. Y el número máximo de trigramas será α^3. Por lo tanto para α = 27 tendremos 729 bigramas y 19683 trigramas diferentes como máximo.

En la tabla 4 se muestran los 10 bigramas y trigramas más frecuentes. Las frecuencias no se han obtenido del vocabulario sino de las apariciones en el corpus.

La lista de los bigramas mas frecuentes no es muy distinta entre nombres y apellidos, pero hay un caso singular, el bigrama *EZ* que es la terminación de 8 de los 10 apellidos más frecuentes, pero que es poco frecuente en nombres. Los trigramas más frecuentes son realmente muy distintos entre nombres y apellidos y entre estos y el lenguaje natural.

En la tabla 5 damos algunos de los resultados numéricos obtenidos del análisis de los bigramas y trigramas en *APELLIDOS*: el número máximo de n-gramas posibles, el número de los encontrados, el % que suponen los más frecuentes y el número de n-gramas que aparecen una sola vez (con toda probalilidad son erróneos). Además, como ya hemos hecho para las letras (apartado 5.3) calculamos la entropía y el alfabeto equivalente, pero ahora el alfabeto no serán letras sino bigramas o trigramas.

	bigramas	trigramas
posibles	729	19683
existentes (difere)	655	7387
el más frecuente	3,28%	1,02%
los 10 más frecue	23%	7,6%
los 100 " "	84,4%	42,2%
los 1000 " "	100%	91,7%
Frecuencia = 1	100	893
Entropía $H(\alpha)$	7,1 bits	9,6 bits
Alfab. equivl. α_e	136 bigrs.	797 trigrs.

Tabla 5: Cifras sobre bigramas y trigramas

El porcentaje de trigramas existentes en *APELLIDOS*, respecto a los posibles, es del 37%, muy superior a lo que suele ocurrir en textos en lenguaje natural, donde en la mayoría de las lenguas no se sobrepasa el 4%.

8 Comentarios finales

El objetivo básico del trabajo que hemos presentado, es obtener datos y funciones que ayuden en trabajos de investigación sobre métodos de búsqueda aproximada de personas en bases de datos. Aunque las únicas conclusiones que se persiguen son esas cifras y funciones, de los resultados se deduce que los antropónimos, especialmente los apellidos, tienen ciertas características que les diferencian del resto del lenguaje natural: mayor entropía, formas de frecuencias muy bajas, mucha mayor diversidad de trigramas, etc. Las razones hay que buscarlas

seguramente en la diversidad de los orígenes de los antropónimos, tanto en el espacio como en el tiempo. Además, en los apellidos apenas tiene repercusión la evolución de la lengua. Los apellidos son fósiles del lenguaje. Los nombres son más cercanos al lenguaje natural, seguramente porque no vienen dados por via hereditaria, son más libres y de uso más familiar y cotidiano.

Referencias

[Ala95] Alameda, J.R.: *Diccionario de frecuencias de las unidades lingüísticas del castellano*, 2 vols. Universidad de Oviedo, 1995.

[Cro99] Croft, B.: Text Based Information Systems. *http://ciir.cs.umass.edu/cmpsci646/*

[Fair69] Fairthhorne, R.A.: Empirical Hyperbolic Distributions (Bradford-Zipf-Mandelbrot) for Bibliometric Description and Prediction. *Journal of Documentation* vol 25, nº4 Diciembre 1969.

[Hil96] Hild, H y Waibel, A.: Recognition of Spelled Names Over the Telephone. *Proceed. ICSLP96*, vol 1.

[USC96] Word, D.L. y Perkins, R.C.: Building a Spanish Surname List for the 1990's- A New Approach to an Old Problem. *Technical working paper no.13* March 1996. US Census Bureau