**LSI-02-61-R**

# ANÁLISIS DE LOS ERRORES QUE SE COMETEN EN LOS ANTROPÓNIMOS

Rafael Camps Paré
Departamento LSI
Universitat Politècnica de Catalunya
Barcelona

1.- Introducción

2.- Operaciones de edición y distancia de

3.- *Corpus*

4.- Normalización, partículas y estructura

5.- Distribución de frecuencias de la distancia de edición
    5.1.- Apellidos
    5.2.- Nombres

6.- Distribución de las operaciones para $\delta = 1$ : Las parejas-uno
    6.1.- Los tipos $I, D$ y $S$
    6.2.- Letras afectadas y tendencia al error
        6.2.1.- Inserciones y supresiones
        6.2.2.- Substituciones
        6.2.3.- Operaciones de edición en los apellidos norteamericanos

7.- Distribución de las operaciones para $\delta > 1$
    7.1.- Distancia de edición, $\delta = 2$: Operaciones no básicas
    7.2.- Distancias superiores ($\delta > 2$)

8.- Posición y tipo de operación

9.- Longitudes y distancia de edición

10.- Confusiones

## 1.- Introducción

Estamos investigando métodos automáticos apropiados para acceder a la información de una persona, dentro de una BD de un SI, conociendo el antropónimo aunque este conste con deformaciones o variantes. Así por ejemplo, si buscamos en una BD la información de una persona de la que creemos que sus apellidos son *Blasco Ximenis*, puede ocurrir que en la BD esa persona figure como *VELASCO GIMENEZ*. Nos interesará pues, que el sistema informático reconozca que *Blasco Ximenis* y *VELASCO GIMENEZ* son probablemente deformaciones, variantes o errores uno de otro. Para esa investigación será indispensable disponer de datos reales sobre errores y deformaciones en las BD de los SI. Hemos analizado un *corpus* de errores real, obteniendo los resultados que presentamos en este informe.

En la sección 2, recordamos los conceptos *operación de edición* y *distancia de edición* que utilizaremos para el análisis de los errores o diferencias. En la sección 3 presentamos los ficheros que forman el *corpus* estudiado. En el 4 hablamos de la normalización de formatos y grafías. Y en las secciones 5 a 10 mostramos los resultados del análisis. En concreto, veremos como se distribuyen las operaciones de edición según las distancias de edición, las letras afectadas, su posición, la influencia de la longitud en los errores y por último la posibilidad de confusiones entre antropónimos.

## 2.- Operaciones de edición y distancia de edición

Se llaman *operaciones de edición* las operaciones de manipulación de caracteres, que permiten transformar una cadena en otra. Se suelen considerar tres tipos de operaciones; *inserción* de un carácter, *eliminación* de un carácter y *substitución* de un carácter por otro. Aquí notaremos esos tipos de operaciones como `I` , `D` y `S` respectivamente.

Para explicar los resultados del análisis en términos de operaciones de edición, usaremos la siguiente notación. Para referenciar un carácter genérico (una variable del tipo caracter y de longitud = 1) utilizaremos los signos *a, b, c, ..* Los caracteres propiamente dichos, literales, apareceran en mayúsculas. La inserción de *a*, la notaremos como `I` *(a)*. La eliminación de *a* la notaremos como `D` *(a)*. La substitución de *a* por *b*, la notaremos como `S` *(a,b)*. Por ejemplo, la transformación de la cadena *GIMENEZ* en la cadena *JIMNEEZ* se puede hacer con tres operaciones de edición; una `I`, una `D`, y una `S`. Una secuencia de operaciones posible (no la única) sería la siguiente:

*GIMENEZ* → *JIMNEEZ*

1º `S` *(G,J)*
2º `D` *(E)*
3º `I` *(E)*

En esta notación no indicamos la posición en la que se realiza la operación. Así por ejemplo, la `D` *(E)* del ejemplo anterior no nos dice a cual de las dos *E* se refiere. Y la `I` *(E)* no indica en que posición se realiza la inserción de la *E.* En este trabajo no necesitaremos utilizar una notación más detallada. Si hubiéramos escogido una notación en la que se indicara la posición, en el caso de las supresiones no haría falta citar el símbolo que se suprime, pero aquí nos interesará saber cual es el símbolo suprimido.

Para un alfabeto de $\alpha$ símbolos (o caracteres), el número de operaciones distintas posibles será $\alpha(\alpha+1)$ ya que hay $\alpha$ operaciones distintas del tipo `I` *(a),* $\alpha$ del tipo `D` *(a)* y $\alpha^2-\alpha$ del tipo `S` *(a,b).* El número de operaciones posibles para $\alpha$=32, será por lo tanto 1056, y para $\alpha$=27 será 702.

La transformación de una cadena en otra puede ser hecha, en general, por más de una secuencia de operaciones de edición. Incluso puede haber varias secuencias válidas con un mismo número de operaciones. Por ejemplo, en *GIMENEZ* → *JIMNEEZ* también sería válida la siguiente secuencia de operaciones:

`S` *(G,J)*
`S` *(E,N)* esta operación es en la posición 4
`S` *(N,E)* esta operación es en la posición 5

Se llama *distancia de edición* entre dos cadenas, al número mínimo de operaciones de edición necesario para transformar una de las cadenas en la otra [Nav01]. En LSI-02-6-R hay una definición más formal. En nuestro ejemplo, la distancia de edición valdría 3 y la podríamos expresar así :

$$\delta(\textit{GIMENEZ}, \textit{JIMNEEZ}) = 3$$

Mientras no digamos lo contrario, supondremos que las operaciones son solo de los tres tipos básicos que acabamos de describir, `I`, `D` y `S`. No se acepta más de una operación en una misma posición: en nuestro ejemplo anterior, el de las tres substituciones, la tercera substitución no se hace sobre la misma posición en la que se ha hecho la segunda.

Así definida la distancia de edición, todas las operaciones tienen el mismo peso o coste: la unidad. En realidad esta no es más que la forma más simple de distancia de edición., es la distancia más estudiada y más utilizada tanto en aplicaciones textuales como en biología molecular y otras aplicaciones. Pero en ciertos casos interesará utilizar distancias con costes diferentes, como se puede ver en LSI-02-6-R.

También se han propuesto distancias de edición con tipos de operaciones diferentes a los tres considerados básicos `I`, `D` y `S`. Es posible imaginar un sinnúmero de tipos distintos de operaciones. Por ejemplo, substitución de dos caracteres contiguos por otro, desplazamiento de un carácter dos posiciones hacia la izquierda, etc. En nuestro estudio consideraremos también las operaciones de *intercambio* de dos caracteres en posiciones no contiguas, `X` *(a,b)*, y las operaciones de *transposición,* `T` *(a,b)*, equivalentes a las de intercambio pero de caracteres en posiciones contiguas. Si añadiéramos estos dos tipos de operaciones, `X` y `T`, al conjunto de operaciones básicas para el cálculo de la distancia, en nuestro ejemplo tendríamos:

$$\delta(\textit{GIMENEZ}, \textit{JIMNEEZ}) = 2$$

ya que la secuencia mínima constaría de las dos operaciones siguientes del nuevo juego de 5 tipos de operaciones ( `I`, `D`, `S`, `X`, `T` ):

`S` *(G,J)*
`T` *(E,N)*

Pero en este estudio del corpus de errores no añadiremos los tipos `X` y `T` al conjunto básico, solo estudiaremos su frecuencia (ver 7.1).

En las aplicaciones en las que se pretende la corrección automática de errores (por ejemplo en los correctores ortográficos) una de las cadenas es la "correcta" y la otra es la "errónea" o variante. Entonces se considera que la transformación va en el sentido de la cadena variante a la correcta. En otros trabajos, como en este, las operaciones de edición se definen en sentido contrario; transforman la cadena correcta en la variante. En ciertas aplicaciones, como por ejemplo en la mayor parte de las relacionadas con apellidos, no se puede hablar de cadena correcta y cadena errónea. En general cualquier cadena puede ser un apellido correcto. Precisamente uno de los orígenes de nuevos apellidos es el error en la inscripción en el registro de nacimientos, pues la versión errónea se convierte en un nuevo apellido "correcto". Por otra parte en muchos casos se considerarán correctas distintas grafías de un mismo antropónimo. En esas aplicaciones el orden de la pareja de cadenas es indiferente, por lo que la operación `I` *(a)* sería equivalente a la `D` *(a)* , y la operación `S` *(a,b)* sería equivalente a la `S` *(b,a)*. Cuando aquí estemos asumiendo esa simetría, utilizaremos solo los dos tipos de *operaciones simétricas* siguientes;

`ID` *(a)* que equivale indistintamente a una `I` *(a)* o una `D` *(a)*

`SS` *(a,b)* que equivale tanto a una `S` *(a,b)* como a una `S` *(b,a)*

En la `SS` *(a,b)*, pondremos la *a* y la *b* en el orden lexicográfico. Así por ejemplo, la pareja *GIMENEZ / GIMENNES* a la que le corresponden las operaciones `I` *(N)* y `S` *(Z,S)*, también le podremos asociar las operaciones simétricas `ID` *(N)* y `SS` *(S,Z)*. En cualquier caso la distancia de edición sigue siendo $\delta$ (*GIMENEZ , GIMENNES)* = 2.

El número de operaciones simétricas distintas será la mitad del de no simétricas. Por lo tanto, para $\alpha$ = 32 habría 528 operaciones, `ID` *(a)* y `SS` *(a,b)*, posibles. Para $\alpha$ = 32 habría 351.

## 3.- *Corpus*

Para el análisis experimental de errores, variantes o anomalias en los antropónimos, es necesario disponer de un *corpus* de datos reales. Para el caso de errores en el lenguaje general, existen algunos *corpus* de libre uso, pero para los nombres o apellidos no existe prácticamente nada y mucho menos para el contexto español. Es extremadamente difícil conseguir *corpus* representativos de los errores existentes en las BD de los SI de las empresas o administraciones públicas. Las razones son múltiples y van desde la confidencialidad hasta el posible desprestigio por la difusión de su origen (son frecuentes las empresas e instituciones que tienen en sus BD más de un 20% de antropónimos erróneos).

En este análisis hemos utilizado como *corpus*, diversos ficheros entre los que destacan los tres que designamos con los nombres *MARKET*, *UNIVERS* y *CENSUSA*.

El fichero *MARKET* se ha obtenido como fruto de los trabajos de depuración de errores en una BD de una empresa de marketing por correo, con 2,5 millones de personas residentes en España. El fichero *MARKET* contiene básicamente errores encontrados durante los procesos de detección de personas duplicadas. Sus registros constan de seis campos: tres para el nombre y los dos apellidos antes de las correcciones, y tres con las versiones corregidas. Del fichero hemos obtenido 11610 parejas (antes/después de la corrección) de apellidos y 5437 parejas de nombres de pila.

El fichero *UNIVERS* se ha obtenido pidiendo a unos 3000 profesores de diversas universidades españolas, todos aquellos errores, variantes o anomalias, que recordaran se hubieran cometido con sus nombres o apellidos. Tan solo respondieron 360 personas, aportando en total 1074 parejas de apellidos y 306 de nombres. Las dimensiones de este *corpus* son suficientemente grandes para poder considerar fiables los resultados de su análisis.

No existen en la literatura (o no los hemos sabido encontrar) trabajos similares al nuestro, para así poder comparar los resultados obtenidos con los antropónimos de otros contextos culturales. Pero hemos encontrado 1300 parejas de apellidos (75 son de origen español) y 3300 de nombres, utilizadas por el *Bureau of the Census* de los EUA para hacer pruebas de *matching* aproximado [Win95]. A ese fichero le llamaremos *CENSUSA*.

Además de esos ficheros, hemos utilizado algún que otro fichero especial para cuestiones específicas puntuales como veremos en su momento.

## 4.- Normalización, partículas y estructura

En nuestro *corpus* original, las tres partes de un antropónimo español completo (nombre de pila, primer apellido y segundo apellido) están separadas en campos diferentes. Dentro de cada campo podrá haber varias palabras ya que, aparte de los errores, existen nombres y apellidos que están compuestos por varios elementos.

Habitualmente se dispone de tres campos, como ocurre en nuestro caso. Pero en algunas BD, los antropónimos se almacenan en un solo campo o en dos. Entonces puede haber o no, características que permitan reconocer cada parte. Por ejemplo cuando todo se incluye en un solo campo, es frecuente que se siga la norma de poner los apellidos en primer término y el nombre al final precedido de una coma y un espacio. Pero ese tipo de normas pueden no aplicarse o aplicarse mal. Como en nuestro caso todos los ficheros tienen campos separados, no disponemos de cifras sobre los errores de estructura que se producen en los campos compartidos. En lo que sigue supondremos campos separados.

Las empresas acostumbran a someter sus datos a procesos de normalización, pero en cualquier caso, en el momento de la comparación de dos términos puede ocurrir que uno de ellos, o los dos, no esté normalizado. Por ello los procedimientos de comparación de nombres o apellidos, suelen ir precedidos por la conversión de los dos términos a comparar a una representación normalizada. Por ejemplo, en algunos casos la letra *Ñ* estará codificada como / y en otros como ¥ por lo que habrá que convertir ambos signos a *Ñ*. En algunos casos todos los caracteres serán mayúsculas, en otros solo lo será la letra inicial de cada palabra, etc. Hay que tener en cuenta que en las BD de los SI, los datos suelen tener orígenes muy diversos y por lo tanto haberse introducido con formatos y codificaciones no unificadas. Esto es especialmente cierto en datos, como los antropónimos, que no acostumbran a pasar por ningún control.

El proceso de normalización dependerá de cada entorno concreto, pero suele ser parecido al que a continuación describimos, que es al que hemos sometido nuestros ficheros antes del análisis de errores.

- Eliminación de los acentos. En la mayoría de BD de los SI casi no se introducen o se introducen mal (especialmente en el caso de los antropónimos)

- Paso de minúsculas a mayúsculas

- Unificación de la codificación de letras especiales. En nuestro caso solo hemos unificado la codificación de la *Ñ* (aparecía como *¥* y como */*). Podría presentarse la conveniencia de unificar la codificación de la letra *Ç* para los casos en que este símbolo, que puede aparecer en algún antropónimo catalán, se haya introducido usando codificaciones diversas. En nuestro *corpus* no hemos encontrado ni una sola aparición de *Ç* pues se supone que se introdujo como *C* o *S*.

- Se convierten al símbolo *@* todos los símbolos que no son letras de nuestros alfabetos (incluyendo en esos alfabetos la *Ñ* y la *Ç*) ni son los separadores habituales de palabras dentro de un campo. En nuestro caso hemos tomado como separadores habituales el espacio, el guión y el punto. En un SI seguramente no convertiríamos los signos especiales a *@* sino a espacio en blanco, pero aquí queremos hacer estadísticas que nos digan cuantos caracteres especiales hay y por lo tanto habrá que distinguirlos de los espacios. El alfabeto final tiene 32 símbolos.

- Se compactan a un solo espacio los espacios contiguos y se eliminan todos los espacios que pudieran existir al inicio y al final de cada campo

Ejemplo: *Deme¥éz de #Concca-d.ar ==> DEMEÑEZ DE @CONCCA-D.AR*

Por lo tanto si en nuestro procedimiento de análisis de errores apareciera una pareja de apellidos que en su fichero original fuera *Da Vita Rª Ro¥án / Tavita R3Roñan,* las diferencias que se detectarían serían tan solo tres; el cambio de la *D* por la *T* y las desapariciones de espacios. Las otras cuatro diferencias habrían desaparecido en la normalización que precede a la comparación, pues ya se habrían hecho las transformaciones siguientes:

$$á \rightarrow a \qquad ¥ \rightarrow Ñ \qquad ª \rightarrow @ \qquad 3 \rightarrow @$$

En los sistemas en los que hay una presencia significativa de personas cuya lengua usa otros símbolos, habrá que tener en cuenta su correcta transliteración.

También es frecuente que el proceso de normalización se encargue de estandarizar o deshacer abreviaturas. Por ejemplo, el paso de *J.Mª.* a *J MARIA* , o el paso de *FDEZ.* a *FERNANDEZ*. En nuestro caso hemos mantenido las abreviaturas tal como están en los ficheros originales. Pero en algunas pruebas las hemos eliminado, como ya se indicará en su momento. Las abreviaturas se encuentran casi exclusivamente en apellidos (o nombres) compuestos pues pretenden aligerar su excesiva longitud.

Algunas empresas incluyen en sus procesos de normalización, la eliminación de partículas (*DE LA ROSA* pasa a *ROSA*). Otras concatenan las partículas a la palabra siguiente (*DE LA ROSA* pasa a *DELAROSA*). Otras colocan las partículas al final (*DE LA ROSA* pasa a *ROSA DE LA*), etc. En nuestro caso no hemos aplicado ninguna normalización a las partículas, por lo que entran en el proceso de comparación tal como están en los ficheros originales. Por lo tanto:

$$\delta \text{ (De la Rosa , Delarosa)} = 2$$

En el fichero *MARKET*, que contiene parejas de apellidos, en un 8% de ellas aparecen partículas.

El que haya tres campos (nombre, primer apellido, segundo apellido) no garantiza que su contenido sea el previsto. Así podría aparecer el nombre en el campo previsto para el segundo apellido o podrían estar los apellidos cambiados,.. Por ello en algunas

correcciones la distancia podrá llegara ser tan grande como la más larga de las dos cadenas.

El conjunto de parejas del fichero *MARKET* tales que tengan: intercambio entre los dos apellidos, o algún apellido compuesto o alguna abreviatura en apellidos, suponen el 18% del total.

Ocurre a veces que algún campo aparece vacío, normalmente por desconocimiento de su valor. En la mayoría de personas extranjeras será normal que no aparezca nada en el segundo apellido. En la BD original de la que obtuvimos *MARKET* había un 38% de personas con el nombre de pila vacío, un 24% con el segundo apellido vacío y solo un 0,2% con el primer apellido vacío. Estas elevadas cifras de ausencias no son en absoluto inhabituales en muchos SI españoles. Para nuestras estadísticas de distancias no hemos incluido en *MARKET* las correcciones que daban un valor a un campo que estaba vacío. En estos casos la distancia de edición hubiera sido la longitud del valor dado al campo.

## 5.- Distribución de frecuencias de la distancia de edición

La mayor parte de trabajos sobre corrección ortográfica automática, búsqueda aproximada de palabras en textos, etc, dan por buena la cifra del 80% para los errores con una distancia de edición $\delta = 1$, que se cometen en la escritura de un texto. Es la cifra que dió en el año 1964 Damerau [Dam64]. En algunos trabajos se citan porcentajes superiores, por ejemplo el 94% [Pol84]. Con porcentajes tan elevados parece razonable escoger como criterio de semejanza que la distancia de edición sea $\delta = 1$. Pero en principio no se debería aceptar como válido, sin más, ese porcentaje para cualquier problema relativo a palabras y mucho menos para el caso de unas palabras tan particulares como los antropónimos. Por ello hemos estudiado la distribución de la distancia de edición en nuestro *corpus* y tal como era de esperar se han obtenido porcentajes bastante más pequeños: entre el 60% y el 70%, lo cual nos obligará a buscar otros criterios de semejanza.

### 5.1.- Apellidos

Para el análisis de las distancias de edición entre parejas (antes/después) de apellidos de nuestro *corpus*, mezclamos en un solo fichero las parejas del primer apellido y las del segundo, por lo que una persona con dos apellidos tendrá dos parejas en este fichero. El análisis del fichero resultante, da los resultados que aparecen en las figuras1 y 2. Recordamos que cada una de operaciones $T$ y $X$ cuenta aquí como dos operaciones básicas (una `D` y una `I`) .

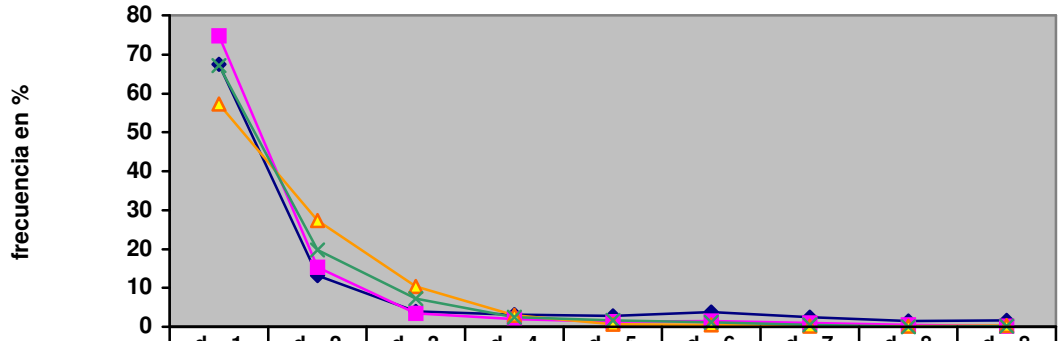
Del conjunto de las 11610 parejas de apellidos contenido en *MARKET* hemos eliminado todas las parejas (el 18%) cuyo error era producido por ..

- faltan o sobran palabras en el caso de apellidos compuestos por más de una palabra (ejemplo: *GOMEZ DEL MORAL / GOMEZ MORAL*)

- abreviaturas (ejemplo: *GLEZ / GONZALEZ*)

- intercambios entre primer y segundo apellidos (ejemplo: una misma persona puede tener la pareja *GARCIA / CLIMENT* obtenida de la corrección del primer apellido y la pareja *CLIMENT / GARCIA* la corrección del segundo)

**Figura *1*: Apellidos. Distribución de la distancia**




| | d = 1 | d = 2 | d = 3 | d = 4 | d = 5 | d = 6 | d = 7 | d = 8 | d > 8 |
|---|---|---|---|---|---|---|---|---|---|
| MARKET | 67,5 | 13,2 | 3,9 | 3,1 | 2,8 | 3,8 | 2,5 | 1,5 | 1,6 |
| MARKET-B | 74,7 | 15,3 | 3,5 | 1,9 | 1,3 | 1,4 | 1 | 0,5 | 0,2 |
| UNIVERS | 57,3 | 27,4 | 10,3 | 3 | 0,7 | 0,5 | 0,2 | 0,2 | 0,3 |
| CENSUSA | 67,1 | 19,8 | 7,2 | 2,4 | 1,6 | 1,2 | 0,5 | 0 | 0,1 |

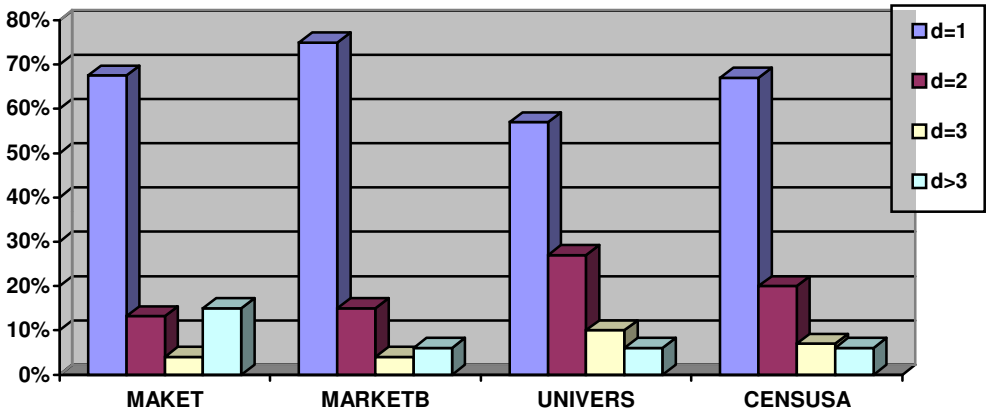
Al fichero resultante le llamamos *MARKET-B* y tiene 9538 parejas. La parejas eliminadas tienen en muchos casos distancias grandes, por lo que las frecuencias de las distancias pequeñas ($\delta < 3$) son más altas en *MARKET-B* que en *MARKET* . Y más bajas para las distancias grandes ($\delta > 3$) como se ve en la figuras 1 y 2.

Las medias de las distancias de edición y sus desviaciones tipo son:

*MARKET* : m = 2.0 σ = 1,97 *MARKET-B* : m = 1.5 σ = 1,23
*UNIVERS:* m = 1.7 σ = 1,07 *CENSUSA*: m = 1.6 σ = 1,08

Hay que hacer notar que *MARKET* contiene un número indeterminado, aunque importante, de parejas producidas por correcciones automáticas realizadas por la empresa de marketing originaria durante los procesos de normalización. Esas correcciones consisten en; substituir el apellido que no aparece en una lista de apellidos "conocidos", por aquel apellido conocido, si existe, que se diferencie del desconocido en tan solo un carácter. Estas correcciones no se hicieron sistemáticamente ni se conoce su

**Figura *2*: Apellidos**
**Distancias**




cuantía. Solo podemos afirmar que el número de parejas de este tipo se sitúa entre el 5% y el 25% del total. La experiencia de la empresa es que la inmensa mayoría de las correcciones automáticas son acertadas, aunque algunas corresponden a apellidos que

son correctos pero que el proceso de normalización ha supuesto incorrectos por no estar en su lista de "conocidos". En definitiva, la frecuencia de la $\delta = 1$ , tanto en *MARKET* como en *MARKET-B*, está algo sobredimensionada, y por lo tanto es solamente un límite superior del valor real.

En el fichero *UNIVERS* no hay ningún caso de abreviatura ni de intercambio y las parejas con algún apellido compuesto son tan solo el 4%.

En el fichero *CENSUSA* tampoco hay ningún caso de abreviatura ni de intercambio. Hay tan solo un 5% de parejas con apellidos compuestos, pero esas parejas tienen distancias cortas ya que siempre se trata de que falta/sobra una partícula corta (ejpls; VAN, MAC).

En los resultados mostrados en las figuras 1 y 2, para *MARKET* y *MARKET-B*, las parejas de apellidos están ponderadas con la frecuencia de su aparición. Así por ejemplo la pareja *GARCIA / GARICA* "pesa" 35 por aparecer ese número de veces. Si solo contamos las parejas diferentes, o sea las formas, la distancia media sube ligeramente.

El profesor Zobel tiene a disposición pública [Zob96] tres juegos de prueba con un total de 1265 parejas de apellidos utilizadas por él en su trabajo y que son el resultado de la apreciación de semejanza por parte de tres jueces humanos. Los resultados que hemos obtenido de su análisis son :

$\delta = 1$ .... 15.9%
$\delta = 2$ .... 33.4%
$\delta = 3$ .... 31.1%
$\delta > 3$ ... 19.6%

La enorme diferencia entre estas cifras y las obtenidas en nuestros *corpus* es debida, sin duda, a que en nuestro casdo el corpus es real y en el caso de Zobel son parejas resultado de la apreciación de jueces humanos. En general, los jueces humanos tienden a interesarse más por la semejanza fonética que por la semejanza gráfica.

## 5.2.- Nombres

Para el análisis de errores en los nombres de pila, también hemos utilizado la información contenida en *MARKET, UNIVERS* y *CENSUSA*. En el caso de *MARKET* se trata de un conjunto muy especial ya que consta csi exclusivamente de mujeres.

Como consecuencia de los procesos de normalización (ver sección 4) la mayoría de las parejas de nombres del fichero *MARKET* eran (o contenían) *MARÍA* o *M*. Para que no interfieran en el resultado las hemos eliminado, con lo que el total de parejas analizadas es de 5437. En la población española hay aproximadamente un 15% de mujeres cuyo nombre contiene la palabra *MARIA* (ver LSI-02-01-R).

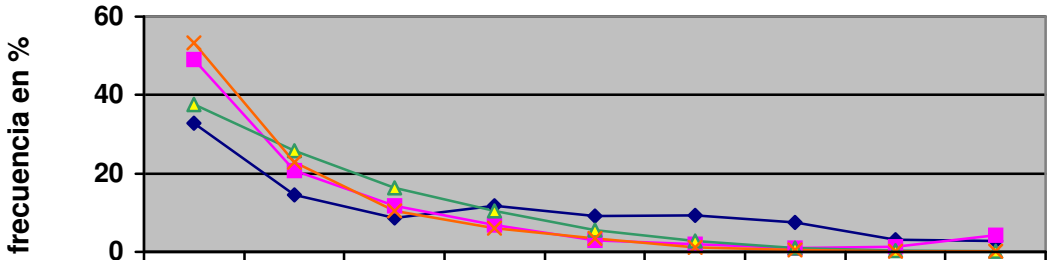
Hemos utilizado también un fichero llamado *NOMSCIM* con 11762 parejas de nombres de pila. El fichero se ha obtenido a partir del diccionario de nombres que utiliza el CIM (Centre d'Informàtica Municipal de Barcelona) pero de acuerdo con lo que se explica a continuación.

En el diccionario del CIM se van registrando todas las cadenas (correctas o no) que van apareciendo nuevas en la entrada de datos o en las consultas. Ese diccionario contenía en el momento de nuestro análisis 14100 nombres básicos, o sea correctos oficialmente, y 52000 variantes. Esas variantes pueden ser errores, formas diferentes (por ejemplo; hipocorísticos), lenguas diferentes, etc. Cuando desde cualquier puesto de trabajo se

introduce un nombre que no está en el diccionario, se intenta averiguar (por medios humanos) si es un nombre nuevo, básico, o bien es una variante (por error u otra causa) de uno básico ya existente. En cualquiera de los dos casos se registra en el diccionario. En muchos casos de nombres extraños, por ejemplo de personas extranjeras, ocurre que no se logra detectar si se trata o no de una variante, y se registra como nombre básico. El hecho de que la mitad de los nombres introducidos como básicos no tengan variante alguna, se explica en parte por esto.

El CIM tiene la política de considerar variantes casi todos los compuestos. En su diccionario, solo el 2% de los nombres básicos son compuestos y en cambio lo son el 60% de las variantes. Por ejemplo: *JOSE JUAN* es considerado una variante de *JUAN*. Por ello nuestro fichero *NOMSCIM* se ha formado con todas aquellas parejas "nombre básico / variante" de ese diccionario eliminando las que tienen algún compuesto. Tras la eliminación de compuestos, *NOMSCIM* tiene 11762 parejas.

**Figura *3*: *Nombres***
**Distribución de la distancia**




| | d = 1 | d = 2 | d = 3 | d = 4 | d = 5 | d = 6 | d = 7 | d = 8 | > 8 |
|---|---|---|---|---|---|---|---|---|---|
| MARKET | 32,8 | 14,5 | 8,6 | 11,8 | 9,2 | 9,4 | 7,5 | 3,1 | 2,8 |
| UNIVERS | 49 | 20,7 | 11,8 | 6,9 | 3 | 2 | 1 | 1,3 | 4,3 |
| NOMSCIM | 37,6 | 25,8 | 16,3 | 10,5 | 5,6 | 2,7 | 1 | 0,4 | 0,1 |
| CENSUSA | 53,3 | 22,9 | 10,4 | 6 | 3,4 | 1,2 | 0,5 | 0,1 | 0,15 |

El análisis de parejas de nuestro *corpus* para el caso de nombres de pila, revela que las distancias pequeñas son mucho menos frecuentes que en los apellidos. Ver la figura 3. Veamos algunas de las razones que justifican este comportamiento distinto al de los apellidos:

- A diferencia de los apellidos, en los nombres existen muchas variantes, como por ejemplo los hipocorísticos (*PAQUITA/FRANCISCA, PEPE/JOSE, LOLA/DOLORES*). En *MARKET,* aproximadamente un 14% de las parejas son de este tipo.

- El repertorio de los nombres, así como el de sus variantes, es bastante limitado y en su mayor parte les es conocido a las personas que registran manualmente la información. Por ello no se suelen cometer errores por desconocimiento de la ortografía del nombre, que es precisamente una de las principales causas de distancia 1. En los nombres, la inmensa mayoría de los casos de distancia 1 son errores de tecleado, aunque hay algunos casos de pequeñas variantes lingüísticas (*JOSEP / JOSE, ESTHER / ESTER, JAVIER / XAVIER*) y de confusión entre dos nombres válidos pero casi iguales.

- En los nombres se suelen utilizar muchas abreviaturas, algunas de distancia corta (suelen ser hipocorísticos como *NURI / NURIA*) pero la mayoría son

abreviaturas radicales que dejan solo la primera letra. Y se producen errores por interpretación errónea de abreviaturas (interpretación de *J.* como *JOSE* cuando quizás debería ser *JUAN*).

- Aunque los nombres compuestos se dan en España con mucha menor frecuencia que en los países hispano-americanos, se producen gran cantidad de problemas por ausencia de alguna palabra o por abreviaturas ( *JUAN JOSE / JOSE / JUAN / JUANJO / J.JOSE / JJ* ). En *MARKET* y en *UNIVERS* hay parejas con compuestos, pero en *NOMSCIM* y *CENSUSA* no las hay. En el caso de *NOMSCIM,* si no hubiéramos eliminado los compuestos, los resultados del análisis nos darían las siguientes frecuencias, muy bajas, para las distancias pequeñas:

$\delta = 1 \rightarrow 12\%$ $\delta = 2$ o $3 \rightarrow 24\%$ $\delta > 3 \rightarrow 54\%$

- El hecho de que en *CENSUSA* la frecuencia de los casos con $\delta = 1$ sea elevada, puede explicarse porque hay pocos hipocorísticos y ninguna abreviatura ni compuestos.

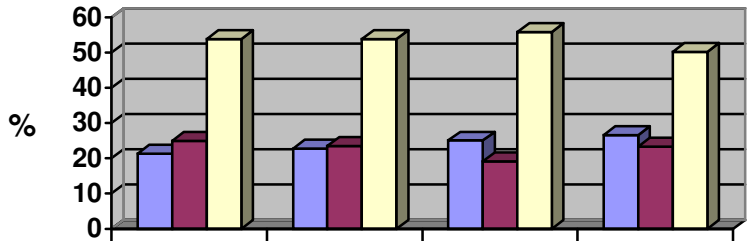
## 6.- Distribución de las operaciones para $\delta = 1$: las parejas-uno

En esta sección estudiaremos con un cierto detalle las parejas de apellidos cuya distancia de edición es $\delta = 1$, a las que llamaremos *parejas-uno* (alrededor del 60% del *corpus*).

### 6.1.- Los tipos I D y S

Del análisis de los tres tipos de operaciones básicos, I D y S, para todas las parejas-uno, se deduce que el número de inserciones de un carácter es muy similar al de supresiones y ligeramente inferior a la mitad del de substituciones. En la figura 4 se da para cada fichero, la frecuencia, en porcentaje, de parejas-uno.

**Figura *4*: Apellidos. Parejas-uno.**




| | MARKET | MARKET-B | UNIVERS | CENSUSA |
|---|---|---|---|---|
| I | 21,2 | 22,7 | 25 | 26,5 |
| D | 24,9 | 23,4 | 19,1 | 23,2 |
| S | 53,8 | 53,8 | 55,9 | 50,2 |

Las únicas cifras sobre la distribución de los tres tipos de operación parejas-uno, que disponemos de otros estudios son las obtenidas por Tagliacozzo [Tag70] relativas a nombres de autores en una BD de bibliografía médica, y las de Joseph (citadas en

[Pol82]) sobre texto libre en inglés. Las cifras de esos trabajos las hemos ajustado para que se correspondan a las mismas condiciones de medida que las nuestras:

| | I | D | S |
|---|---|---|---|
| Tagliacozzo | 17,5% | 33,7% | 48,8% |
| Joseph | 20,8% | 42,5% | 36,7% |

## 6.2.- Letras afectadas y tendencia al error

Estudiaremos aquí la frecuencia con la que los distintos caracteres son sujetos de los distintos tipos de operaciones de edición. Las frecuencias relativas podremos considerarlas una aproximación a la probabilidad de que se produzcan esas operaciones.

Los caracteres que no son letras (en nuestro caso el guión, el punto, el espacio y *@*) y las letras *Ñ* y *Ç*, plantean una enorme variabilidad de comportamiento dependiendo de como se haya hecho la normalización manual o automática (ver sección 4). Por ello nos limitaremos aquí a las letras y excluiremos la *Ñ* y la *Ç*. El alfabeto resultante tiene pues 26 símbolos.
Advertimos que la presencia de la letra *K* es casi nula en los ficheros españoles que hemos utilizado.

En un contexto básicamente vasco, según qué reglas ortográficas se siguieran, habría habido muchas más apariciones de la letra *K*.

Para cada letra y tipo de operación de edición, definimos las siguientes magnitudes:

- Una frecuencia $F$, como $$F = \frac{apariciones\,de\,la\,letra\,en\,la\,operación}{número\,de\,parejas\,uno}$$

- Una medida de la tendencia al error, $T$, de una letra en un tipo de operación, como la frecuencia $F$ condicionada a la frecuencia $Fr$ de aparición de la letra en el conjunto (ponderado) de apellidos españoles. Como valor de las frecuencias $Fr$ hemos adoptado las obtenidas en LSI-02-01-R a partir del fichero *APELLIDOS* (conjunto de 1,6 millones de españoles). Por lo tanto, expresando $T$ en porcentaje tendremos:

$$T = \frac{F}{Fr \times \sum_{\forall letra} \frac{F}{Fr}} \times 100$$

### 6.2.1.- Inserciones y supresiones

En el estudio de la distribución de frecuencias de las operaciones `ID` *(a)* (ver sección 2) en las parejas-uno de *MARKET-B* se encuentran prácticamente los mismos resultados que en *MARKET*. Entre *MARKET* y *UNIVERS* hay algunas diferencias que ya iremos señalando, pero hemos decidido presentar los resultados obtenidos juntando las parejas-uno de ambos conjuntos, formando así un conjunto de 8449 parejas-uno al que llamaremos *MARKUNO* y que por lo tanto tendrá solamente parejas con $\delta = 1$.

Hemos tomado las operaciones de inserción y supresión conjuntamente, `ID` *(a)*, porque en el caso de apellidos no suele tener sentido el considerar un término de la pareja como

erróneo y el otro como correcto. En la tabla 1 aparecen las frecuencias $F$ de las letras, determinadas a partir de *MARKUNO,* junto a las frecuencias $Fr$ determinadas a partir de *APELLIDOS*. Todas las frecuencias aparecen en forma de %.

Considerando la frecuencia $F$, la letra que más aparece en inserciones o supresiones es la *S*. El 8,6% de las parejas-uno son supresiones o inserciones de esta letra. Dicho de otra forma, aproximadamente el 19% de las inserciones o supresiones lo son de la letra *S*. Le sigue la letra *A* con el 7,2% de las parejas-uno. Entre la *S* y la *A* representan el 35% de las inserciones o supresiones. Luego, más lejos, aparecen las letras *R*, *N*, *L*, *I.*

Si consideramos la tendencia al error $T$, la letra *S* (letra frecuente pues tiene $Fr$ = 5,3%) sigue siendo problemática. También destacan la letra *H* que es de frecuencia media y algunas letras poco frecuentes como *K*, *Y, W* .

| | | ID*(a)* | | S*(ab)* y S*(ba)* | |
|---|---|---|---|---|---|
| | *Fr%* | *F%* | *T%* | *F%* | *T%* |
| **A** | 14,8 | 7,2 | 1,3 | 6,1 | 0,8 |
| **B** | 2 | 0,3 | 0,4 | 1,9 | 1,7 |
| **C** | 3,7 | 0,7 | 0,6 | 2,1 | 1,1 |
| **D** | 2,6 | 1,1 | 1,1 | 1,5 | 1,1 |
| **E** | 10,6 | 4,3 | 1,1 | 3,6 | 0,7 |
| **F** | 0,9 | 0,4 | 1,1 | 0,5 | 1 |
| **G** | 3 | 0,4 | 0,4 | 1,9 | 1,2 |
| **H** | 1 | 1 | 2,7 | 0,8 | 1,5 |
| **I** | 5,8 | 3 | 1,4 | 2,4 | 0,7 |
| **J** | 0,6 | 0,3 | 1,2 | 1,1 | 3,4 |
| **K** | 0,04 | 0,2 | 12,9 | 0,3 | 11,9 |
| **L** | 6,7 | 3,3 | 1,3 | 2,6 | 0,7 |
| **M** | 3,3 | 0,5 | 0,4 | 2,1 | 1,2 |
| **N** | 5,9 | 3,4 | 1,5 | 3,2 | 1 |
| **O** | 8,3 | 2,4 | 0,8 | 4,8 | 1,1 |
| **P** | 2 | 0,4 | 0,5 | 0,9 | 0,9 |
| **Q** | 0,3 | 0,2 | 1,5 | 0,3 | 1,5 |
| **R** | 10,5 | 4,9 | 1,3 | 2,8 | 0,5 |
| **S** | 5,3 | 8,6 | 4,4 | 3,1 | 1,1 |
| **T** | 3,4 | 1,5 | 1,2 | 1,1 | 0,6 |
| **U** | 3,1 | 1,7 | 1,4 | 2 | 1,2 |
| **V** | 1,5 | 0,3 | 0,5 | 1,3 | 1,6 |
| **W** | 0,02 | 0,1 | 3,2 | 0,1 | 1,7 |
| **X** | 0,1 | 0,05 | 1,2 | 0,2 | 3,7 |
| **Y** | 0,4 | 0,3 | 2,2 | 1,3 | 6,1 |
| **Z** | 3,8 | 1,8 | 1,3 | 2,2 | 1,1 |

**Tabla 1: Letras afectadas. Frecuencia y tendencia**

Separando las operaciones `D` *(a)* de las `I` *(a)* se ve que apenas hay diferencias entre ellas. Las dos principales son, que la letra *S* tiene algo más de probabilidades de ser suprimida que de ser insertada, y que a las letras *E* y *A* les ocurre al revés, o sea que es más probable su inserción que su supresión.

Recuérdese que como hemos dicho en la sección 2, las operaciones no las hemos definido como operaciones de corrección de errores, sino de transformación de una cadena en otra.

Podría interesarnos entrar en el detalle de las operaciones `I` *(a)* y `D` *(a)* considerando el carácter anterior y posterior al *a*. Entonces las podríamos llamar *operaciones detalladas* y notarlas así

`I` *(a,c,d)* `D` *(a,c,d)* siendo *c* el carácter anterior y *d* el posterior a *a*

Para notar el inexistente "carácter anterior" al primero del apellido podríamos usar el símbolo *[* y el símbolo *]* para el "carácter siguiente" al último. Así, *c* podria valer *[* y *d* podria valer *]*.

El número de posibilidades del par *cd* es $\alpha^2+2\alpha$, por lo tanto el número total de posibles operaciones detalladas de inserción o supresión será

$$2\alpha(\alpha^2+2\alpha) = 2\alpha^3 + 4\alpha^2$$

Esto supone en nuestro caso 36856 posibilidades. Pero nuestra muestra de parejas de apellidos con una inserción o supresión, no llega a los 4000 ejemplares. Por lo tanto no consideramos de suficiente confianza el resultado que obtuviéramos de un análisis de las operaciones detalladas. Nos limitaremos a citar aquí las operaciones más frecuentes en *MARKUNO* :

`I` *(S, A, ] )* `I` *(S, O, ] )* `I` *(S, E, ] )* `I` *(N, O, ] )* `I` *(R, E, ] )* `I` *(R, A, ] )*
`D` *(S, O, ] )* `D` *(Z, E, ] )* `D` *(U, G, E)* `D` *(S, A, ] )* `D` *(S, E, ] )* `D` *(L, L, A)*

El valor más frecuente para el carácter *d* (el posterior) es *]* y para el *c* (el anterior) es *A*. Dicho de otra manera; la letra tras la cual hay más probabilidades de que se hagan inserciones o supresiones, es la letra *A*. Y la posición más probable es al final del apellido.

### 6.2.2.- Substituciones

Analizaremos ahora la presencia de las distintas letras en las operaciones de substitución. Para cada letra *a* contaremos todas las parejas-uno con substitución en las que la letra está involucrada, tanto si es la letra substituida como si es la substituyente. Consideremos el conjunto de parejas siguiente

| | |
|---|---|
| *GIMENEZ / JIMENEZ* | `S` *( G, J )* o bien `SS` *( G, J )* |
| *OJIN / OGIN* | `S` *( J, G )* o bien `SS` *( G, J )* |
| *AQUILO / AGUILO* | `S` *( Q, G )* o bien `SS` *( G, Q )* |

En este conjunto de tres parejas, la letra *G* está involucrada en todas; la *J* en dos y la *Q* solo en una.

Los resultados de este análisis de las substituciones en *MARKUNO,* aparecen también en la tabla 1. Como en cada operación de substitución están involucrados dos caracteres, la operación cuenta dos veces. Por ello hemos dividido las frecuencias por 2.

Vemos en esa tabla que la letra *A* es la que tiene más probabilidades de aparecer en las substituciones, ya sea como letra substituida ya sea como substituyente. El 6,1% de las

parejas-uno son substituciones de, o por, la letra *A*. Dicho de otra forma, la letra *A* está involucrada en el 12% aproximadamente de las substituciones. Le siguen a cierta distancia las letras *O, E, N, S, R, L* y *Z* (en ese orden).

Si observamos la tendencia *T*, vemos que las letras más propensas a ser substituidas o a substituir, son: en primer lugar la letra *K* (recordar que este resultado es poco fiable por la escasa presencia de esta letra) seguida de *Y, X , J* y algo más lejos las letras *W, B*, *V, H* y *Q*.

Hemos analizado también las operaciones simétricas SS*(a,b)* y hemos encontrado que las parejas *ab* más frecuentes son *AO* y *AE*. Como estamos hablando de operaciones simétricas, cuando hablamos de SS *( A, O )* estamos refiriéndonos tanto al cambio de *A* por *O*, como al de *O* por *A* (ver la sección 2).

Las operaciones SS *( A, O )* y SS *( A, E )* suponen, en conjunto, el 11% de las parejas de apellidos con substitución. A estas dos parejas de letras les siguen *MN, SZ, OU, BV* y *EI.* Estas cinco suponen el 16,9% de las substituciones. Y algo más lejos aparecen las parejas *LR, NR, EO, CL, JG, HM* y *OI*, que entre las siete suponen el 11% de las substituciones.

Como en el caso de las inserciones y supresiones, podría interesarnos entrar en el detalle de las operaciones S *(a,b)* considerando el carácter anterior al *a*, y el posterior al *b*. Esas operaciones detalladas podríamos notarlas así

S *(a,b,c,d)* *c* es el carácter anterior al *a*
*d* es el carácter posterior al *b*

Recordar que *d* puede valer *]* y *c* puede valer *[* . Por lo tanto:

- el número de parejas *ab* posibles es $\alpha^2-\alpha$
- el número de parejas *cd* posibles es $\alpha^2+2\alpha$
- el número total de posibles operaciones detalladas de susbstitución es
$(\alpha^2 - \alpha)(\alpha^2 + 2\alpha)$

Esto supone en nuestro caso 473200 posibilidades. Pero nuestra muestra de parejas de apellidos con una substitución no llega a los 5000 ejemplares, por lo que no consideramos de suficiente confianza el resultado que obtuviéramos de un análisis de las operaciones detalladas. Nos limitaremos a citar aquí las operaciones más frecuentes en *MARKUNO:*

S *( Z, S, E, ] )* S *( S, Z, E, ] )* S *( O, A. R, ] )* S *( M, N, [, A )*
S *( A, O, R, ] )* S *( A, O, N, ] )* S *( A, O, L, ] )* S *( E, A, B, R )*

El valor más frecuente del carácter *d* (el posterior) es *]* y el de *c* (el anterior) es *[* . Dicho de otra forma, las posiciones en las que las substituciones son más probables son el principio y el final del apellido.

### 6.2.3.- Operaciones de edición en los apellidos norteamericanos

Del análisis de las parejas-uno del *corpus CENSUSA*, se deduce que las operaciones simétricas más frecuentes en EUA son:

ID *( S )* ID *( E )*

SS *( A, O )*  SS *( A, E )*

Vemos que son más o menos las mismas que en el caso español. Estas operaciones por si solas, suponen ya el 20% de las parejas de *CENSUSA*.

Las operaciones detalladas más frecuentes son:

D *( L, L, ] )*  D *( S, N, ] )*  D *( R, E, ] )*  D *( S, E, ] )*
S *( S, Z, E, ] )*  S *( R, N, E, ] )*

Al igual que en el caso español, la *S* es la letra más afectada, pero no le sigue la *A* sino la *E*. Algo más lejos siguen la *L, R, N* y *A*.

## 7.- Distribución de las operaciones para $\delta > 1$

Acabamos de evaluar en la sección 6 las parejas-uno (aquellas cuya distancia de edición es 1) o sea el 65% aproximadamente de los errores o variantes contenidos en *MARKET* y *UNIVERS*. Veamos ahora la distancia media del resto de parejas (no parejas-uno):

*UNIVERS* $\rightarrow$ $\delta = 2{,}5$
*MARKET* $\rightarrow$ $\delta = 4{,}2$
*MARKET-B* $\rightarrow$ $\delta = 3{,}2$

Recuérdese que en *MARKET* hay un 18% de apellidos compuestos, abreviaturas e intercambios de apellidos (ver sección 4).

### 7.1.- Distancia de edición $\delta = 2$ : Operaciones no básicas

Las parejas de apellidos con $\delta = 2$ son en *MARKET* el 13% , en *MARKET-B* el 15% y en *UNIVERS* el 27%.

En la sección 2 hemos definido las operaciones, no básicas, de intercambio y de transposición, *X* y *T*. Son transformaciones que en términos de operaciones básicas tienen una $\delta = 2$. En nuestros *corpus* su frecuencia conjunta en relación al número total de parejas, oscila entre el 2% y el 3%, cifra suficientemente alta para que esas operaciones sean tenidas en cuenta. Si separamos las operaciones *T* de las operaciones *X*, obtenemos los siguientes porcentajes respecto al total de parejas con $\delta = 2$:

| | *UNIVERS* | *MARKET* | *MARKET-B* |
|---|---|---|---|
| *T*..... | 6,8% | 17,2% | 15% |
| *X*..... | 1,3% | 2,2% | 2,3% |

En las parejas con $\delta = 2$, pero que no son ni *T* ni *X*, hemos analizado qué otras transformaciones de dos caracteres se dan con mayor frecuencia y hemos visto que son:

- conversiones: *LL* $\longleftrightarrow$ *Y*  *TX* $\longleftrightarrow$ *CH*  *CH* $\longleftrightarrow$ *G*
- eliminación/inserción del bigrama *LA* al principio del apellido

Hemos analizado también esas parejas (las que tienen $\delta = 2$ pero que no corresponden a operaciones *T* ni *X* ) en términos de las tres operaciones básicas ( I, D, S ) y el

resultado muestra una distribución casi idéntica a la obtenida en las parejas-uno. Por ello las frecuencias dadas en la tabla 1 las tomaremos como válidas para todas las parejas, no solo las parejas-uno.

También hemos obtenido algunas cifras de *CENSUSA* respecto a $\delta = 2$. En ese fichero norteamericano, las operaciones del tipo `T` son el 18% de las parejas de $\delta = 2$. Las conversiones de $\delta = 2$ encontradas con más frecuencia son:

| | | | | |
|---|---|---|---|---|
| *CH* → *G* | *GU* → *GO* | *GU* → *Q* | *PH* → *F* | *CK* → *X* |
| *ER* → *U* | *ER* → *A* | *SCH* → *S* | *SH* → *TCH* | |

## 7.2.- Distancias superiores ($\delta > 2$)

El número de operaciones básicas `I` *(a),* `D` *(a),* `S` *(a,b)* posible es $\alpha(\alpha+1)$ y por lo tanto es 702 en nuestro caso ya que $\alpha = 26$. Las parejas con distancia superior a 2, son en nuestro *corpus* de un 10% a un 15% aproximadamente. Su número no es suficientemente grande para que los resultados de su análisis en términos de las 702 operaciones básicas, sean muy fiables. Pero ocurre además, que la determinación de las operaciones de edición de una pareja puede ser imposible, ya que a partir de $\delta = 3$ no existe una secuencia única de operaciones básicas equivalente (ya hemos visto en la sección 2 con el ejemplo de la transformación *GIMENEZ* → *JIMNNEZ*, que tanto se podía hacer con tres `S` , como con una `S` una `D` y una `I`).

La mayoría de inserciones o supresiones de partículas tienen $\delta = 3$ (dos letras y un espacio). En *MARKET-B* (no tiene compuestos ni abreviaturas ni intercambios de apellidos) el 20% de los casos con $\delta = 3$ son inserciones o supresiones de las tres letras iniciales o de las tres finales.

Las parejas con distancia 4 o 5, tienen exclusivamente errores relacionados con apellidos compuestos o bien errores por tratarse de un contexto lingüístico con características gráficas y fonéticas muy distintas a las de la lengua del receptor. Ejemplo de error entre distintas lenguas españolas: *ETXEBARRIA / ECHEVERRIA* ($\delta = 4$). Ejemplo de error por lengua extranjera: *SCHAEIDT / SAINZ* ($\delta = 5$).

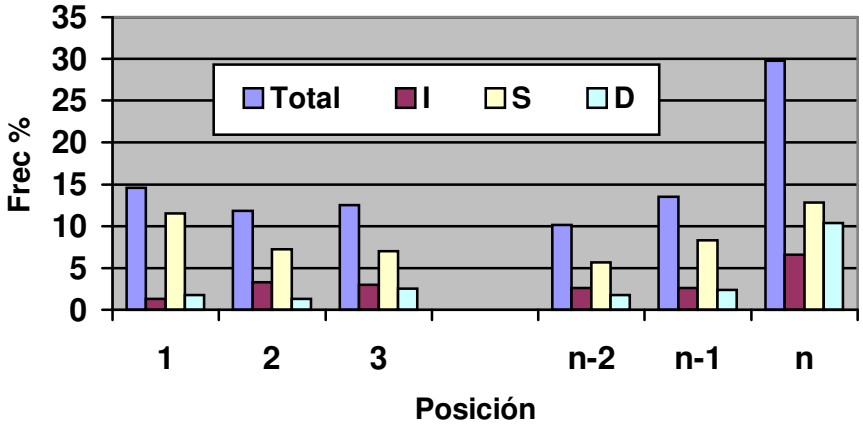
Prácticamente todos los casos con $\delta > 5$, son producidos por abreviaturas, apellidos compuestos (no partículas) o intercambios. Las excepciones suelen ser problemas con el catalán o el vasco, como por ejemplo *AMETXAZURRA / AMEZARRUCHA* ($\delta = 6$).

Nos ha parecido interesante investigar la presencia de transposiciones (contiguas o no) de parejas de caracteres. Serían casos especiales de $\delta = 4$. Por ejemplo entre *SARALO* y *SALORA* hay una transposición de dos parejas contiguas y entre *SARALO* y *LORASA* la hay no contigua. Pues bien, para nuestra sorpresa no hemos encontrado ni un solo caso de esas transposiciones en ninguno de los ficheros que hemos estudiado.

## 8.- Posición y tipo de operación

Tradicionalmente se ha considerado que la probabilidad de error en una posición de una palabra es tanto mayor cuanto más a la derecha esté esa posición. Por ejemplo, es típico considerar que la letra inicial apenas sufre errores y que las posiciones finales son las más problemáticas. Ya hemos visto al final de 6.2.2 que esto no es así. Un análisis de la posición de los errores en nuestro *corpus MARKUNO,* nos indica que los supuestos tradicionales no son del todo correctos. Veámoslo en la figura 5.

**Figura *5*: Posición de los errores**




Las substituciones, `S`, son muy probables en la primera y última posiciones, y un poco menos en la penúltima. Tanto las inserciones, `I`, como las supresiones, `D`, son muy probables en la última posición. El total de `I` más `D`, está en todas las posiciones alrededor del 5%, excepto en la última posición en la que se dispara al 18%. La posición más afectada es, con mucha diferencia, la última pero le sigue la primera, en contra del supuesto tradicional.

En la función de distancia propuesta en LSI-02-06-R, el coste de las operaciones depende de la posición y de las letras afectadas. Hemos hecho un análisis detallado de las operaciones por posición y letra. Damos a continuación algunas de las conclusiones más importantes:

- Casi todas las operaciones `I`, `D` o `S` de las letras *S* y *Z*, se dan al final de palabra.

- En la 1ª posición, las operaciones más probables son la `I` *(A)* y las substituciones de consonantes, especialmente `SS` *( M , N )* y `SS` *( B, V )*. Como se ve en LSI-02-01-R , la *M* y la *B* y en general las consonantes, son letras frecuentes en la 1ª posición de los apellidos.

- En la 2ª posición destaca la `D` *(A)* y las substituciones de vocales ( excepto la *I* ). Esto se corresponde con el hecho de que las vocales son las letras más frecuentes en la 2ª posición.

- En la última posición, las operaciones más probables son las `ID` *(S).* Precisamente la letra *S* es frecuente en la posición final.

## 9.- Longitudes y distancia de edición

Es razonable pensar que cuanto más largo sea un apellido, mayor sea su distancia de edición a sus deformaciones o variantes. En [Tag70] se daban cifras en este sentido. Damos a continuación la distancia media por cada longitud, obtenida a partir de *MARKET-B*. Conviene recordar que en ese fichero de parejas no hay compuestos ni abreviaturas ni intercambios. La longitud es la del primer miembro de la pareja (en el caso de *MARKET-B* es el "correcto").

| longitud: | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\delta$ media: | - | - | 1,3 | 1,3 | 1,4 | 1,5 | 1,4 | 1,5 | 1,7 | 1,8 | 2,6 | 3,6 | 5,8 | 6,7 | 7,2 |

Para apellidos de menos de 10 caracteres la influencia de la longitud sobre el número de errores no es muy grande. Pero a partir de esa longitud la influencia es enorme, a pesar de que en *MARKET-B* no hay compuestos ni abreviaturas. En LSI-02-01-R vimos que la inmensa mayoría de apellidos (más del 90%) tienen una longitud entre 4 y 9 .

## 10.- Confusiones

Para la lengua inglesa se suele decir [Kuk92] que alrededor del 15% de los errores producen palabras existentes (que están en el diccionario). En el caso de los nombres de pila de nuestro *corpus*, el 50% de los errores produce nombres que ya existen en la BD. En muchos de esos casos el error es una confusión entre nombres, ya que la mayor parte de los nombres de pila son perfectamente conocidos por la mayoría de personas. Para el caso de los apellidos no existe este conocimiento, pero la influencia de un apellido muy conocido puede llevar a confundirlo. Para hacernos una idea de la importancia de la confusión en el caso de los apellidos, hemos mirado cuales de los apellidos erróneos no son palabras raras sino apellidos más frecuentes que los propios apellidos correctos. Esas frecuencias han sido consultadas en el fichero *APELLIDOS* (ver LSI-02-01-R) que contiene los apellidos de unos 1,6 millones de españoles.

Hemos encontrado que el 5% de los apellidos erróneos en *MARKET* y el 20% en *UNIVERS,* son más frecuentes que su corrección. En el 50% de esos casos, el erróneo tiene una frecuencia más de 10 veces superior al correcto y en algunos casos se llega a varios miles de veces superior. Por ejemplo; son muchas las parejas en las que los frecuentísimos apellidos *GARCIA*, *MARTINEZ*, *GONZALEZ*, *GOMEZ, HERNANDEZ,* o *RUIZ,* eran incorrectos y han sido corregidos, respectivamente, por *GRACIA*, *MARTINES*, *GONZALO*, *POMES, HERNANZ,* o *URIZ,* apellidos que tienen frecuencias inferiores en varios órdenes de magnitud.

## REFERENCIAS

**[Dam64]** Damerau, F.J.: A technique for computer detection and correction of spelling errors. *Comm. ACM 7*,3 (Mar. 1964) 171-176.

**[Kuk92]** Kukich, K.: Spelling Correction for the Telecommunications Network for the Deaf. *Commun.ACM 35*,5 (1992), 80-90.

**[Nav01]** Navarro, G.: A guided tour to approximate string matching. *ACM Computing Surveys 33,*1 (Marzo 2001) 32-88

**[Pol82]** Pollock, J.J.: Spelling error detection and correction by computer: some notes and a bibliography. *J. Doc. 38*,4. (Dec. 1982).

**[Pol84]** Pollock J.J. Zamora A.: Automatic spelling correction in scientific and scholarly text. *Comm. ACM 27*,4. (April 1984) 358-368.

**[Tag70]** Tagliacozzo, R. et al.: Orthographic error patterns of author names in catalog searches. *Journal of Library Automation, 3* (1970) 93-101.

**[Win95]** Winkler, W.E.: Matching and record-linkage, cap.20 de *Business Survey Methods* (1995) 355-384. John Wiley & Sons.

**[Zob96]** Zobel, J. & Dart, P.: Phonetic string matching: Lessons from Information Retrieval. *Proceed.19th Annual Intl. ACM SIGIR Conf. on R&D in IR*. (Aug.1996) 166-173.