



①9



OFICINA ESPAÑOLA DE
PATENTES Y MARCAS

ESPAÑA

①1 Número de publicación: **2 316 474**

⑤1 Int. Cl.:
G06F 17/27 (2006.01)

①2

TRADUCCIÓN DE PATENTE EUROPEA

T3

⑨6 Número de solicitud europea: **01970463 .4**

⑨6 Fecha de presentación : **26.09.2001**

⑨7 Número de publicación de la solicitud: **1325429**

⑨7 Fecha de publicación de la solicitud: **09.07.2003**

⑤4 Título: **Método para normalizar caja.**

③0 Prioridad: **26.09.2000 SE 2000103433**

④5 Fecha de publicación de la mención BOPI:
16.04.2009

④5 Fecha de la publicación del folleto de la patente:
16.04.2009

⑦3 Titular/es: **Hapax Limited**
Caparo House, 103 Baker Street
London W1U 2LN, GB

⑦2 Inventor/es: **Ejerhed, Eva**

⑦4 Agente: **Elzaburu Márquez, Alberto**

Aviso: En el plazo de nueve meses a contar desde la fecha de publicación en el Boletín europeo de patentes, de la mención de concesión de la patente europea, cualquier persona podrá oponerse ante la Oficina Europea de Patentes a la patente concedida. La oposición deberá formularse por escrito y estar motivada; sólo se considerará como formulada una vez que se haya realizado el pago de la tasa de oposición (art. 99.1 del Convenio sobre concesión de Patentes Europeas).

DESCRIPCIÓN

Método para normalizar caja.

5 Campo del invento

El presente invento se refiere en general al campo del análisis, procesado y almacenamiento de texto en lenguaje natural por ordenador, y más concretamente a un método para diferenciar las distinciones no significativas de las distinciones significativas de las letras de la caja alta (mayúsculas) y de la caja baja (minúsculas) en una serie de tipos de palabras de entrada de entrada de un texto en lenguaje natural.

Antecedentes del invento

15 Cuando se analiza, se procesa y se almacena texto en lenguaje natural surgen varios problemas relativos a la caja a la que pertenezcan las letras del texto. Por ejemplo, cuando se almacenan los tipos de palabra de un texto largo en una base de datos, surge la cuestión de si una diferencia únicamente en cuanto a la caja de la que proviene una letra es relevante, o no, cuando se diferencian tipos de palabra.

20 Los sistemas conocidos para analizar, procesar y almacenar tipos de palabra tienen dos soluciones generales para manipular las distinciones en cuanto a la caja. Las dos soluciones son: (1) prescindir de toda diferenciación en cuanto a la caja alta y caja baja en los tipos de palabra única (insensibilidad a la caja), o (2) preservar todas las diferenciaciones en cuanto a la caja (sensibilidad a la caja). La primera solución da por resultado inventarios más pequeños de tipos de palabra, a costa de pérdida de la información que llevan consigo las diferenciaciones en cuanto a la caja, y la segunda solución da por resultado la conservación de la información en cuanto a la caja a costa de unos inventarios mayores de tipos de palabra.

En el documento US-A-5 485 372 se describe un método para establecer si una palabra es un nombre común, frente a si es un nombre propio u otra palabra que se debe escribir con mayúscula siempre. Las palabras son etiquetadas como palabras que se deben escribir con mayúscula siempre con vistas a ese establecimiento, y las palabras que no han de escribirse con mayúscula siempre no se escriben con mayúscula.

Sumario del invento

35 Un objeto del presente invento es resolver el problema de la pérdida de información asociada a la insensibilidad a la caja, y el problema de los grandes inventarios de tipos de palabra asociados con la sensibilidad a la caja, respectivamente, al tiempo que se conservan las ventajas de esas dos soluciones. Este objeto se consigue por un método para diferenciar automáticamente las variantes significativas de las no significativas de la caja alta y de la caja baja en una serie de tipos de palabra de entrada, de acuerdo con las reivindicaciones que se acompañan.

40 El invento está basado en el reconocimiento de que la información local, tal como la ocurrencia y la localización de las letras de la caja alta (mayúsculas) en tipos de palabra, juntamente con la información global, tal como la ocurrencia de tipos de palabra que solamente difieran con respecto a la caja a la que pertenezcan una o más letras, pueden ser usadas para determinar si la diferenciación en cuanto a la caja a la que pertenezcan las letras es significativa, o no.

45 De acuerdo con un aspecto del invento, se propone un método para diferenciar automáticamente las distinciones significativas de las no significativas de las cajas alta y baja, en una serie de tipos de palabras de entrada, por medio de un ordenador. De acuerdo con el método, se asigna un tipo de palabra de entrada a uno de una serie de grupos locales separados basados en la caja y en la posición de las letras que constituyen el tipo de palabra. Además, dicho tipo de palabra de entrada se vuelve a asignar a uno de los grupos globales separados, en base a cuáles sean los grupos locales a que se asignen las variantes en cuanto a la caja del tipo de palabra de entrada. Finalmente, se normalizan las cajas para dicho tipo de palabra de entrada, de acuerdo con reglas predeterminadas asociadas con el grupo global a que haya sido asignado dicho tipo de palabra de entrada.

55 De acuerdo con este aspecto del invento, un gran número de tipos de palabra que hayan sido identificados en una base de datos de texto muy largo, son dados de entrada a un ordenador. Los tipos de entrada tal como aparecen en la base de datos del texto, es decir, que se mantienen las cajas a que pertenezcan las letras de los tipos de palabra. Por consiguiente, las muestras de palabra en la base de datos del texto que sean idénticas, excepto por lo que se refiere a la caja a la que pertenezcan una o más letras, serán dadas de entrada como dos tipos de palabra diferentes, mientras que las muestras de palabra contenidas en la base de datos del texto que sean idénticas también en cuanto a la caja a la que pertenezcan las letras, serán dadas de entrada como un tipo de palabra. El método, que se ejecuta de un modo totalmente automático, por medio de un ordenador, hace luego uso tanto de la información local como de la información global relativa a las cajas a que pertenezcan los tipos de palabra. La información local es la de las cajas y las posiciones de las letras que constituyen los tipos de palabra, tal como la caja de la letra inicial y la caja de las letras no iniciales. En cuanto a la información global, el hecho de que haya tipos de palabra que difieran unos de otros solamente con respecto a la caja a la que pertenezcan una o más letras, se usa para el invento. Estos tipos de palabra son variantes de la caja de un tipo de palabra común. Está admitido que, determinando qué variantes de caja diferentes hay para un tipo de palabra común, es posible determinar, con un razonable nivel de certeza, si la diferencia en cuanto a la caja entre las variantes de caja es significativa, o no, y de no serlo a qué variante de la caja deberán ser normalizadas

las variantes de caja. El término “asignado” empleado en “asignado a una serie de grupos locales separados” y en “asignado a una serie de grupos globales separados”, deberá ser interpretado en un sentido general, de modo que no solamente un agrupamiento real del los tipos de palabras de entrada, sino también un reconocimiento más teórico de que hay tipos diferentes de tipos de palabra, en términos de las propiedades locales y globales de interés. Además, entre las reglas predeterminadas se incluyen también las reglas que detectan cuándo no se ha de hacer normalización alguna, lo que ocurre cuando las cajas a que pertenezcan las letras en los tipos de palabra sean consideradas como significativas. De este modo, se preservan las cajas para aquellos tipos de palabras de entrada que no tengan variante de caja alguna, y para aquellos tipos de palabras de entrada que tengan variables de caja para las cuales la diferencia en cuanto a la caja sea considerada significativa, mientras que se normalizan las cajas para los tipos de palabras de entrada para los cuales la diferencia en cuanto a la caja se considere que no es significativa. Una ventaja de este método es la de que se disminuye el número de tipos de palabra que, por ejemplo, deban ser almacenadas en una base de datos. Al mismo tiempo, se preserva la información contenida en la caja cuando se considera que la caja es significativa. Por consiguiente, deberá disminuir el tamaño de la base de datos, con lo cual se disminuirán los costes de la base de datos y se aumentará la velocidad de consulta en la base de datos.

El método es general, independiente del lenguaje, y aplicable a conjuntos de caracteres de lenguas para las cuales la ortografía normal diferencia entre las cajas alta o baja a la que pertenezcan las letras. El método es de aplicación en cuanto a los procedimientos de indización y de consulta en sistemas para recuperación de la información, y a los componentes de análisis de léxico de sistemas para análisis de textos.

En una realización del método de acuerdo con el invento, las variantes en cuanto a la caja de un tipo de palabra de entrada son normalizadas a una variante de caja dada, que es predeterminada para el grupo global dado del tipo de palabra de entrada. Así, para cada grupo global hay una forma de caja que se considera que es la forma normal, y todas las variantes de caja de un tipo de palabra de un grupo global dado son normalizadas a esa forma normal. Esto está basado en el reconocimiento de que los diferentes tipos de palabra, tales como los nombres propios, los acrónimos, los nombres comunes, etc., ocurrirán en un cierto conjunto de variantes de caja en un texto en lenguaje natural, y que el conjunto de variantes de caja de un tipo de palabra que se encuentren en un texto en lenguaje natural largo es indicativo de que tipo de tipo de palabra es el tipo de palabra.

En otra realización del método de acuerdo con el invento, cada tipo de palabra de entrada está asociado con una frecuencia que indica el número de ocurrencias del tipo de palabra de entrada en el texto en lenguaje natural. Las variantes en cuanto a la caja de un tipo de palabra de entrada son luego normalizadas de acuerdo con reglas predeterminadas asociadas con (a) el grupo global a que haya sido asignado el tipo de palabra de entrada, y (b) la frecuencia de las variantes en cuanto a la caja del tipo de palabra de entrada. Por consiguiente, en esta realización la información adicional relativa al número de veces que haya ocurrido cada tipo de palabra en el texto en lenguaje natural se usa para la determinación de si un tipo de palabra de entrada deba ser normalizado, y cómo hacerlo. Por ejemplo, la información relativa a la frecuencia de cada variante de caja de un tipo de palabra puede indicar que la normalización por defecto asociada con el grupo global de las variantes de caja no deberá ser aplicada. Por consiguiente, incluso aunque haya una forma, en términos de caja, que se considere que sea la forma normal a la cual deban ser normalizadas todas las variantes de caja de un tipo de palabra, esto no debe hacerse en algunos casos. Por ejemplo, ese podría ser el caso en que una variante de caja que se considere que sea la forma normal, tenga una frecuencia que sea significativamente menor que la frecuencia de otra variante de caja. Esto está basado en el reconocimiento de que, incluso aunque el conjunto de variantes de caja que tenga un tipo de palabra en un texto en lenguaje natural indica qué tipo de tipo de palabra es el tipo de palabra, hay excepciones a esto. Esas excepciones pueden identificarse considerando también para ello la frecuencia de cada variante de caja. Esto potencia las características de actuación del método, en términos de la corrección de la normalización.

En todavía otra realización del método de acuerdo con el invento, cada tipo de palabra de entrada está asociado con una posición en la frase que indica si el tipo de palabra de entrada ocurrió en una posición interna de una frase, y/o en una posición inicial de la frase en el texto en lenguaje natural. Las variantes de caja e un tipo de palabra de entrada son entonces normalizadas de acuerdo con reglas predeterminadas relativas al grupo global del tipo de palabra de entrada y a las posiciones en la frase de las variantes de caja de dicho tipo de palabra de entrada. También en esta realización, la información relativa a cada grupo específico de variantes de caja puede ser ponderada cuando se determine si debiera ser normalizado un tipo de palabra de entrada, y cómo hacerlo. Por ejemplo, la información relativa a la posición en la frase de cada variante de caja de un tipo de palabra puede indicar que la normalización por defecto asociada con el grupo global de las variantes de caja no deberá ser aplicada. Por consiguiente, incluso aunque haya una forma de caja que se considere que sea la forma normal a la cual deban ser normalizadas todas las variantes de caja de un tipo de palabra, esto no debe hacerse en algunos casos. Por ejemplo, cuando una variante de caja con una letra inicial de la caja alta y otra variante de caja con una letra inicial de la caja baja aparezcan ambas en posiciones internas de frases en el texto natural, ello indica que la diferencia de caja es significativa y que no deberá efectuarse normalización alguna. Esto está basado en el reconocimiento de que, incluso aunque el conjunto de variantes de caja de un tipo de palabra indique de qué clase de tipo de palabra es el tipo de palabra, hay excepciones a esto. Esas excepciones pueden identificarse considerando también para ello en qué posiciones de la frase ha ocurrido cada variante de caja. Esto potencia las características de actuación del método, en términos de preservar las diferencias en cuanto a caja que sean significativas.

Breve descripción de los dibujos

En lo que sigue se ilustra el presente invento, a modo de ejemplo y no de limitación, con referencia a los dibujos que se acompañan, en los cuales:

La Figura 1 es un organigrama de una primera realización de un método de acuerdo con el invento;

La Figura 2 es un organigrama de una segunda realización de un método de acuerdo con el invento; y

La Figura 3 es un organigrama de una tercera realización de un método de acuerdo con el invento.

Descripción detallada del invento

En la Figura 1 se ha representado un organigrama de una primera realización de un método de acuerdo con el invento. Una serie de tipos de palabra se han identificado en una base de datos de texto largo que comprende texto en lenguaje natural. Los tipos de palabra son luego dados de entrada a un ordenador o similar. Los tipos de palabra son dados de entrada tal como aparecen en la base de datos del texto, es decir, que se mantienen las cajas a las que pertenezcan las letras de los tipos de palabra. Por consiguiente, dos muestras de palabra en la base de datos del texto que sean idénticas excepto por lo que se refiere a la caja a que pertenezcan una o más letras, serán dadas de entrada como dos tipos de palabra diferentes, mientras que dos muestras de palabra de la base de datos del texto que sean idénticas también en cuanto a la caja a que pertenezcan las letras, serán dadas de entrada como un tipo de palabra. Los tipos de palabra que difieran entre sí únicamente con respecto a la caja a que pertenezcan una o más letras, serán denominados en lo que sigue como variantes de caja de un tipo de palabra común. De acuerdo con la realización del método, cada tipo de palabra que empiece con un carácter alfabético se asigna a uno de los cuatro grupos locales separados, en el paso 110. Un tipo de palabra se asigna a un grupo local sobre la base de la caja a la que pertenezca la letra inicial del tipo de palabra y la caja a que pertenezcan las letras no iniciales del tipo de palabra. Más concretamente, en el paso 115A, cada tipo de palabra que tenga una letra inicial de la caja alta y ninguna letra no inicial de la caja baja, se asigna a un primer grupo local (LG1). En el paso 115B, cada tipo de palabra que tenga una letra inicial de la caja alta y al menos una letra no inicial de la caja baja se asigna a un segundo grupo local (LG2). En el paso 115C, cada tipo de palabra que tenga una letra inicial de la caja baja y ninguna letra no inicial de la caja alta se asigna a un tercer grupo local (LG3). En el paso 115D, cada tipo de palabra que tenga una letra inicial de la caja baja y al menos una letra no inicial de la caja alta, se asigna a un cuarto grupo local (LG4). Después de la identificación de la información local, es decir, de la información que puede obtenerse simplemente considerando cada tipo de palabra en sus contextos locales de ocurrencia, se reasigna cada tipo de palabra a uno de cuatro grupos globales separados, en el paso 120. Se reasigna un tipo de palabra a un grupo global sobre la base de los grupos locales a los cuales hayan sido asignadas las variantes de caja del tipo de palabra. La identificación de las variantes de caja, es decir, de los tipos de palabra que son iguales entre sí, excepto por lo que se refiere a la caja a que pertenezcan una o más letras, puede hacerse de varios modos diferentes que son evidentes para quien sea experto en la técnica. Cuando se hayan encontrado todas las variantes de caja para un tipo de palabra común, se identifican los grupos a los cuales se han asignado las variantes de caja. Si al menos una variante de caja de un tipo de palabra está asignada al primer grupo local, y al menos una variante de caja del tipo de palabra está asignada al segundo grupo local, y ninguna variante de caja del tipo de palabra está asignada al tercer grupo local, entonces se asigna el tipo de palabra a un primer grupo global (GG1), en el paso 125A. Si al menos una variante de caja está asignada al primer grupo local, y al menos una variante de caja está asignada del tipo de palabra al tercer grupo local, y ninguna variante de caja del tipo de palabra está asignada al segundo grupo local, entonces se asigna el tipo de palabra a un segundo grupo global (GG2), en el paso 125B. Si al menos una variante de caja de un tipo de palabra está asignada al segundo grupo local, y al menos una variante de caja del tipo de palabra está asignada al tercer grupo local, y ninguna variante de caja del tipo de palabra está asignada al primer grupo local, entonces se asigna el tipo de palabra al tercer grupo global (GG3), en el paso 125C. Si al menos una variante de caja de un tipo de palabra está asignada al primer grupo local, y al menos una variante de caja del tipo de palabra está asignada al segundo grupo local, y al menos una variante de caja del tipo de palabra está asignada al tercer grupo local, entonces se asigna el tipo de palabra al cuarto grupo global (GG4), en el paso 125D. De este modo, se identifica la información global, es decir, la información que puede obtenerse analizando para ello la ocurrencia de un tipo de palabra y las variantes de caja del tipo de palabra en una base de datos de todo un texto. Después se usa la información global en el paso 130, cuando se normalizan las cajas para los tipos de palabra de acuerdo con las reglas predeterminadas asociadas con cada grupo global. Las reglas predeterminadas están basadas en el conocimiento empírico del significado de las cajas para los tipos de palabra que ocurren como ciertas variantes de caja en una base de datos del texto, y de la forma normal más probable en términos de cajas para esos tipos de palabra. En esta realización, cada tipo de palabra de entrada asignado al primer grupo global se normaliza de acuerdo con la variante de caja del tipo de palabra que esté asignada al segundo grupo local. Cada tipo de palabra de entrada asignado al segundo grupo global se normaliza de acuerdo con la variante de caja del tipo de palabra que haya sido asignado al tercer grupo local. Cada tipo de palabra de entrada asignado al tercer grupo global se normaliza de acuerdo con la variante de caja del tipo de palabra que esté asignada al tercer grupo local. Cada tipo de palabra de entrada asignado al cuarto grupo global se normaliza de acuerdo con la variante de caja del tipo de palabra que esté asignada al segundo grupo local.

Para que sirvan de ejemplo de la realización descrita con referencia a la Figura 1, se dan en lo que sigue cuatro ejemplos diferentes asociados con los diferentes grupos globales. En un primer ejemplo, se han dado de entrada en el método dos tipos de palabra “CALIFORNIA” y “California”. Estos tipos de palabra son variantes de caja de un tipo de palabra común. La variante de caja “CALIFORNIA” se asigna al primer grupo local, y la variante de caja “California” se asigna al segundo grupo local. Por consiguiente, las dos variantes de caja son ambas asignadas al primer grupo global y serán normalizadas a la variante de caja que esté asignada al segundo grupo local, es decir, a la variante de caja “California”. Una razón para que la normalización sea a la variante de caja que está asignada al segundo grupo local es la de que los tipos de palabra asignados al primer grupo local son usualmente nombres que han ocurrido en una línea de cabecera (primer grupo local) y en texto ordinario (segundo grupo local). Por consiguiente, la diferencia de caja entre las dos variantes de caja no se considera que sea significativa, y las variantes de caja que son asignadas al primer grupo global son normalizadas a la variante de caja que está asignada al segundo grupo local, el cual se considera que es la forma normal.

En un segundo ejemplo, se han dado de entrada en el método los dos tipos de palabra “SUMARIO” y “sumario”. En este ejemplo, la variante de caja “SUMARIO” es asignada al primer grupo local y la variante de caja “sumario” es asignada al tercer grupo local. Por consiguiente, las dos variantes de caja están ambas asignadas al segundo grupo global y serán normalizadas a la variante de caja que esté asignada al tercer grupo local, es decir, a la variante de caja “sumario”. Una razón por la que la normalización ha de ser a la variante de caja que esté asignada al tercer grupo local es la de que los tipos de palabra asignados al segundo grupo global son usualmente nombres comunes, verbos, preposiciones, etc., que han ocurrido en una línea de cabecera (primer grupo local) y en el texto ordinario (tercer grupo local). Por consiguiente, la diferencia de caja entre las dos variantes de caja no se considera que sea significativa, y las variantes de caja que son asignadas al segundo grupo global son normalizadas a la variante de caja que esté asignada al tercer grupo local, el cual se considera que es la forma normal.

En un tercer ejemplo, se han dado de entrada para el método dos tipos de palabra “A menudo” y “a menudo”. En este ejemplo, la variante de caja “A menudo” se asigna al segundo grupo global, y la variante de caja “a menudo” se asigna al tercer grupo local. Por consiguiente, las dos variantes de caja son ambas asignadas al tercer grupo global y serán normalizadas a la variante de caja que esté asignada al tercer grupo local, es decir, a la variante de caja “a menudo”. Una razón por la que la normalización es a la variante de caja que esté asignada al tercer grupo local es la de que los tipos de palabra asignados al tercer grupo global son usualmente nombres comunes, verbos, preposiciones, etc., que han ocurrido tanto en una posición inicial de una frase (segundo grupo local), como en una posición interna de una frase (tercer grupo local). Por consiguiente, se considera que la diferencia de caja no es significativa, y las variantes de caja que son asignadas al tercer grupo global son normalizadas a la variante de caja que esté asignada al tercer grupo local, el cual se considera que es la forma normal.

Finalmente, en un cuarto ejemplo, se han dado de entrada para el método los tres tipos de palabra “ALICE”, “Alice” y “alice”. En este ejemplo, la variante de caja “ALICE” se asigna al primer grupo local, la variante de caja “Alice” se asigna al segundo grupo local, y la variante de caja “alice” se asigna al tercer grupo local. Por consiguiente, las tres variantes de caja serán todas asignadas al cuarto grupo global, y serán normalizadas a la variante de caja que esté asignada al segundo grupo local, es decir, a la variante de caja “Alice”. Una razón por la que la normalización es a la variante de caja que esté asignada al segundo grupo local es la de que los tipos de palabra asignados al cuarto grupo global son usualmente los nombres propios que han ocurrido en una línea de cabecera (primer grupo local), en el texto ordinario (segundo grupo local), y en, por ejemplo, una dirección de un correo electrónico (tercer grupo local). Por consiguiente, se considera que la diferencia de caja no es significativa, y las variantes de caja que son asignadas al cuarto grupo local son normalizadas a la variante de caja que esté asignada al segundo grupo local, el cual se considera que es la forma normal.

En la Figura 2 se ha representado un organigrama de una segunda realización de un método de acuerdo con el invento. Como en la realización descrita con referencia a la Figura 1, una serie de tipos de palabra que hayan sido identificados en una base de datos de texto largo, que comprenda texto en lenguaje natural, son dados de entrada a un ordenador. De acuerdo con esta segunda realización, cada tipo de palabra de entrada está también asociado a una frecuencia que indica el número de ocurrencias del tipo de palabra de entrada en el texto en lenguaje natural. Los pasos de asignación de los tipos de palabra a los grupos locales y a los grupos globales no difieren de la primera realización. Por consiguiente, no se describen aquí con mayor detalle. La diferencia con respecto a la primera realización se produce en el paso 230, en donde se normaliza la caja. En esta realización, la caja de un tipo de palabra de entrada se normaliza de acuerdo con reglas predeterminadas, asociadas con (1) el grupo global del tipo de palabra de entrada, y (2) la frecuencia de cada variante de caja del tipo de palabra de entrada. Más concretamente, para los tipos de palabras de entrada que son asignados al primer grupo global, al segundo grupo global, o al tercer grupo global, se normaliza la caja a la variante de caja del tipo de palabra de entrada que tiene la máxima frecuencia. En el caso de enlaces de baja frecuencia entre las variantes de caja, se normaliza la caja a la forma normal por defecto del respectivo grupo global. En los casos de enlaces de alta frecuencia, la diferencia entre variantes de caja es significativa, y se bloquea la normalización. Para los tipos de palabras de entrada que sean asignados al cuarto grupo global hay dos alternativas. Si la frecuencia de la variante de caja que esté asignada al segundo grupo local es mayor que la frecuencia de la variante de caja que esté asignada al primer grupo global, entonces se normaliza la caja a la variante de caja que esté asignada al segundo grupo local. Si la frecuencia de la variante de caja que esté asignada al segundo grupo local es menor que la frecuencia de la variante de caja que esté asignada al primer grupo local, entonces se normaliza la caja del tipo de palabra de entrada a la variante de caja que esté asignada al primer grupo local. El tratamiento del segundo par consistente en la variante de caja asignada al primer grupo local y la variante de caja asignada al tercer grupo local, es análogo.

Para que sirvan de ejemplo de la realización descrita con referencia a la Figura 2, se dan en lo que sigue dos ejemplos diferentes asociados con grupos locales diferentes. En un primer ejemplo, se han dado de entrada para el método dos tipos de palabra “UNESCO” y “Unesco”. Estos tipos de palabra son variantes de un tipo de palabra común. La variante de caja “UNESCO” se asigna al primer grupo local, y la variante de caja “Unesco” se asigna al segundo grupo local. En este ejemplo, las dos variantes de caja están asociadas con sus respectivas frecuencias, y la frecuencia de la variante de caja “UNESCO” es mayor que la frecuencia de la variante de caja “Unesco”. Por consiguiente, incluso aunque las variantes de caja estén asignadas al primer grupo local, para el cual la forma normal por defecto es la variante de caja asignada al segundo grupo local, serán normalizadas en cambio a la variante de caja que esté asignada al primer grupo local, es decir, a la variante de caja “UNESCO”. Esto se debe al hecho de que las respectivas frecuencias de las variantes de caja pesan más que las reglas predeterminadas asociadas con el grupo global. En este ejemplo, las variantes de caja son acrónimos para los cuales la variante de caja asignada al primer grupo local se considera que es la forma normal.

Como un segundo ejemplo, se han dado de entrada para el método los tres tipos de palabra “ATM”, “Atm”, y “atm”. En este ejemplo, las tres variantes de caja están asignadas con sus frecuencias, y la frecuencia de la variante de caja “ATM” es mayor que la frecuencia de la variante de caja “Atm”. Así, incluso aunque las variantes de caja sean asignadas al cuarto grupo global, serán normalizadas a la variante de caja que esté asignada al primer grupo local, es decir, a la primera variante de caja “ATM”. Esto es debido al hecho de que las respectivas frecuencias de las variantes de caja pesan más que las reglas predeterminadas asociadas con el grupo global. En este ejemplo, las variantes de caja son acrónimos, para los cuales se considera que la variante de caja asignada al primer grupo global es la forma normal.

En la Figura 3 se ha representado un organigrama de una tercera realización de un método de acuerdo con el invento. Como en la realización descrita con referencia a la Figura 1, una serie de tipos de palabra que han sido identificados en una base de datos de texto largo, que comprende texto en lenguaje natural, son dados de entrada en un ordenador. De acuerdo con esta tercera realización, cada tipo de palabra de entrada está también asociado a una posición en una frase, indicadora de si el tipo de palabra de entrada ha ocurrido en una posición interna de una frase y/o en una posición inicial de una frase, en el texto en lenguaje natural. Los pasos de asignar los tipos de palabra a los grupos locales y a los grupos globales no difieren de los de la primera realización. Por consiguiente, no se describen aquí con mayor detalle. La diferencia con respecto a la primera realización tiene lugar en el paso 330, en donde la caja es normalizada. En esta realización, la caja de cada tipo de palabra de entrada se normaliza de acuerdo con reglas predeterminadas asociadas con (1) el grupo global del tipo de palabra de entrada, y (2) la posición en la frase de cada variante de caja del tipo de palabra de entrada. Más concretamente, las cajas de un tipo de palabra de entrada son normalizadas de acuerdo con las mismas reglas que en la realización descrita con referencia a la Figura 1, con dos excepciones. Si se asigna un tipo de palabra de entrada al tercer o al cuarto grupo global, no se efectuará la normalización si la variante de caja asignada al segundo grupo global está asociada con una posición en la frase que indique que el tipo de palabra de entrada ocurrió en una posición interna de la frase, en el texto en lenguaje natural.

Para que sirva de ejemplo de la realización descrita con referencia a la Figura 3, se dará en lo que sigue un ejemplo. En el ejemplo se han dado de entrada para el método los dos tipos de palabra “Bill” y “bill”. Estos tipos de palabra son variantes de caja de un tipo de palabra común. La variante de caja “Bill” se asigna al segundo grupo local y la variante de caja “bill” se asigna al tercer grupo local. En este ejemplo, las dos variantes de caja son asociadas a información de posición en la frase. La información de posición en la frase, acerca de la variante de caja “Bill”, indica que la variante de caja se ha producido en una posición inicial de la frase, y también en una posición interna de la frase, en el texto en lenguaje natural. Por consiguiente, no se efectuará normalización alguna para las variantes de caja “Bill” y “bill”. Esto se debe al hecho de que las posiciones en la frase de esas variantes de caja indican que no deberán usarse las reglas predeterminadas asociadas con el grupo global. En este ejemplo, la diferencia de cajas entre las variantes de caja conduce información que deberá ser preservada. Más concretamente, la variante de caja “Bill” pudiera ser tanto un nombre propio como un nombre común. Si, por otra parte, han sido dados de entrada para el método los dos tipos de palabra “Coche” y “coche”, y la información de posición en la frase acerca de la variante de caja “Coche” indica que esa variante de caja solamente se produce en una posición inicial de la frase, mientras que la información de posición de la frase acerca de la variante de caja “coche” indica que esa variante de caja solamente ocurre en una posición interna de la frase, entonces se usarían las reglas de la realización descrita con referencia a la Figura 1, y se normalizan las dos variantes de caja a la variante de caja que esté asignada al tercer grupo global, es decir, a la variante de caja “coche”.

Las realizaciones descritas en lo que antecede pueden ser implementadas en un programa de ordenador que comprenda instrucciones ejecutables por ordenador, para ejecutar los pasos. El programa de ordenador puede entonces ser almacenado en cualquier medio legible por ordenador, y las realizaciones pueden entonces ser realizadas por medio de un ordenador para fines generales, con acceso a esos medios. Además, las realizaciones pueden ser implementadas también directamente, en equipo físico, tal como en uno más procesadores de ordenador que estén dispuestos para ejecutar los pasos.

REIVINDICACIONES

1. Un método para diferenciar automáticamente distinciones significativas de distinciones no significativas de caja alta y caja baja en una serie de tipos de palabras de entrada de un texto en lenguaje natural por medio de un ordenador, **caracterizado** por los pasos de:

asignar (110) un tipo de palabra de entrada a uno de una serie de grupos locales separados (LG1, LG2, LG3, LG4) en base a la caja y a la posición de las letras que constituyan el tipo de palabra de entrada;

asignar (120) dicho tipo de palabra de entrada a uno de una serie de grupos globales separados (GG1, GG2, GG3, GG4) en base a qué variantes de caja de grupos locales de los tipos de palabras de entrada estén asignados; y

normalizar (139) las cajas para dicho tipo de palabra de entrada de acuerdo con reglas predeterminadas asociadas con el grupo global al que esté asignado dicho tipo de palabra de entrada.

2. El método de acuerdo con la reivindicación 1, en el que el paso de normalizar las cajas comprende el paso de:

normalizar las cajas de dicho tipo de palabra de entrada de acuerdo con las cajas de la variante de caja de dicho tipo de palabra de entrada que esté asignada a un grupo local que esté predeterminado para el grupo global al que esté asignado dicho tipo de palabra de entrada.

3. El método de acuerdo con la reivindicación 1 ó 2, en el que el paso de asignar un tipo de palabra de entrada a uno de una serie de grupos locales separados comprende el paso de:

asignar un tipo de palabra de entrada a uno de una serie de grupos locales separados en base a la caja de la letra inicial de dicho tipo de palabra de entrada y a la caja de las letras no iniciales de dicho tipo de palabra de entrada.

4. El método de acuerdo con la reivindicación 1 ó 2, en el que el paso de asignar un tipo de palabra de entrada a uno de una serie de grupos locales separados, comprende los pasos de:

asignar un tipo de palabra de entrada a una serie de grupos locales separados en base a la caja de la letra inicial de dicho tipo de palabra de entrada y a si hay, o no, alguna letra no inicial de dicho tipo de palabra de entrada que sea de una caja diferente a la de la letra inicial del tipo de palabra de entrada.

5. El método de acuerdo con la reivindicación 1 ó 2, en el que el paso de asignar un tipo de palabra de entrada a uno de una serie de grupos locales separados comprende los pasos de:

asignar a un primer grupo local un tipo de palabra de entrada que tenga una letra inicial de la caja alta y ninguna letra no inicial de la caja baja;

asignar a un segundo grupo local un tipo de palabra de entrada que tenga una letra inicial de la caja alta y al menos una letra no inicial de la caja baja;

asignar a un tercer grupo local un tipo de palabra de entrada que tenga una letra inicial de la caja baja y ninguna letra no inicial de la caja alta; y

asignar a un cuarto grupo local un tipo de palabra de entrada que tenga una letra inicial de la caja baja y al menos una letra de la caja alta.

6. El método de acuerdo con la reivindicación 5, en el que el paso de asignar dicho tipo de palabra de entrada a grupos globales separados, comprende los pasos de:

asignar dicho tipo de palabra de entrada a un primer grupo global si una variante de caja de dicho tipo de palabra de entrada está asignada a dicho primer grupo local, una variante de caja de dicho tipo de palabra de entrada está asignada a dicho segundo grupo local, y no hay ninguna variante de caja de dicho tipo de palabra de entrada que esté asignada a dicho tercer grupo local;

asignar dicho tipo de palabra de entrada a un segundo grupo global, si una variante de caja de dicho tipo de palabra de entrada está asignada a dicho primer grupo local, una variante de caja de dicho tipo de palabra de entrada está asignada a dicho tercer grupo local, y no hay ninguna variante de caja de dicho tipo de palabra de entrada que esté asignada a dicho segundo grupo local;

asignar cada tipo de palabra de entrada a un tercer grupo global, si una variante de caja de dicho tipo de palabra de entrada está asignada a dicho segundo grupo local, una variante de caja de dicho tope está asignada a dicho tercer grupo local, y no hay ninguna variante de caja de dicho tipo de palabra de entrada que esté asignada a dicho primer grupo local; y

ES 2 316 474 T3

asignar cada tipo de palabra de entrada a un cuarto grupo global, si una variante de caja de dicho tipo de palabra de entrada está asignada a dicho primer grupo local, una variante de caja de dicho tipo de palabra de entrada está asignada a dicho segundo grupo local, y una variante de caja de dicho tipo de palabra de entrada está asignada a dicho tercer grupo local.

5

7. El método de acuerdo con la reivindicación 6, en el que el paso de normalizar las cajas comprende los pasos de:

10 normalizar las cajas de dicho tipo de palabra de entrada de acuerdo con las cajas de la variante de caja de dicho tipo de palabra de entrada que esté asignada a dicho segundo grupo local, si dicho tipo de palabra de entrada está asignado a dicho primer grupo global;

15 normalizar las cajas de dicho tipo de palabra de entrada de acuerdo con las cajas de la variante de caja de dicho tipo de palabra de entrada que esté asignada a dicho tercer grupo local, si dicho tipo de palabra de entrada está asignado a dicho segundo grupo local;

20 normalizar las cajas de dicho tipo de palabra de entrada de acuerdo con las cajas de la variante de caja de dicho tipo de palabra de entrada que esté asignada a dicho tercer grupo local, si dicho tipo de palabra de entrada está asignado a dicho tercer grupo global; y

normalizar las cajas de dicho tipo de palabra de entrada de acuerdo con las cajas de la variante de caja de dicho tipo de palabra de entrada que esté asignada a dicho segundo grupo global, si dicho tipo de palabra de entrada está asignado a dicho cuarto grupo global.

25 8. El método de acuerdo con la reivindicación 1, en el que cada uno de los tipos de palabras de entrada está asociado con un indicador de frecuencia que indique el número de ocurrencias del tipo de palabra de entrada en dicho texto en lenguaje natural, y en el que el paso de normalizar comprende el paso de:

30 normalizar las cajas para dicho tipo de palabra de entrada de acuerdo con reglas predeterminadas asociadas con el grupo global al que esté asignado dicho tipo de palabra de entrada y con los indicadores de frecuencia con los que estén asociadas las variantes de caja de dicho tipo de palabra de entrada.

35 9. El método de acuerdo con la reivindicación 6, en el que cada uno de los tipos de palabras de entrada está asociado con un indicador de frecuencia que indica el número de ocurrencias del tipo de palabra de entrada en dicho texto en lenguaje natural, y en el que el paso de normalizar comprende el paso de:

40 normalizar las cajas de dicho tipo de palabra de entrada de acuerdo con las cajas de la variante de caja de dicho tipo de palabra de entrada que esté asociada con el indicador de máxima frecuencia, si dicho tipo de palabra de entrada está asignado a dicho primer grupo global, a dicho segundo grupo global, o a dicho tercer grupo global;

45 normalizar las cajas de dicho tipo de palabra de entrada de acuerdo con las cajas de la variante de caja de dicho tipo de palabra de entrada que esté asignada al segundo grupo local, si dicho tipo de palabra de entrada está asignado a dicho cuarto grupo global y la variante de caja de dicho tipo de palabra de entrada que esté asignada a dicho segundo grupo local está asociada con un indicador de frecuencia que presente una indicación mayor que la del indicador de frecuencia con el que esté asociada la variante de caja de dicho tipo de palabra de entrada que está asignada al primer grupo local; y

50 normalizar las cajas de dicho tipo de palabra de entrada de acuerdo con las cajas de la variante de caja de dicho tipo de palabra de entrada que esté asignada al primer grupo local, si dicho tipo de palabra de entrada está asignado a dicho cuarto grupo global y la variante de caja de dicho tipo de palabra de entrada que esté asignada a dicho segundo grupo local está asociada con un indicador de frecuencia que presente una indicación menor que la del indicador de frecuencia con el que esté asociada la variante de caja de dicho tipo de palabra de entrada que está asignada al primer grupo local.

55 10. El método de acuerdo con la reivindicación 1, en el que cada uno de dichos tipos de palabras de entrada está asociado con un indicador de posición en la frase que indica si el tipo de palabra de entrada ocurrió en una posición interna de una frase y/o en una posición inicial de una frase en dicho texto en lenguaje natural, y en el que el paso de normalizar comprende el paso de:

60 normalizar las cajas para dicho tipo de palabra de entrada de acuerdo con reglas predeterminadas asociadas con el grupo global al que esté asignado dicho tipo de palabra de entrada y con el indicador de la posición en la frase al que estén asociadas las variantes de caja de dicho tipo de palabra de entrada.

65 11. El método de acuerdo con la reivindicación 6, en el que cada uno de dichos tipos de palabras de entrada está asociado con un indicador de posición en la frase que indique si el tipo de palabra de entrada ocurrió en una posición interna de una frase y/o en dicho texto en lenguaje natural, y en el que el paso de normalizar comprende los pasos de:

ES 2 316 474 T3

normalizar las cajas de dicho tipo de palabra de entrada de acuerdo con las cajas de la variante de caja de dicho tipo de palabra de entrada que esté asignada a dicho segundo grupo local, si dicho tipo de palabra de entrada está asignado a dicho primer grupo global;

- 5 normalizar las cajas de dicho tipo de palabra de entrada de acuerdo con las cajas de la variante de caja de dicho tipo de palabra de entrada que esté asignada a dicho tercer grupo local, si dicho tipo de palabra de entrada está asignado a dicho segundo grupo global;

- 10 normalizar las cajas de dicho tipo de palabra de entrada de acuerdo con las cajas de la variante de caja de dicho tipo de palabra de entrada que esté asignada a dicho tercer grupo local, si dicho tipo de palabra de entrada está asignado a dicho tercer grupo global y la variante de caja de dicho tipo de palabra de entrada que esté asignada a dicho segundo grupo local no está asociada con un indicador de posición en la frase que indique que el tipo de palabra de entrada ocurrió en una posición interna de una frase en dicho texto en lenguaje natural; y

- 15 normalizar las cajas de dicho tipo de palabra de entrada de acuerdo con las cajas de la variante de caja de dicho tipo de palabra de entrada que esté asignada a dicho segundo grupo local, si dicho tipo de palabra de entrada está asignado a dicho cuarto grupo global y la variante de caja de dicho tipo de palabra de entrada que esté asignada a dicho segundo grupo local no está asociada con un indicador de posición en la frase que indique que el tipo de palabra de entrada ocurrió en una posición interna de una frase en dicho texto en lenguaje natural.

- 20 12. El método de acuerdo con una cualquiera de las reivindicaciones 1 - 11, que comprende además el paso de:

- almacenar dichos tipos de palabras de entrada con cajas normalizadas en unos medios de almacenamiento electrónico.

- 25 13. Un procesador de ordenador dispuesto para ejecutar los pasos reseñados en cualquiera de las reivindicaciones 1 - 12.

- 30 14. Un medio legible por ordenador que tiene instrucciones ejecutables por ordenador para que un ordenador ejecute los pasos indicados en cualquiera de las reivindicaciones 1-12.

- 15 15. Un programa de ordenador que comprende instrucciones ejecutables por ordenador para que un ordenador ejecute los pasos indicados en cualquiera de las reivindicaciones 1-12.

35

40

45

50

55

60

65

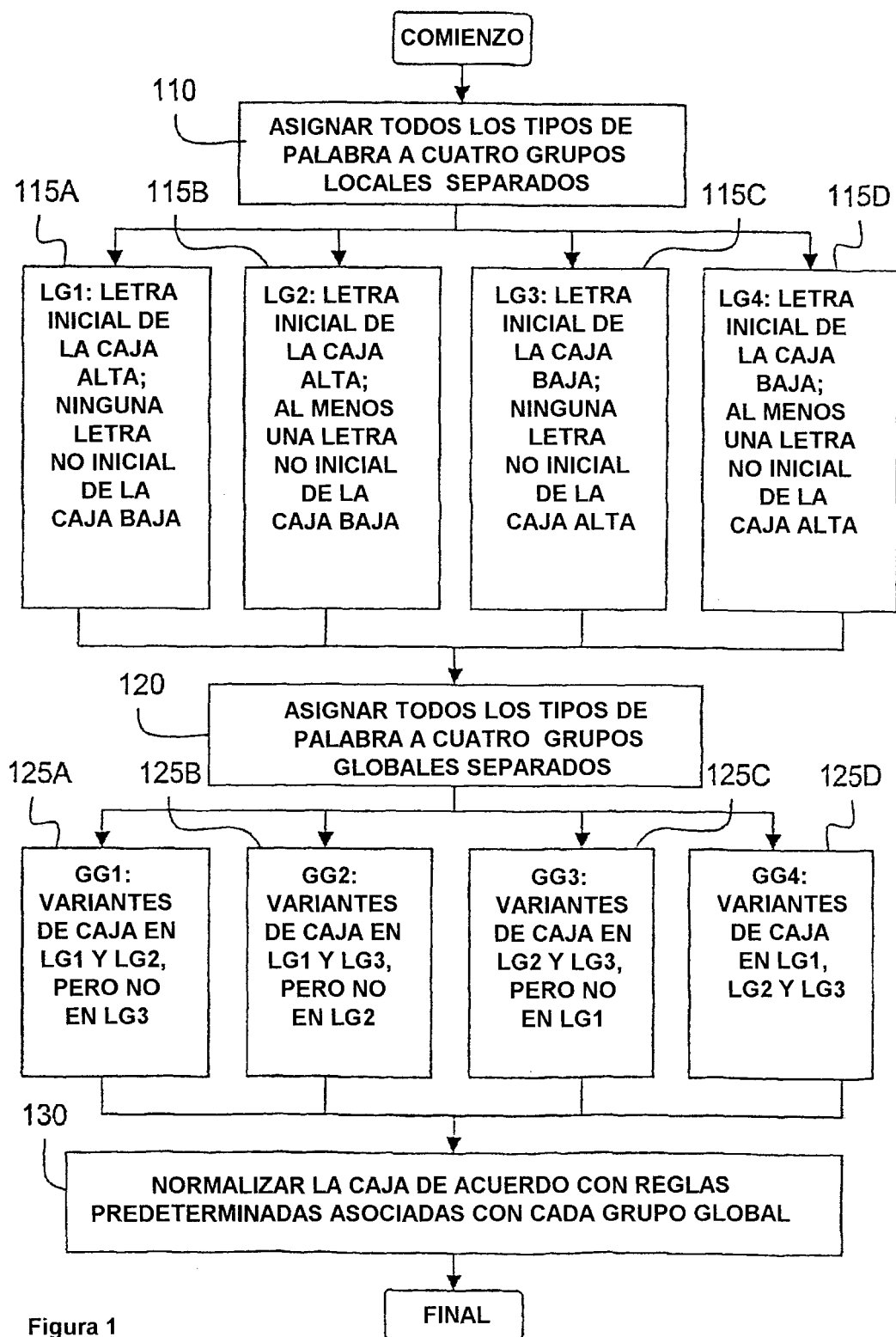


Figura 1

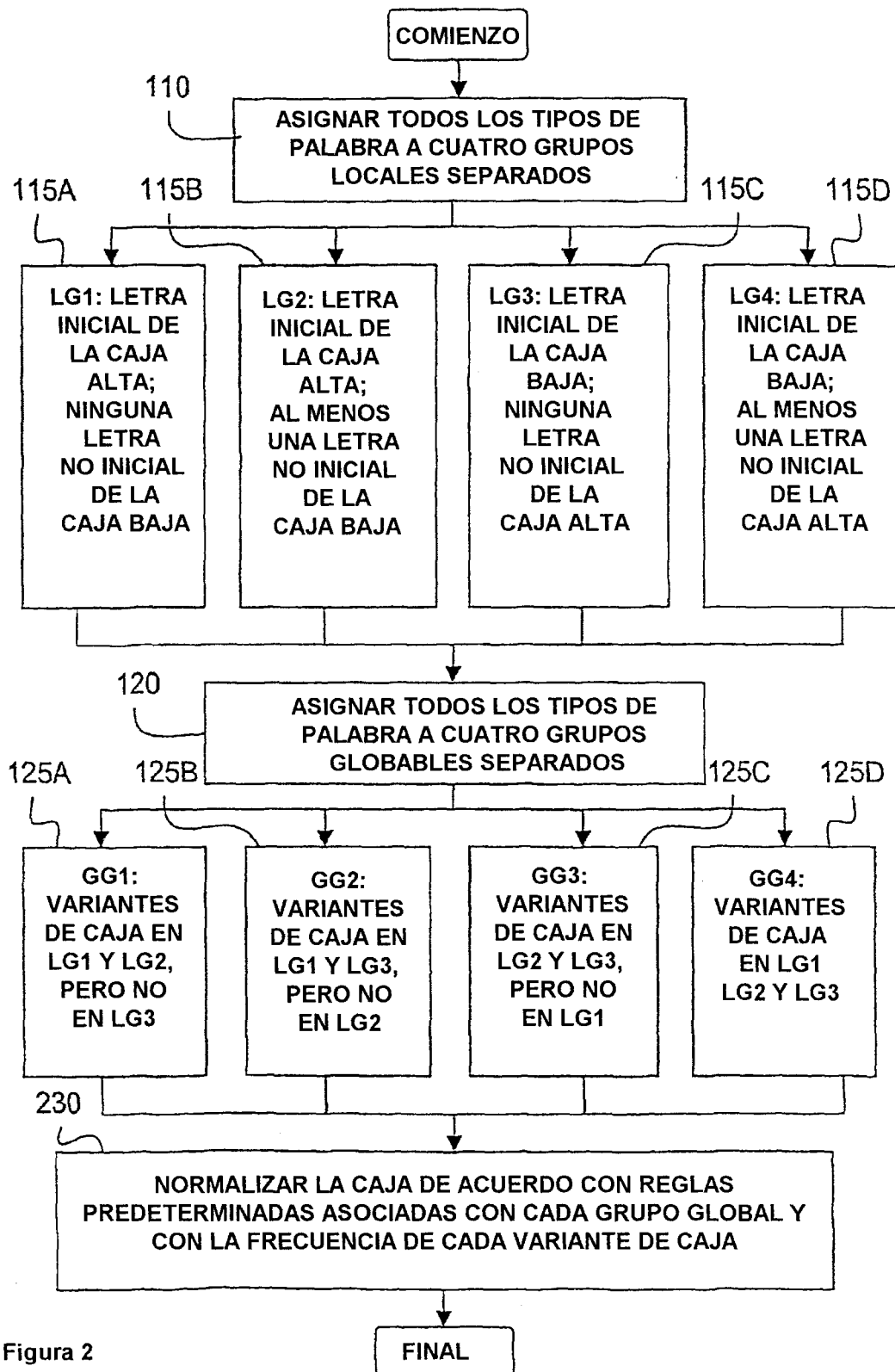


Figura 2

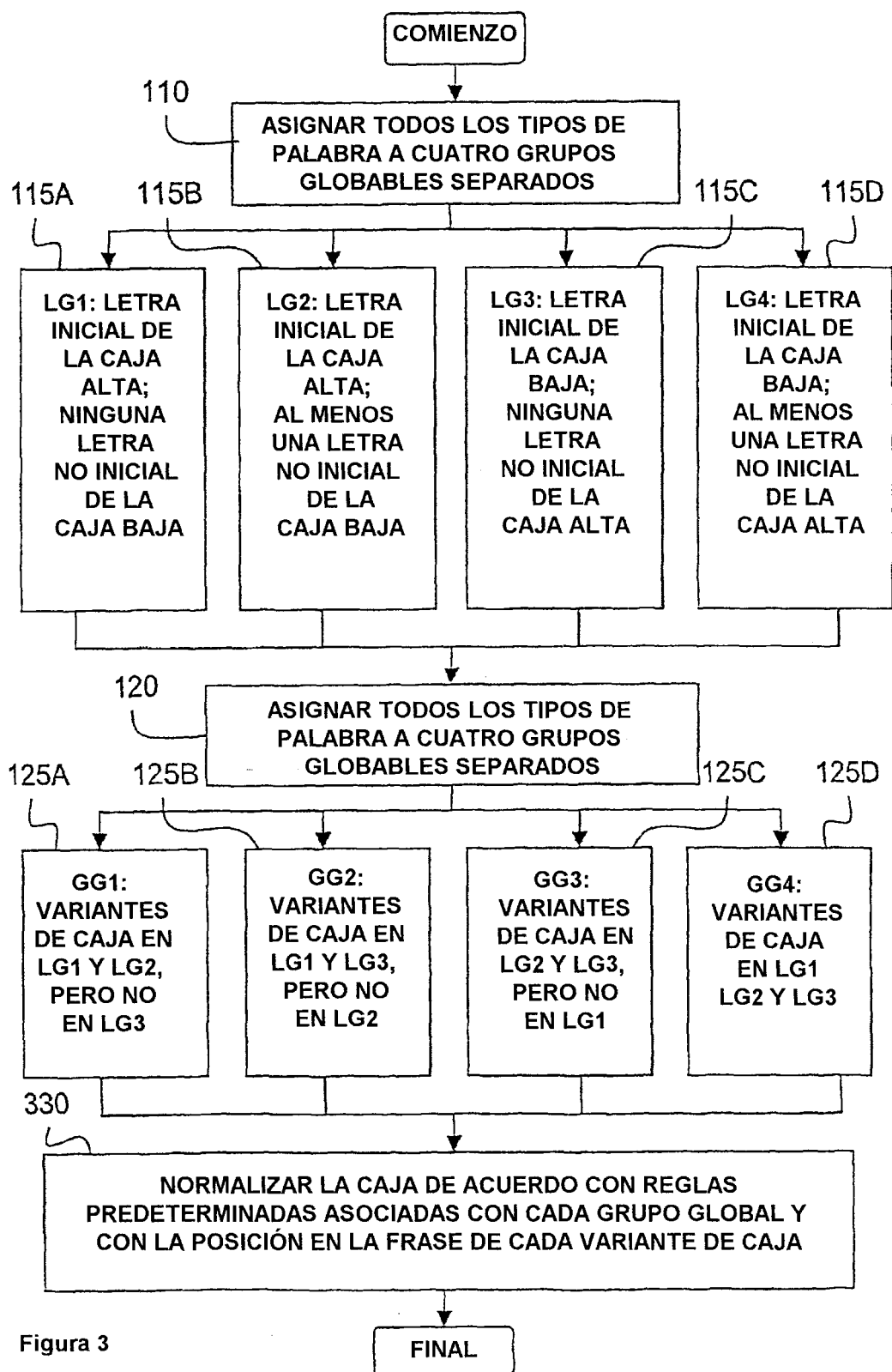


Figura 3