(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 386 631**

(51) Int. CI.:
G06F 17/27 (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Número de solicitud europea: **03008804 .1**
(96) Fecha de presentación: **23.04.2003**
(97) Número de publicación de la solicitud: **1367501**
(97) Fecha de publicación de la solicitud: **03.12.2003**

(54) Título: **Léxico con datos divididos en secciones y procedimiento de uso del mismo**

(30) Prioridad:
30.04.2002 US 137456

(45) Fecha de publicación de la mención BOPI:
24.08.2012

(45) Fecha de la publicación del folleto de la patente:
24.08.2012

(73) Titular/es:
MICROSOFT CORPORATION
ONE MICROSOFT WAY
REDMOND, WASHINGTON 98052-6399, US

(72) Inventor/es:
Finnegan, James P.;
Huttenhower, Curis E.;
Potter, Douglas W. y
Powell, Kevin R.

(74) Agente/Representante:
Carpintero López, Mario

DESCRIPCIÓN

Léxico con datos divididos en secciones y procedimiento de uso del mismo

Antecedentes de la invención

La presente invención versa acerca del tratamiento del lenguaje o de textos. Más en particular, la presente invención versa acerca de una estructura de datos mejorada para almacenar un léxico y de un procedimiento mejorado de uso del mismo.

El tratamiento del lenguaje o de textos abarca muchos tipos de sistemas. Por ejemplo, analizadores sintácticos, verificadores ortográficos, verificadores gramaticales, divisores silábicos, procesadores de lenguaje natural o sistemas de comprensión y sistemas de traducción automática son solo algunos de los tipos de sistemas que se encuentran dentro de esta amplia categoría.

Un componente común e importante de muchos sistemas de tratamiento del lenguaje o de textos es el léxico. Generalmente, el léxico es una estructura de datos que contiene información sobre palabras. Por ejemplo, el léxico puede almacenar indicaciones de información sintáctica y semántica. Ejemplos incluyen si la palabra es un sustantivo, un verbo, un adjetivo, etc. Además, en el léxico pueden guardarse diferentes tipos de información lingüística. A menudo es útil almacenar otra información útil para el tipo particular de tratamiento del lenguaje, tal como almacenar información sobre la palabra que contribuya a su análisis sintáctico. En otros léxicos distintos, pueden resultar útiles indicaciones de si la palabra es un nombre propio, una ubicación geográfica, etc.

En operación, tras recibir una cadena de palabras de entrada, el sistema de tratamiento del lenguaje o de textos accede al léxico para obtener la información almacenada con respecto a cada una de las palabras. Habiendo recogido la información sobre cada una de las palabras de la cadena de entrada, el sistema de tratamiento del lenguaje o de textos procesa la cadena de entrada, lo que puede incluir resolver cualquier ambigüedad que pueda existir con base en la información de las palabras. Por ejemplo, en un sistema de tratamiento de lenguaje natural, el léxico asigna partes de la oración a cada una de las palabras de la cadena de entrada. A continuación, un analizador sintáctico decide cuáles de las asignaciones de partes de la oración son apropiadas y construye una estructura a partir de la cadena de entrada, que puede pasarse entonces a un componente semántico para su interpretación.

Comúnmente, cada entrada del léxico comprende un solo objeto binario grande. Aunque la información es accesible, este formato no permite fácilmente un acceso localizado a información léxica usada comúnmente sin tener que leer la entrada completa. Si hay que leer del léxico toda la información perteneciente a una entrada de palabra, hacen falta más memoria y tiempo de procesamiento, particularmente si solo se necesita una parte pequeña de la información para la entrada de la palabra.

También resulta difícil modificar información léxica o añadirla. Específicamente, para modificar la información del léxico o añadir información adicional, el autor del léxico debe replicar todos los bits, los atributos u otra información dentro de cada entrada, luego modificar la información deseada o añadir a la misma mientras se mantiene la integridad y la organización de una estructura de datos muy compleja.

Existe, así, la necesidad de una estructura mejorada de datos de léxico que aborde una, algunas o todas las desventajas presentadas en lo que antecede.

El documento EP-A2-0 539 965 da a conocer un diccionario electrónico usado típicamente en sistemas de traducción. El diccionario electrónico incluye un fichero de cabeceras, un fichero de punteros que tiene una porción de almacenamiento de banderas de corrección y una porción de almacenamiento de punteros, un fichero de información de palabras y un fichero de corrección de la información de palabras.

El documento US-A-6 138 087 da a conocer un sistema para almacenar y recuperar experiencia y conocimiento de lenguaje natural a través de procedimientos y un aparato. El diccionario está organizado de tal forma que cada palabra almacenada contiene: una entrada de texto que corresponde a la palabra; un número de representación que se usa para representar la palabra de texto; un conjunto de conjuntos de palabras de sintaxis, cada uno con una parte asociada de la oración; una dirección para un proceso de selección de función del conjunto de palabras y un código de función asociado; una lista de anomalías gramaticales asociadas divididas por conjuntos de palabras; punteros a tablas comunes para seleccionar códigos de inflexión y/o a otras tablas comunes relacionadas con un conjunto de palabras, tales como preposiciones de modificación de un sustantivo concreto.

C. J. WELLS, L. J. EVETT, P. E. WHITBY Y R. J. WHITROW: “Fast dictionary look-up for contextual word recognition” PATTERN RECOGNITION [en línea], vol. 23, nº 5, 1990, páginas 501-508, XP002407899 Gran Bretaña, recuperado de Internet: URL: http://portal.acm.org/citation.cfm?id=83050> se refiere a una consulta rápida de diccionario para el reconocimiento contextual de palabras en la que el léxico está representado como un trie.

A. NTOULAS ET AL.: “Use of a morphosyntactic lexicon as the basis for the implementation of the Greek Wordnet”, LECTURE NOTES IN COMPUTER SCIENCE, 2000, XP002407900 Springer Berlín/Heidelberg se refiere al uso de

un léxico morfosintáctico como base para la implementación de la WordNet Griega, en la que el léxico está representado como un trie.

FREDKIN E: "TRIE MEMORY", COMMUNICATIONS OF THE ASSOCIATION FOR COMPUTING MACHINERY, ACM, NUEVA YORK, NY, EE. UU., vol. 3, nº 9, agosto de 1960 (1960-08), páginas 490-499, XP002271883 ISSN: 0001-0782 se refiere a memoria de tipo trie como forma de almacenar y recuperar información.

El objeto de la presente invención es proporcionar un procedimiento mejorado de obtención de información de palabras accediendo a un léxico, así como un correspondiente medio legible por ordenador.

Este objeto se resuelve por medio de la materia de las reivindicaciones independientes.

Las reivindicaciones dependientes definen realizaciones preferentes.

Un aspecto de la presente invención es un léxico de palabras almacenado en un medio legible por ordenador que tiene información de palabras adaptada para su uso en un sistema de tratamiento del lenguaje. El léxico incluye una sección de lista de palabras para almacenar una pluralidad de palabras y una pluralidad de secciones de datos para almacenar información de palabras de la pluralidad de palabras. Las varias secciones de datos están separadas entre sí y de la sección de lista de palabras. Para acceder a la información de palabras, se proporciona una sección de índices que almacena punteros que apuntan a datos en la pluralidad de secciones de datos. Una identificación de qué puntero usar es una función de la correspondiente palabra en la sección de lista de palabras.

La estructura mejorada del léxico permite flexibilidad y eficiencias no disponibles previamente. La sección de índices y la pluralidad de secciones de datos permiten que el léxico se adapte para amoldarse a las necesidad de un sistema de tratamiento del lenguaje según los recursos disponibles del ordenador. En una realización adicional, la estructura del léxico permite que la información de palabras se clasifique o agrupe con base en una clasificación. Por ejemplo, la clasificación puede basarse en la parte de la oración de la entrada de la palabra, tal como si la entrada de la palabra puede ser un sustantivo, un verbo, un adjetivo, etc. La información de la palabra puede ser objeto de acceso selectivo en función de la clasificación. En el aspecto ejemplar, se proporcionan indicadores en punteros para indicar la clasificación de la correspondiente información de la palabra.

Otros aspectos de la presente invención incluyen un procedimiento implementado por ordenador para almacenar información de palabras en una pluralidad de secciones de datos, para almacenar información de punteros en la sección de índices y para almacenar la lista de palabras en la sección de lista de palabras, teniendo información la lista de palabras para identificar los correspondientes punteros relacionados con la palabra seleccionado. De forma similar, otro aspecto es acceder a la información de palabras usando la estructura de datos para el léxico proporcionada en lo que antecede.

La estructura del léxico descrita en lo que antecede es particularmente útil cuando es deseable obtener información de varios léxicos, lo cual es otro aspecto adicional de la presente invención. En general, los datos de múltiples léxicos para una entrada de palabra particular pueden combinarse, ignorarse o seleccionarse según se desee accediendo de forma selectiva a secciones de datos de cada uno de los léxicos.

Breve descripción de los dibujos

La Fig. 1 es un diagrama de bloques de un sistema de tratamiento del lenguaje o de textos.
La Fig. 2 es un diagrama de bloques de un entorno ejemplar para implementar la presente invención.
La Fig. 3 es una representación gráfica de un entorno ejemplar para implementar la presente invención.
La Fig. 4 es una representación gráfica de la recuperación de información en una pluralidad de léxicos o de acceso a la misma.

Descripción detallada del aspecto ilustrativo

La Fig. 1 ilustra en general un sistema 10 de tratamiento del lenguaje o de textos que recibe una entrada 12 de lenguaje, comúnmente en forma de cadena de texto, y procesa la entrada 12 de lenguaje para proporcionar una salida 14 de lenguaje, también comúnmente en forma de una cadena de texto. Por ejemplo, el sistema 10 de tratamiento del lenguaje puede estar implementado como un verificador ortográfico, un verificador gramatical o un procesador de lenguaje natural, por nombrar solo algunos. Según apreciarán los expertos en la técnica, el sistema 10 de tratamiento del lenguaje puede ser una aplicación dedicada o un módulo o componente accesible por otro sistema o incluido en el mismo.

En general, el sistema de tratamiento del lenguaje incluye un analizador 20 de textos y un léxico 22. El analizador 20 de textos representa esquemáticamente componentes o módulos que reciben la entrada 12, acceden y obtienen información del léxico 22 y procesar la información de palabras para proporcionar la salida 14. Un aspecto de la presente invención es una estructura de datos mejorada para el léxico 22 para proporcionar de forma eficiente la información necesaria al analizador 20 de textos según requiera su aplicación. En vista de que el léxico 22 es un componente separado que puede ser usando en muchos sistemas de tratamiento del lenguaje y con muchas formas de analizadores de textos, se describirá la interacción general del analizador 20 de textos con el léxico 22, pero no

se describirán los detalles específicos relativos a las diversos formas de los analizadores de textos, porque tal descripción no es necesaria para una comprensión de la presente invención.

Antes de una exposición detallada adicional de la presente invención, puede resultar útil una visión general de un entorno operativo. La FIG. 2 ilustra un ejemplo de un entorno 50 de un sistema informático adecuado en el que la invención puede ser implementada. El entorno 50 de un sistema informático es solo un ejemplo de un entorno informático adecuado y no se pretende sugerir ninguna limitación en cuanto al alcance de uso o la funcionalidad de la invención. ni debería interpretarse que el entorno informático 50 tenga ninguna dependencia ni requerimiento en cuanto a uno cualquiera o una combinación de componentes ilustrados en el entorno ejemplar 50 de un sistema informático.

La invención es operativa con numerosos entornos o configuraciones adicionales de sistemas informáticos de uso general o uso especial. Ejemplos de sistemas, entornos y/o configuraciones informáticas bien conocidos que pueden ser adecuados para su uso con la invención incluyen, sin limitación, ordenadores personales, ordenadores servidores, dispositivos de mano o portátiles, sistemas multiprocesadores, sistemas basados en microprocesadores, decodificadores, electrónica programable de consumo, PC en red, microordenadores, ordenadores centrales, entornos informáticos distribuidos que incluyan cualquiera de los sistemas o dispositivos anteriores, y similares.

La invención puede ser descrita en el contexto general de instrucciones ejecutables por ordenador, tales como módulos de programa que son ejecutados por un ordenador. En general, los módulos de programa incluyen rutinas, programas, objetos, componentes, estructuras de datos, etc., que llevan a cabo tareas particulares o implementan tipos particulares de datos abstractos. La invención también puede ser puesta en práctica en entornos informáticos distribuidos, en los que las tareas son realizadas por dispositivos remotos de procesamiento que están enlazados a través de una red de comunicaciones. En un entorno informático distribuido, los módulos de programa pueden estar situados en medios de almacenamiento de ordenador tanto locales como remotos, incluyendo dispositivos de almacenamiento de memoria. Las tareas llevadas a cabo por los programas y los módulos son descritas en lo que sigue con la ayuda de figuras. Los expertos en la técnica pueden implementar la descripción y las figuras como instrucciones ejecutables por procesador, que pueden ser escritas en cualquier forma de medio legible por ordenador.

Con referencia a la FIG. 2, un sistema ejemplar para implementar la invención incluye un dispositivo informático de uso general en forma de ordenador 60. Los componentes del ordenador 60 pueden incluir, sin limitación, una unidad 70 de proceso, una memoria 80 del sistema y un bus 71 del sistema que acopla diversos componentes del sistema, incluyendo la memoria del sistema, a la unidad 70 de proceso. El bus 71 del sistema puede ser de cualquiera de varios tipos de estructuras de bus, incluyendo un bus de memoria o controlador de memoria, un bus de periféricos y un bus local usando cualquiera de una variedad de arquitecturas de bus. A título de ejemplo y no de limitación, tales estructuras incluyen el bus de la arquitectura industrial normalizada (ISA), el bus de arquitectura microcanal (MCA), el bus ISA mejorado (EISA), el bus local de la Asociación de Normativa Electrónica sobre Vídeo (VESA) y un bus de interconexión de componentes periféricos (PCI), también denominado bus de entresuelo.

Típicamente, el ordenador 60 incluye una variedad de medios legibles por ordenador. Los medios legibles por ordenador pueden ser cualquier medio disponible que pueda ser objeto de acceso por el ordenador 60 e incluyen tanto medios volátiles como no volátiles, medios extraíbles y no extraíbles. A título de ejemplo y no de limitación, los medios legibles por ordenador pueden comprender medios de almacenamiento de ordenador y medios de comunicaciones. Los medios de almacenamiento de ordenador incluyen tanto medios volátiles como no volátiles, extraíbles y no extraíbles, implementados en cualquier procedimiento o tecnología para el almacenamiento de información tal como instrucciones, estructuras de datos, módulos de programa u otros datos legibles por ordenador. Los medios legibles por ordenador incluyen, sin limitación, RAM, ROM, EEPROM, memoria flash o cualquier otra tecnología de memoria, CD-ROM, discos versátiles digitales (DVD) u otro almacenamiento en disco óptico, casetes magnéticas, cinta magnética, almacenamiento en disco magnético u otros dispositivos de almacenamiento magnético, o cualquier otro medio que pueda ser usado para almacenar la información deseada y que pueda ser objeto de acceso por parte del ordenador 50.

Los medios de comunicaciones implementan típicamente instrucciones, estructuras de datos, módulos de programa u otros datos legibles por ordenador en una señal modulada de datos, tal como una onda portadora u otro mecanismo de transporte, e incluyen cualquier medio de distribución de información. La expresión “señal modulada de datos” significa una señal una o más de cuyas características han sido fijadas o cambiadas para cifrar la información de la señal. A título de ejemplo y no de limitación, los medios de comunicaciones incluyen medios cableados, tales como una red cableada o una conexión cableada directa, y medios inalámbricos tales como acústicos, de RF, infrarrojos y otros medios inalámbricos. También deberían incluirse dentro del alcance la los medios legibles por ordenador las combinaciones de cualquier de los anteriores.

La memoria 80 del sistema incluye medios de almacenamiento de ordenador en forma de memoria volátil y/o no volátil, tal como memoria 81 de solo lectura (ROM) y memoria 82 de acceso aleatorio (RAM). Típicamente, en la ROM 81 se almacena un sistema básico 83 de entrada/salida (BIOS) que contiene las rutinas básicas que contribuyen a transferir información entre elementos dentro del ordenador 60, tal como durante el arranque.

Típicamente, la RAM 82 contiene datos y/o módulos de programa que son inmediatamente accesibles a la unidad 70 de proceso y/o que están operados ya por ella. A título de ejemplo y no de limitación, la FIG. 2 ilustra el sistema operativo 84, programas 85 de aplicación, otros módulos 86 de programa y datos 87 de programa.

El ordenador 60 también puede incluir otros medios de almacenamiento de ordenador extraíbles/no extraíbles volátiles/no volátiles. A título de ejemplo únicamente, la FIG. 2 ilustra una unidad 91 de disco duro que lee de medios magnéticos no volátiles no extraíbles o escribe en ellos, un disco magnético no volátil 102 y una unidad 105 de disco óptico que lee de un disco óptico extraíble no volátil 106, tal como un CD ROM u otro medio óptico, o escribe en él. Otros medios de almacenamiento de ordenador extraíbles/no extraíbles volátiles/no volátiles que pueden ser usados en el entorno operativo ejemplar incluyen, sin limitación, casetes de cinta magnética, tarjetas de memoria flash, discos versátiles digitales, cinta de vídeo digital, RAM de estado sólido, ROM de estado sólido y similares. La unidad 91 de disco duro está conectada típicamente al bus 71 del sistema a través de una interfaz de memoria no extraíble, tal como la interfaz 90, y una unidad 101 de disco magnético y una unidad 105 de disco óptico están típicamente conectadas al bus 71 del sistema por medio de una interfaz de memoria extraíble, tal como la interfaz 100.

Las unidades y sus medios asociados de almacenamiento de ordenador presentados en lo que antecede e ilustrados en la FIG. 2 permiten el almacenamiento de instrucciones, estructuras de datos, módulos de programa y otros datos legibles por ordenador para el ordenador 60. En la FIG. 2, por ejemplo, la unidad 91 de disco duro se ilustra almacenando el sistema operativo 94, programas 95 de aplicación, otros módulos 96 de programa y datos 97 de programa. Obsérvese que estos componentes pueden ser los mismos o diferentes que el sistema operativo 84, los programas 85 de aplicación, otros módulos 86 de programa y los datos 87 de programa. El sistema operativo 84, los programas 85 de aplicación, otros módulos 86 de programa y los datos 87 de programa reciben aquí números diferentes para ilustrar que, como mínimo, son copias diferentes.

Un usuario puede introducir órdenes e información en el ordenador 60 a través de dispositivos de entrada tales como un teclado 112, un micrófono 113, un tablero 114 de escritura manuscrita, y un dispositivo 111 de puntero, tal como un ratón, una bola de mando o una almohadilla táctil. Otros dispositivos de entrada (no mostrados) pueden incluir una palanca de mando, un mando para juegos, una antena parabólica, un escáner o similares. Estos y otros dispositivos de entrada están a menudo conectados con la unidad 70 de proceso a través de una interfaz 110 de entrada de usuario que está acoplada al bus del sistema, pero pueden estar conectados por medio de otra interfaz y otras estructuras de bus, tal como un puerto paralelo, un puerto de juegos o un bus serie universal (USB). Un monitor 141 u otro tipo de dispositivo de visualización también está conectado al bus 71 del sistema a través de una interfaz, tal como una interfaz 140 de vídeo. Además del monitor, los ordenadores también pueden incluir otros dispositivos periféricos de salida, tales como altavoces 147 y la impresora 146, que pueden estar conectados a través de una interfaz 145 de periféricos de salida.

El ordenador 60 puede operar en un entorno de red usando conexiones lógicas a uno o más ordenadores remotos, tal como un ordenador remoto 130. El ordenador remoto 130 puede ser un ordenador personal, un dispositivo de mano, un servidor, un dispositivo de encaminamiento, un PC de red, un dispositivo del mismo nivel u otro nodo común de red, y típicamente incluye muchos de los elementos descritos en lo que antecede, o todos, relativos al ordenador 60. Las conexiones lógicas representadas en la FIG. 2 incluyen una red 121 de área local (LAN) y una red 123 de área amplia (WAN), pero también pueden incluir otras redes. Tales entornos de red son habituales en despachos, redes de ordenadores de ámbito empresarial, intranets e Internet.

Cuando se usa en un entorno de red LAN, el ordenador 60 está conectado a la LAN 121 a través de una interfaz o un adaptador 120 de red. Cuando se usa en un entorno de red WAN, el ordenador 60 incluye típicamente un módem 122 u otro medio para establecer comunicaciones en la WAN 123, tal como Internet. El módem 122, que puede ser interno o externo, puede estar conectado al bus 71 del sistema a través de la interfaz 110 de entrada de usuario o de otro mecanismo apropiado. En un entorno de red, los módulos de programa representados relativos al ordenador 60, o porciones de los mismos, pueden ser almacenados en el dispositivo remoto de almacenamiento de memoria. A título de ejemplo y no de limitación, la FIG. 2 ilustra programas remotos 135 de aplicación que residen en el ordenador remoto 130. Se apreciará que las conexiones de red mostradas son ejemplares y que pueden usarse otros medios de establecimiento de un enlace de comunicaciones entre ordenadores.

Debiera entenderse que el analizador 20 de textos puede residir en el ordenador 60 o en cualquier ordenador que se comunique con el ordenador 60, tal como el ordenador remoto 130. De forma similar, el léxico 22 puede residir en el ordenador 60 en cualquiera de los dispositivos de almacenamiento descritos en lo que antecede, o ser accesible a través de un enlace de comunicaciones adecuado.

La Fig. 3 es una representación gráfica del léxico 22. En el aspecto ejemplar ilustrado, el léxico 22 incluye una sección 160 de cabecera, una sección 162 de lista de palabras, una sección 164 de tabla índice, una sección de índices 166, dos o más secciones 168 de datos de léxico (en el presente documento, por ejemplo, 16 secciones 168A, 168B, 168C, 168D, 168E, 168F, 168G, 168H, 168I, 168J, 168K, 168L, 168M, 168N, 1680, 168P), y una sección 170 de pila de cadenas.

Generalmente, la sección 160 de cabecera almacena información en cuanto a la estructura del léxico 22. Por ejemplo, la sección 160 de cabecera puede incluir información en cuanto al nombre y la versión del léxico. La

sección 160 de cabecera puede también incluir información en cuanto al desplazamiento en memoria y el tamaño de cada una de las secciones 162, 164, 166, 168A-168P y 170. La sección 162 contiene la lista de palabras del léxico 22. Puede usarse cualquier formato para implementar la lista de palabras en la sección 162. Un formato particularmente útil comprende almacenar una lista de palabra en una estructure “trie”, que es una técnica de estructura de datos bien conocida. Las ventajas de este formato incluyen ser capaz de determinar fácilmente cuántas palabras pueden comenzar con un prefijo particular, lo que puede resultar útil, por ejemplo, en el reconocimiento de la escritura manuscrita y cuando es necesario determinar la probabilidad de que el usuario haya escrito una letra particular. Este formato también permite que se conozca el sentido de la escritura tanto hacia la derecha como hacia la izquierda. Tal como se ha indicado en lo que antecede, pueden usarse otras formas de listado de palabras en la sección 162. Por ejemplo, puede usarse una simple tabla o lista. En otro aspecto adicional, puede usarse una técnica de “diferencias” para almacenar la lista de palabras en la que se almacene la diferencia en símbolos o caracteres entre palabras sucesivas.

Antes de describir la sección 164, puede resultar útil describir primero la sección 166 y su relación con la pluralidad de secciones 168. Según se expuso en la sección de antecedentes, los léxicos actuales requieren que se lea toda la información relativa a una entrada de palabra particular, aunque puede que solo se desee una porción de la información. Las secciones 168A-168P permiten que los datos de cada entrada de palabra en el léxico estén organizados como se desee, de modo que la información relacionada del léxico puede agruparse generalmente de forma conjunta. Por ejemplo, puede usarse una de las secciones 168A-168P para almacenar información relativa a la verificación ortográfica, mientras que otra sección almacena información relativa a clasificaciones lingüísticas normalizadas. Generalmente, la sección 166 de índices proporciona punteros (por ejemplo, agrupados en conjuntos) a los datos almacenados en las secciones 168A-168P en función de entradas de palabra de la sección 16 de lista de palabras. En otras palabras, la sección 162 (por ejemplo, trie) de lista de palabras determina directa o indirectamente el punto o los puntos de acceso (desplazamientos) a la sección 166 de índices. Generalmente, el procedimiento para obtener información de palabras incluye acceder a la sección de lista de palabras en función de una palabra dada para determinar una identificación de puntero para la sección de índices. Usando la identificación de puntero se obtiene un puntero para la palabra en la sección de índices. El puntero se usa entonces para determinar qué sección de datos de la pluralidad de secciones de datos tiene información sobre la palabra dada en qué lugar está situada la información en la sección de datos. Así, para una entrada de palabra particular presente en la sección 162, los correspondientes datos de léxico almacenados en las secciones 168A-168P para esa palabra pueden ser objeto de acceso selectivo a través de la sección 166 de índices, no requiriéndose por ello que toda la información de palabra para una palabra dada sea procesada o, al menos, leída.

En un aspecto particularmente útil, el puntero o el conjunto de punteros en la sección 166 de índices para las secciones 168A-168P para cada entrada de palabra en la sección 162 están clasificados por su parte de oración (“POS”) en cuanto a si la entrada de palabra puede ser un sustantivo, un verbo, un adjetivo, etc. De esta manera, los datos para una POS de una entrada de palabra son una serie de punteros a la información de POS en las secciones 168A-168P. Así, para una entrada de palabra con dos POS, hay dos conjuntos diferenciados de punteros en la sección 166. Un conjunto indicaría la ubicación de la información sobre la primera POS (por ejemplo, la forma sustantiva de la entrada) y el segundo conjunto de punteros indicaría la ubicación de información sobre la otra POS (por ejemplo, la forma verbal de la entrada). En este punto, debería entenderse que pueden usarse otras formas de clasificación además de la POS, dependiendo del lenguaje al que esté dirigido el léxico 22. Por ejemplo, en vez de usar partes de la oración, para los idiomas japonés o chino pueden usarse clasificaciones de inflexión o tonales. Aunque está ejemplificado en el presente documento cuando la sección 166 de índices proporciona clasificación de POS, no debería considerarse que esta característica sea limitante ni requerida.

También debería hacerse notar que el uso de la palabra “palabra”, tal como se usa en el presente documento, incluye símbolos, ideogramas, logogramas, etc., tal como se usan en idiomas tales como el chino y el japonés. Así, pueden construirse léxicos para estos idiomas usando aspectos de la presente invención y se pretende que estén cubiertos por las reivindicaciones, a no ser que se haga notar lo contrario en el presente documento.

En el aspecto ejemplar, cada puntero incluye información relacionada con a qué sección 168A-168P apunta el puntero, información relacionada con el tipo de POS, y un valor de desplazamiento en el que han de encontrarse los datos relevantes en la sección identificada 168A-168P. Aunque los punteros asociados con una entrada de palabra dada en la sección 162 pueden ser fijos, en el aspecto ejemplar el número de punteros para cada entrada de palabra puede variar de una entrada de palabra a otra. De esta manera, la sección 166 de índices puede ser más compacta y flexible, sin ninguna limitación fija inherente.

A continuación se proporciona una representación de un puntero ejemplar en la sección 166 para una entrada de palabra: X_1: X_2: X_3: X_4:
siendo X_1 una bandera que indica el fin del conjunto de punteros para la entrada de palara, siendo X_2 información que identifica a una de las secciones 168, siendo X_3 información que identifica la POS u otra clasificación, y siendo X_4 un valor que indica el desplazamiento para los datos de léxico identificados por X_2. Usando este formato, los punteros para toda la información para una palabra dada son almacenados secuencialmente, estando identificado el primer puntero directa o indirectamente en función de la sección de lista de palabras y estando puesta la bandera X_1 para el último puntero para indicar el fin de la lista de punteros para la palabra dada. En un aspecto, la sección 166

de índices es una gran matriz de DWORD (cantidades de 4 bytes, alineamiento de palabra de 4 bytes para un acceso rápido). En este aspecto, un byte comprende una bandera de un bit para X_1 que indica el fin del conjunto de punteros, cuatro bits X_2 que indican la sección 168A-168P y tres bits X_3 que indican el tipo de POS. Luego se usan tres bytes para X_4 para proporcionar un valor de desplazamiento de 24 bits para indicar el lugar en el que los datos están almacenados en las secciones 168A-168P. Debiera entenderse que este formato no es más que un ejemplo en el que también pueden usarse otros formatos. De forma similar, no debería considerarse que este ejemplo sea requerido ni limitante. En general, se escoge el formato de los punteros de la sección 166 de índices para indicar la ubicación de los datos en la pluralidad de secciones 168 y, si se desea, una o más clasificaciones de la información de la palabra.

En este punto, debería hacerse notar también que cualquier dato que sea lo bastante pequeño como para caber en la porción de desplazamientos de una entrada de punto de la sección 166 puede ser cifrado directamente en la sección 166 de índices, en vez de en una sección 168A-168P separada. Ejemplos de este tipo de datos incluyen información ortográfica, o datos de probabilidad y frecuencia para la entrada de palabra, cualquiera de los cuales a menudo puede ser fácilmente almacenado en los bits asignados para los valores de desplazamiento de los datos.

Tal como se ha indicado en lo que antecede, la entrada a la sección 166 de índices es una función de la entrada de palabra desde la sección 162. Pueden usarse diversas técnicas para pasar entre las secciones 162 y 166. En un primer aspecto, cada entrada de palabra en la sección 162 podría incluir el requerido desplazamiento a la sección 166. Sin embargo, si la sección 162 comprende una estructura trie, puede ser necesaria la modificación de la estructura del nodo hoja del trie. Da manera alternativa, puede usarse los desplazamientos de los nodos de la estructura trie como desplazamientos a la sección 166 de índices. En el aspecto ejemplo, esto significaría asignar 40 bytes (10 punteros POS) para el conjunto de índices POS para una entrada de palabra. En otro aspecto adicional, el valor de desplazamiento puede estar fijado al final de una entrada de palabra en la sección 166.

En otra realización adicional, la tabla índice 164 está incluida en la estructura del léxico 22. La tabla índice 164 permite el establecimiento de una correlación entre las entradas de palabras y la sección 166 de índices y es particularmente útil cuando los punteros en la sección 166 pueden variar en número de una entrada de palabra a otra. Sin embargo, es posible usar un número de tamaño fijo de punteros en la sección 166 de índices para cada entrada de palabra asociada. Usando esta estructura de la sección 166, no sería necesaria la sección 164 de tabla índice. En este aspecto alternativo, en caso de que se permitan entradas de palabras con más del número fijo de punteros de índice en la sección 166, podría usarse una tabla de desbordamiento.

En este punto debería hacerse notar que los desplazamientos de la sección 162 a la sección 166 de índices y, más en particular, los punteros de la sección 166 que apuntan a los datos en las secciones 168A-168B pueden ser organizados para proporcionar eficiencia y velocidad cuando se recuperan datos del léxico 22. Por ejemplo, los puntos de desplazamiento pueden organizarse para ubicar información de léxico en las secciones 168A-168P adyacente a otra información de palabras usadas frecuentemente o, si se desea, organizarse información asociada o relacionada en las secciones 168A-168B más estrechamente entre sí, de modo que, cuando se almacenen en un dispositivo de almacenamiento de ordenador, tal como un disco duro, un disco flexible o similares, se acorten los tiempos de recuperación de la información.

Los datos de las secciones 168A-168P pueden estar presentes en las mismas o, si se desea, pueden proporcionarse punteros para que refieran las entradas de palabras de la sección 162 a datos contenidos dentro de la misma sección 168A-168P, en otras secciones 168A-168P, y, en el aspecto ilustrativo, también a una pila 170 de cadenas. Se usa la pila 170 de cadenas para proporcionar una única ubicación de almacenamiento para una cadena seleccionada cuyos datos sería preciso que fueran almacenados como múltiples instancias en las secciones 168A-168P. La pila 170 de cadenas puede ser una sola sección o puede incluir subsecciones similares a las secciones 168A-168P. Otras formas de información en las secciones 168 pueden incluir banderas booleanas, valores, listas de palabras en árboles de decisión, etc.

La organización de los datos de entradas de palabras usando una pluralidad de secciones 168A-168P permite que el léxico 22 sea fácilmente adaptado para cumplir las necesidades de una aplicación particular sin consumir grandes cantidades de memoria en el ordenador en el que está implementado. Por ejemplo, el léxico 22 puede leer introduciendo datos en una memoria de acceso rápido como la RAM; sin embargo, si no se necesita un tipo particular de datos en el léxico, pueden omitirse esa sección o secciones de la pluralidad de secciones 168A-168P. Aunque los punteros en la sección 166 de índices pueden ser modificados para que reflejen únicamente aquellas secciones 168A-168P que están presentes, en un aspecto adicional, es innecesaria la modificación, dado que, si las secciones 168A-168P están presentes, se obtiene información, mientras que si la sección no está presente, no se busca información alguna. Las secciones presentes en el léxico pueden estar registradas, por ejemplo, en la cabecera 160 para no causar errores.

Un beneficio particular de la estructura de léxico descrita en el presente documento es que un usuario o un autor del léxico pueden poner cualquier tipo de información sobre una palabra para su recuperación posterior si la entrada 12 (Fig. 1) incluye esa palabra. Además, no es preciso que la información definida por el usuario se entremezcle con

otra información contenida en el léxico, sino que, más bien, puede ser almacenada en una o varias secciones dedicadas en la pluralidad de secciones 168A-168P.

Los que siguen no son más que unos ejemplos de datos de léxico organizados en secciones adecuadas para las secciones 168A-168P. Debería hacerse notar que estos no son más que meros ejemplos en los que los datos del léxico 22 pueden ser organizados de cualquier manera deseada en aras de la conveniencia o de la comprensión. Se ha descubierto que las secciones descritas en el presente documento son particularmente útiles, pero no debería considerarse que sean requeridas ni limitantes.

Sección de datos de morfología: tal información puede incluir pronunciación, así como otras formas de la palabra para tiempos verbales de varias palabras.

Sección estándar de datos de autor: tal información puede incluir datos que indiquen si la entrada de palabra es o no singular, plural o si la palabra es animada o inanimada. La información relativa a las entradas de palabras para esta sección es generalmente información bien conocida sobre la entrada de palabra tal como aquella de la que podría ser autor un lego. De esta manera, esta información puede ser cambiada o alterada fácilmente para amoldarse a los requerimientos de un usuario.

Sección estándar de datos lingüísticos: tal información incluye información lingüística de las entradas de palabras. Aunque tal información no resulta bien conocida para un lego común, los lingüistas pueden entender fácilmente esta información y modificarla según sea necesario.

Sección de datos de análisis sintáctico: tal información incluye datos útiles para el análisis sintáctico de lenguaje natural.

Sección de datos de dominio/tema: tal información está relacionada con códigos de dominio o tema. Por ejemplo, la información puede indicar que las correspondientes palabras están relacionadas con la física, las matemáticas, la geografía, la alimentación, etc.

Sección de datos de ortografía: tal información está relacionada con la verificación ortográfica, por ejemplo, marcas de dialecto, marcas restringidas, etc. Las marcas restringidas indican palabras que son permitidas pero que no se sugieren durante la verificación ortográfica, tal como vulgaridades, acrónimos, términos arcaicos, etc.

Sección de datos de expresión de múltiples palabras: tal información es útil cuando es preciso identificar por separado palabras múltiples como modismos, nombres propios, títulos de un libro o de películas, títulos de cargos, nombres de lugares, etc. Comúnmente, los datos almacenados con respecto a cada entrada de palabra son las palabras que preceden y/o siguen a la palabra en una expresión de múltiples palabras.

Por ejemplo, una de las secciones de la pluralidad de secciones 168A-168P puede incluir pares jerárquicos arbitrarios de valores de nombre para buscar únicamente por el autor de las entradas del léxico. Por ejemplo, si un autor desea añadir información de la Entidad Nombre (NE) sobre una expresión de múltiples palabras (expuesto más arriba), el autor puede añadir en la sección un conjunto de pares de valores basados en cadenas de nombres, que en el formato XML puede ser representado como:

```
<named-entity>
   <app-ne-id>movieFinder::el_día_más_largo</app-ne-id>
   <semantic-type>movieFinder::movieTitle</semantic-type>
   <genre>Drama</genre>
   <URL>http://www.movieFinder.com/fetch-movie-info/the_longest_day</URL>
   <movie-info>
      <date>30 de enero de 1969</date>
      <running-time>137 min.</running-time>
      <studio>20th Century Fox</studio>
   </movie-info>
   <non-rated/>
</named-entity>
```

La sección puede representar, así, pares de nombres de valores simples basados en cadenas de anidamiento arbitrario. El formato no soporta atributos de identificadores de XML, sino que el autor codifica estos como subelementos separados de esa sección. En el ejemplo anterior, los datos para el título de una película tiene una mezcla de datos específicos de aplicación que pueden ser almacenados si se desea.

La estructura del léxico 22 acomoda léxicos escribibles aprovechando el hecho de que no es preciso que cada sección del léxico sea continua con la sección que la precede inmediatamente. En otras palabras, las secciones pueden reservar espacio extra no utilizado para su expansión futura. Las operaciones de actualización del léxico se llevan a cabo escribiendo el o los nuevos valores en las debidas ubicaciones. Obsérvese que si el léxico 22 está implementado como un léxico basado en una DLL (biblioteca de enlace dinámico) o como léxicos basados en

ficheros precompilados (estáticos), que no tienen espacio de reserva, una simple implementación de lista libre encuentra espacio libre de entradas con base en un algoritmo de primer ajuste.

Generalmente, un procedimiento de almacenamiento de información de palabras en el léxico 22 incluye almacenar información de palabras en la pluralidad de secciones 168 de datos, almacenando cada sección de datos información seleccionada sustancialmente diferente sobre las palabras de una lista de palabras; almacenar información de punteros en la sección 166 de índices, que está separada de la pluralidad de secciones 168 de datos, apuntando cada uno de los punteros a datos seleccionados en la pluralidad de secciones 168 de datos; y almacenar la lista de palabras en una sección 162 de lista de palabras separada de la pluralidad de secciones 168 de datos y de la sección 166 de índices, teniendo la lista de palabras información para identificar los correspondientes punteros relacionados con una palabra seleccionada. Si se desea, los valores de identificación pueden almacenarse en la sección 164 de tabla índice, en la que cada valor de identificación corresponde a una palabra de la sección 162 de lista de palabras y está asociado con un puntero en la sección 166 de índices. De forma similar, pueden incluirse indicaciones de clasificación en los punteros para clasificar la información de la palabra.

La estructura del léxico 22 resulta particularmente útil cuando es deseable obtener información de varios léxicos. En general, los datos de múltiples léxicos para una palabra particular pueden combinarse, ignorarse o seleccionarse como se desee. Ejemplos de combinación de información de léxico de varios léxicos existen en una implementación en la que un léxico central o base contiene una primera cantidad de información sobre entradas de palabras, un segundo léxico incluye una segunda cantidad de información sobre las entradas de palabras para un dominio particular y un tercer léxico incluye una tercera cantidad de información sobre las entradas de palabras según determina el usuario.

La Fig. 4 ilustra esquemáticamente cómo puede obtenerse de múltiples léxicos información para una entrada de palabra particular. En la Fig. 4, los léxicos (representados solamente por las secciones de datos, pero, por lo demás, de los que se ilustran algunas de sus secciones, o todas, en la Fig. 3) están organizados en filas e indicados en 180, 181, 182 y 183. Las secciones individuales de datos (correspondientes a las secciones 168) están representadas verticalmente en la Fig. 4 y, en este aspecto ilustrativo, hay un máximo de seis secciones 190, 191, 192, 193, 194 y 195 de datos que pueden ser objeto de acceso en los cuatro léxicos 180-183. Obsérvese que no es necesario que cada léxico 180-183 incluya todas las secciones 190-195 de datos y, en muchos casos prácticos, no existiría tal correspondencia entre todas las secciones de datos de todos los léxicos.

En la Fig. 4, se usa la notación X_y para indicar datos en una sección de léxico, denotando X una de las secciones 190-195 de datos y denotando Y el léxico 180-183. Por ejemplo, el léxico 180 comprende las secciones de datos 190_{180}, 193_{180} y 195_{180}.

Dado que los datos de los léxicos 180-183 han sido organizados en secciones 190-195 que tienen el mismo tipo de contenido, la información puede ser combinada o seleccionada fácilmente en los léxicos 180-183. Puede obtenerse la información para una entrada de palabra dada examinando la información en el primer léxico 180 y luego avanzando a la misma sección de datos de los otros léxicos 181-183 según se necesite. En un aspecto, los datos a recuperar son controlados por un conjunto definido en tiempo de ejecución de tipos de sección deseados. Una variable determina si leer o no leer datos para una entrada de un léxico. Una segunda variable determina si combinar o sobrescribir datos de entradas leídas previamente de la correspondiente sección de los otros léxicos examinados. Esquemáticamente, puede considerarse que los léxicos están “apilados” y que la información se obtiene leyendo la sección 190-195 de datos del léxico superior de la pila y luego proseguir secuencialmente hacia abajo en la pila, siguiendo las reglas en cuanto a si leer o no y en cuanto a si seleccionar, ignorar, sobrescribir o combinar. La estructura de léxico ilustrada en la Fig. 3 permite al implementador escoger cómo se combinan los datos de un tipo dado de sección con los datos de la misma sección en otros léxicos, o los anulan.

En la Fig. 4, la información 186 obtenida de los léxicos 180-183 comprende información correspondiente a las secciones 190_{180}, 191_{181}, 192_{183}, 193_{180}, 194_{183} y $195_{180 + 181 + 182}$. En este ejemplo, los datos para las secciones 190, 191, 193 y 194 se obtienen simplemente examinando los léxicos 180-183 en orden sección a sección hasta que se encuentra un indicador de parada en una de las secciones de datos. Por ejemplo, aunque tanto el léxico 180 como el 183 tienen información en la sección 190, solo se recupera la información del léxico 180 porque se encontró un indicador de parada en la sección 190_{180}. En tiempo de ejecución, esto hace que la información de la sección 190_{183} sea ignorada. En cambio, se combina entre sí la información de las secciones 195_{180}, 195_{181} and 195_{182} para formar la información $195_{180 + 181 + 182}$ porque no se encontró un indicador de parada hasta que se examinó la sección 195_{182}. Si se desea, la información en las secciones de todos los léxicos puede ser combinada, ignorada o, si no, seleccionarse con base en reglas implementadas por el analizador 20 de textos o por un módulo de interfaz, no mostrado, que accede al léxico 22 con base en peticiones del analizador 20 de textos. Por ejemplo, tales reglas pueden indicar que una sección particular de un léxico particular ha de usarse siempre, con independencia de si hay información en secciones correspondientes de otros léxicos. Esto se representa en la Fig. 4, en la que, aunque hay información presente en el léxico 182, y al menos la sección 192 es examinada en primer lugar porque está más alta en la pila, se obtiene la información de la sección 192 del léxico 183. Sin embargo, la selección de datos también

puede realizarse por entrada de palabra, por ejemplo usando los punteros de parada, tal como se ha descrito en lo que antecede.

En resumen, se ha descrito una estructura mejorada de léxico que proporciona flexibilidad y eficiencias no disponibles anteriormente. La sección de índices y la pluralidad de secciones de datos permiten que el léxico se adapte para cuadrar con las necesidades del sistema procesador de textos y/o los recursos disponibles del ordenador. La estructura mejorada de datos también permite que los datos de múltiples léxicos sean objeto de acceso selectivo y/o combinados según se desee.

Aunque la presente invención ha sido descrita con referencia a aspectos preferentes, los expertos en la técnica reconocerán que pueden efectuarse cambios en forma y detalla sin apartarse del ámbito de la invención.

REIVINDICACIONES

1. Un medio de almacenamiento legible por ordenador que tiene una pluralidad de léxicos (180-183) para almacenar información de palabras y adaptado para su uso en un analizador (20) de textos en un sistema (10) de tratamiento del lenguaje, comprendiendo cada léxico:

una sección (162) de lista de palabras para almacenar una pluralidad de palabras;
un conjunto de secciones (168A -168P; 190 - 195) de datos que se corresponde con cada palabra de la lista de palabras, almacenando las secciones de datos información seleccionada diferente sobre la correspondiente palabra de la lista de palabras; y
para cada palabra de la lista de palabras, una pluralidad de punteros almacenados en una tabla (166) de índices aparte de los conjuntos de secciones de datos, apuntando cada uno de los punteros a una sección de datos diferente relacionada con información diferente sobre la palabra correspondiente, incluyendo cada uno de los punteros una primera indicación de a qué sección de datos acceder, una segunda indicación de un valor de desplazamiento relacionado con la información almacenada en el mismo y una tercera indicación de una clasificación de la palabra;
en el que las secciones (190-195) de datos de cada uno de los léxicos (180-183) que tienen información similar son accesibles selectivamente para obtener información de las mismas, y en el que dicha pluralidad de léxicos que tienen secciones de datos con información similar permite:

obtener la información de palabras de al menos dos secciones (195_{180}, 195_{181}, 195_{182}) de datos que tienen información similar para combinar la información ($195_{180\text{-}181+182}$) obtenida de palabras, siendo dichas al menos dos secciones (195_{180}, 195_{181}, 195_{182}) de datos de al menos dos léxicos diferentes (180-182), u
obtener la información de palabras (192_{182}) de al menos dos secciones (192_{182}, 192_{183}) de datos que tienen información similar y luego usar solo la información de palabras obtenida de una sección (192_{182}) de datos, siendo dichas al menos dos secciones (192_{182}, 192_{183}) de datos de al menos dos léxicos diferentes (182-183).

2. El medio de almacenamiento legible por ordenador de la reivindicación 1 en el que la sección de lista de palabras es una estructura de datos trie.

3. El medio de almacenamiento legible por ordenador de la reivindicación 1 en el que la identificación es un valor de desplazamiento almacenado en la sección de lista de palabras.

4. El medio de almacenamiento legible por ordenador de la reivindicación 1 en el que cada léxico comprende, además, una sección (164) de tabla índice para almacenar cada una de las identificaciones correlacionadas con palabras de la sección de lista de palabras, teniendo cada palabra de la sección de lista de palabras una entrada correspondiente en la sección de la tabla índice.

5. El medio de almacenamiento legible por ordenador de la reivindicación 1 en el que una sección de la pluralidad de secciones de datos almacena información relacionada con la verificación ortográfica.

6. El medio de almacenamiento legible por ordenador de la reivindicación 1 en el que una sección de la pluralidad de secciones de datos almacena información relacionada con la morfología.

7. El medio de almacenamiento legible por ordenador de la reivindicación 1 en el que una sección de la pluralidad de secciones de datos almacena información relacionada con la lingüística.

8. El medio de almacenamiento legible por ordenador de la reivindicación 1 en el que una sección de la pluralidad de secciones de datos almacena información que indica que una palabra pertenece a una expresión de múltiples palabras.

9. El medio de almacenamiento legible por ordenador de la reivindicación 1 en el que dos secciones de datos de la pluralidad de secciones de datos almacenan por separado información seleccionada del grupo que consiste en información de verificación ortográfica, información morfológica, información lingüística e información de expresiones de múltiples palabras.

10. Un procedimiento implementado por ordenador para obtener información de palabras accediendo a una pluralidad de léxicos (180-183), estando adaptado cada léxico para ser usado con un analizador (20) de textos en un sistema (10) de tratamiento del lenguaje, en el que cada léxico almacena información de una palabra perteneciente a una pluralidad de palabras, comprendiendo cada léxico:

una sección (162) de lista de palabras que almacena la pluralidad de palabras;
conjuntos de secciones (168A -168P; 190 - 195) de datos, correspondiéndose cada conjunto de secciones de datos con una palabra individual en la sección de lista de palabras, almacenando cada sección de datos entre un conjunto de secciones de datos información seleccionada diferente sobre la correspondiente palabra de la lista de palabras; y

una sección (166) de índices que almacena una pluralidad de punteros aparte de los conjuntos de secciones de datos, correspondiéndose cada pluralidad de punteros con una palabra individual, apuntando cada puntero a datos en una sección de datos, incluyendo cada uno de los punteros una primera indicación de a qué sección de datos acceder, una segunda indicación de un valor de desplazamiento relacionado con la información almacenada en el mismo y una tercera indicación de una clasificación de la palabra, comprendiendo el procedimiento:

acceder selectivamente a secciones (190-195) de datos de cada uno de los léxicos (180-183) que tienen información similar y obtener información de los mismos,
en el que dicha pluralidad de léxicos que tienen secciones de datos con información similar permite:
obtener la información de palabras de al menos dos secciones (195_{180}, 195_{181}, 195_{182}) de datos que tienen información similar para combinar la información ($195_{180+181+182}$) obtenida de palabras, siendo dichas al menos dos secciones (195_{180}, 195_{181}, 195_{182}) de datos de al menos dos léxicos diferentes (180-182), u
obtener la información de palabras (192_{182}) de al menos dos secciones (192_{182}, 192_{183}) de datos que tienen información similar y luego usar solo la información de palabras obtenida de una sección (192_{182}) de datos, siendo dichas al menos dos secciones (192_{182}, 192_{183}) de datos de al menos dos léxicos diferentes (182-183); acceder a la sección de lista de palabras de al menos un léxico de dicha pluralidad de léxicos en función de dicha palabra para determinar una identificación de puntero para la sección de índices usando la identificación de puntero de dicho al menos un léxico para obtener un puntero en la sección de índices;
usar el puntero de dicho al menos un léxico para determinar qué sección de datos de la pluralidad de secciones de datos tiene información sobre la palabra dada y del lugar en el que la información está localizada en la sección de datos.

11. El procedimiento implementado por ordenador de la reivindicación 10 en el que la identificación es un valor de desplazamiento almacenado en la sección de lista de palabras.

12. El procedimiento implementado por ordenador de la reivindicación 10 en el que cada léxico incluye una sección (164) de tabla índice que almacenada cada una de las identificaciones correlacionadas con palabras de la sección de lista de palabras, y en el que acceder a la sección de lista de palabras en función de dicha palabra para determinar una identificación de puntero para la sección de índices incluye, además, el uso de la sección de lista de palabras para encontrar una entrada correspondiente en la sección de la tabla índice que tenga la identificación correspondiente.

13. El procedimiento implementado por ordenador de la reivindicación 13 en el que el acceso selectivo incluye obtener la información de palabras de secciones de datos similares de cada léxico hasta que se localiza un indicador de parada.

14. El procedimiento implementado por ordenador de la reivindicación 13 en el que el acceso selectivo a las secciones de datos incluye el acceso secuencial a los léxicos en un orden seleccionado.

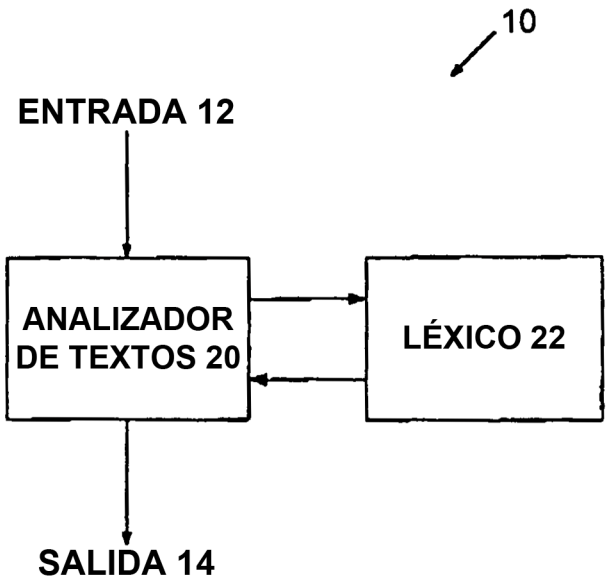

FIG. 1

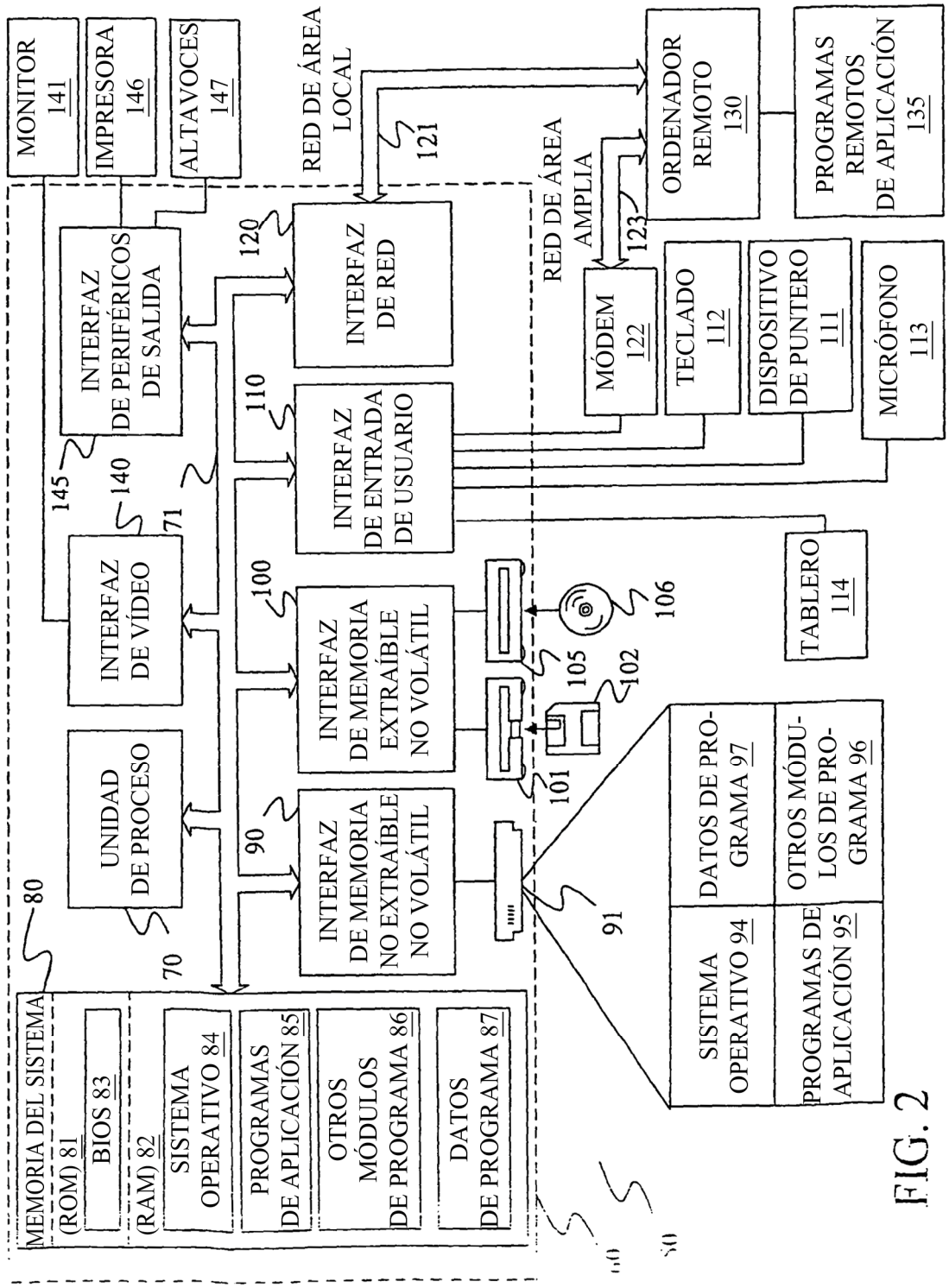
MONITOR 141
IMPRESORA 146
ALTAVOCES 147
RED DE ÁREA LOCAL
121
RED DE ÁREA AMPLIA
ORDENADOR REMOTO 130
PROGRAMAS REMOTOS DE APLICACIÓN 135
INTERFAZ DE PERIFÉRICOS DE SALIDA
INTERFAZ DE RED
120
INTERFAZ DE VÍDEO
INTERFAZ DE ENTRADA DE USUARIO
110
MÓDEM 122
TECLADO 112
DISPOSITIVO DE PUNTERO 111
MICRÓFONO 113
145
140
71
123
INTERFAZ DE MEMORIA EXTRAÍBLE NO VOLÁTIL
INTERFAZ DE MEMORIA NO EXTRAÍBLE NO VOLÁTIL
100
90
UNIDAD DE PROCESO
TABLERO 114
106
105
102
101
80
70
91
MEMORIA DEL SISTEMA (ROM) 81
BIOS 83
(RAM) 82
SISTEMA OPERATIVO 84
PROGRAMAS DE APLICACIÓN 85
OTROS MÓDULOS DE PROGRAMA 86
DATOS DE PROGRAMA 87
SISTEMA OPERATIVO 94
DATOS DE PROGRAMA 97
PROGRAMAS DE APLICACIÓN 95
OTROS MÓDULOS DE PROGRAMA 96

FIG. 2

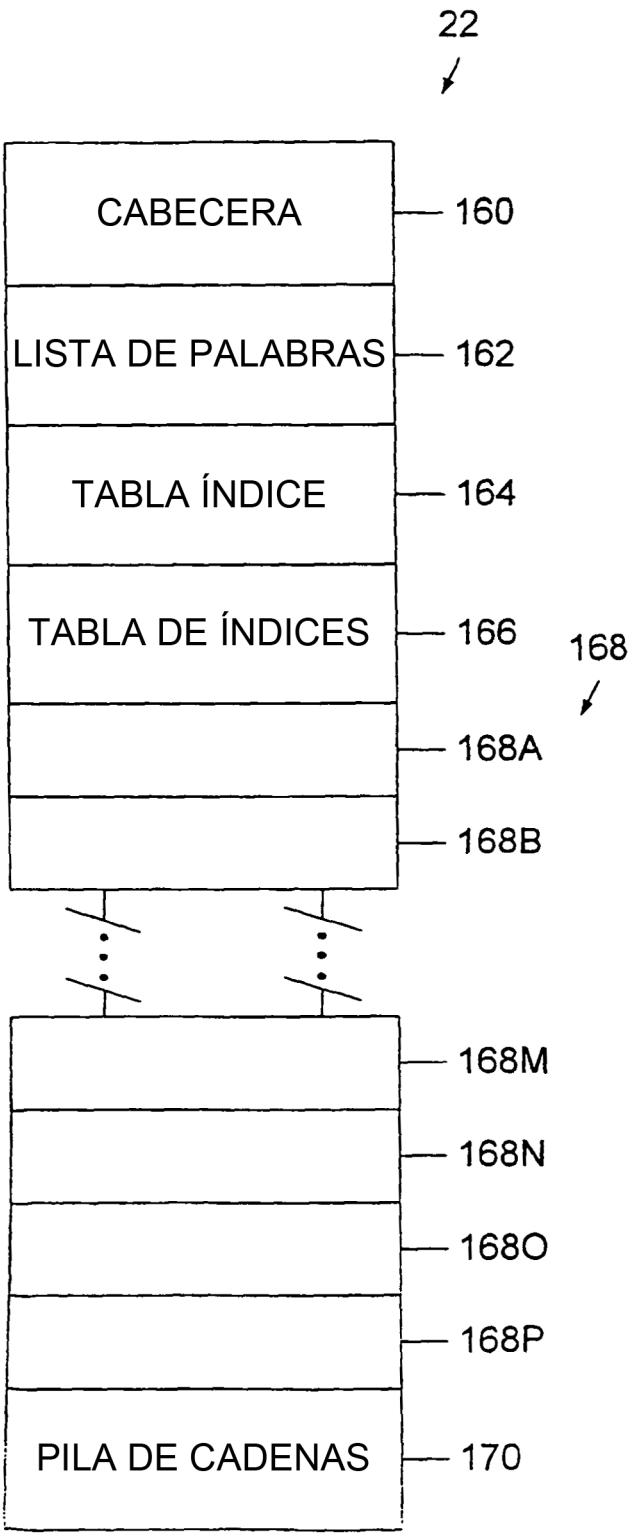
22
CABECERA
160
LISTA DE PALABRAS
162
TABLA ÍNDICE
164
TABLA DE ÍNDICES
166
168
168A
168B
168M
168N
168O
168P
PILA DE CADENAS
170

FIG. 3

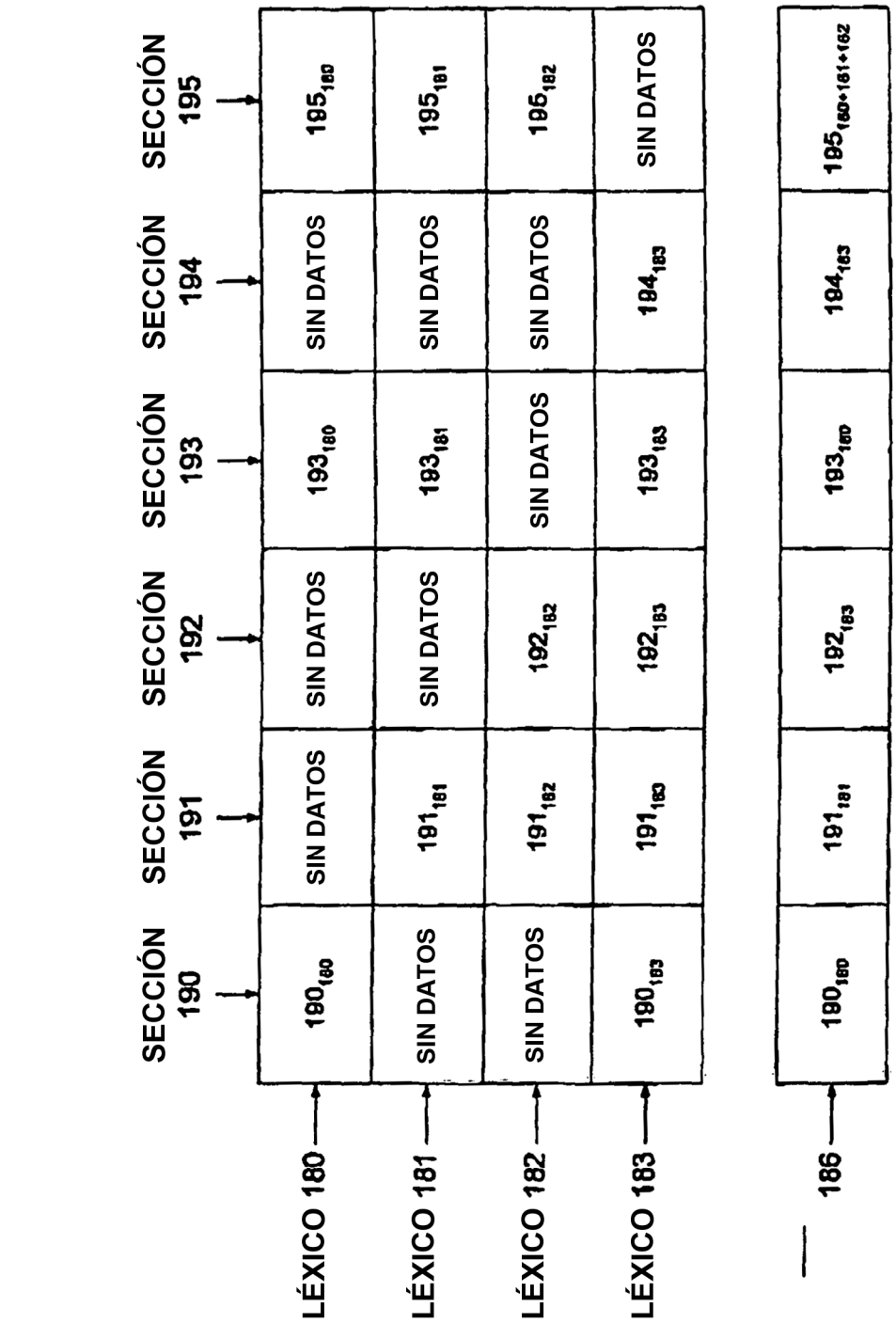
SECCIÓN 190
SECCIÓN 191
SECCIÓN 192
SECCIÓN 193
SECCIÓN 194
SECCIÓN 195
LÉXICO 180
LÉXICO 181
LÉXICO 182
LÉXICO 183
186
SIN DATOS
190_{180}
191_{181}
191_{182}
191_{183}
192_{182}
192_{183}
193_{180}
193_{181}
193_{183}
194_{183}
195_{180}
195_{181}
195_{182}
$195_{(180+181+182}$

FIG. 4