Proyectos de investigación

Octavio Santana Suárez

INFORMACIÓN TEXTUAL:

LÍNEA DE INVESTIGACIÓN Y PROYECTOS DE DESARROLLO

TÍTULO DEL PROYECTO: Gestión de Información Textual.
CENTRO DE INVESTIGACIÓN: Universidad de las Palmas de Gran Canaria.
DIRECTOR DEL PROYECTO: Octavio Santana Suárez.
EQUIPO COLABORADOR: Grupo de Investigación en Estructuras de Datos: M. Díaz, J. C. Rodríguez, D. González, G. Rodríguez, Z. Hernández, A. Ballester.
FECHA DE TERMINACIÓN: 1995.

0. INTRODUCCIÓN

El Grupo de Investigación en Estructuras de Datos de la Universidad de Las Palmas de Gran Canaria, dirigido por el Catedrático de Universidad D. Octavio Santana Suárez, ha estado realizando investigación básica desde 1986 en su campo, y de programas de desarrollo en áreas relacionadas con la recuperación de información textual a partir de 1990.

Nuestra investigación se centra en el problema clásico de los sistemas computacionales de información, la recuperación de información dentro de grandes volúmenes de datos como un problema de búsqueda de palabras. Este problema ha dado lugar a la creación de numerosos algoritmos, así como variantes en el enunciado. Una de ellas, la que aborda nuestro grupo, es considerar que la palabra de entrada puede contener errores con lo que se deben buscar no solamente sus copias exactas, sus repeticiones, sino, también, todas aquellas palabras similares según una cierta distancia de parecido formal. Para ello definimos una distancia, **DIT**, que no solamente presenta un tiempo de cálculo menor que otras convencionales –en particular la clásica distancia de Levenshtein– sino que además nos permite crear algoritmos de búsqueda eficientes a través de una estructura de datos –estructura **S-D**– en la que se organizan las palabras de igual longitud, considerando la frecuencia de sus caracteres.

Sobre la estructura **S-D** hemos planteado además otros tipos de búsquedas complejas, orientando de esta forma la investigación al desarrollo de software que gestione bases de datos documentales.

El proyecto **SOTA** consiste en el desarrollo de un sistema que permita la indización de documentos sin restricciones de formato. La estructura **S-D** se utiliza en **SOTA** como índice para localizar las palabras que se pueden tomar como argumento de búsqueda en la base de datos documental.

Otro de los aspectos de desarrollo de software que hemos seguido ha sido la creación de herramientas de soporte en la producción documental en castellano. El proyecto **FrecText** define un entorno de trabajo interactivo que permite analizar documentos desde un punto de vista estadístico, morfológico, topológico, semántico, entre otros, estando abierto a nuevas incorporaciones que se están desarrollando actualmente como el tratamiento de sinónimos y la corrección ortográfica de errores.

1. CUESTIONES SOBRE LOCALIZACIÓN DE PALABRAS

El almacenamiento de grandes volúmenes de datos textuales en un ordenador es un proceso relativamente sencillo; el problema se

plantea a la hora de manejar y recuperar en un tiempo aceptable la información almacenada –la búsqueda de palabras constituye la parte fundamental de este problema–.

Existen dos aproximaciones a la localización de palabras en textos. En la primera, se efectúa la búsqueda directamente sobre el documento sin ningún tipo de preparación preliminar; si tratan con textos de longitud mediana o las consultas son algo complejas, esta manera de proceder puede resultar muy costosa en cuanto a tiempo de respuesta. Desde la segunda perspectiva, se someten previamente los escritos a un tratamiento; la intención es generar un índice que aumente la eficacia de los accesos al documento.

1.1. *Búsqueda de coincidencias*

La búsqueda de coincidencias de palabras consiste en encontrar las ocurrencias de un patrón en un texto. Los algoritmos clásicos –válidos para documentos cortos– realizan la búsqueda de coincidencias directamente sobre el texto en un tiempo que es función lineal de la longitud del mismo; existen posteriores optimizaciones para ciertos casos.

1.2. *Búsquedas aproximadas. Similitud*

En la práctica, también se necesita a menudo analizar situaciones donde los datos no son del todo correctos. Considérese la situación en la que la palabra de entrada contiene errores de sustitución o presenta un cierto número de diferencias con el patrón, y de todas formas es necesario encontrar sus apariciones en el texto. Se podría suponer que lo que se proporciona como palabra de búsqueda es precisamente algo parecido a lo que previamente ha sido almacenado en algún registro o registros. Tal palabra no coincide exactamente con la de origen porque ha sufrido alguna distorsión. Sin embargo, puede ocurrir que una palabra se parezca mucho e incluso coincida con otra previamente almacenada. El objetivo de la búsqueda es precisamente la recuperación de las palabras candidatas.

2. Búsquedas en bases de datos documentales

Las bases de datos documentales representan grandes volúmenes de datos textuales y requieren métodos de búsqueda más rápidos que los mencionados –tiempo lineal respecto de la longitud del testo. Para ello se someten los escritos a un tratamiento previo, al objeto de construir una estructura auxiliar que aumente la velocidad de las ulteriores consultas. Estos métodos de búsqueda en texto pueden clasificarse en tres categorías: índices lexicográficos, índices basados en dispersiones y técnicas de agrupamiento.

3. Detección y corrección de errores

Si se plantea el hecho de corregir los errores introducidos por la distorsión, mediante el proceso de recuperación se obtiene una tentativa de corrección de errores, y con la recuperación de una palabra no relevante se obtendría un error.

Los métodos de corrección de errores se pueden clasificar en dos categorías: los métodos estadísticos que usan N-gramas, probabilidades de confusión y de ocurrencia; y los métodos de diccionarios que están basados en similitudes o distancias. En general, los métodos estadísticos como el algoritmo de Viterbi clásico y sus modificaciones son más rápidos que los que usan diccionarios, mientras que los métodos de diccionarios pueden obtener mejores tasas de corrección que los estadísticos.

La principal dificultad de los métodos que usan la distancia de Levenshtein entre dos palabras con el objeto de elegir las más parecidas a la de búsqueda de entre las que forman el diccionario, radica en el hecho de que se ha de comparar cada palabra de búsqueda con la base de datos entera, o con gran parte de ella. Existen algunos métodos de corrección que utilizan el digrama o trigrama de frecuencias. Sin embargo, se pone de manifiesto que tales métodos son inferiores a los que utilizan la distancia de Levenshtein.

Los investigadores han tomado un creciente interés por los métodos rápidos de corrección o búsqueda para vocabularios grandes. Una línea es encontrar métodos estadísticos con alta tasa de corrección. Otra línea consiste en aumentar la velocidad de los métodos de diccionario.

4. Investigación básica del grupo

Nuestros trabajos tratan de aspectos teórico-experimentales en torno al problema de

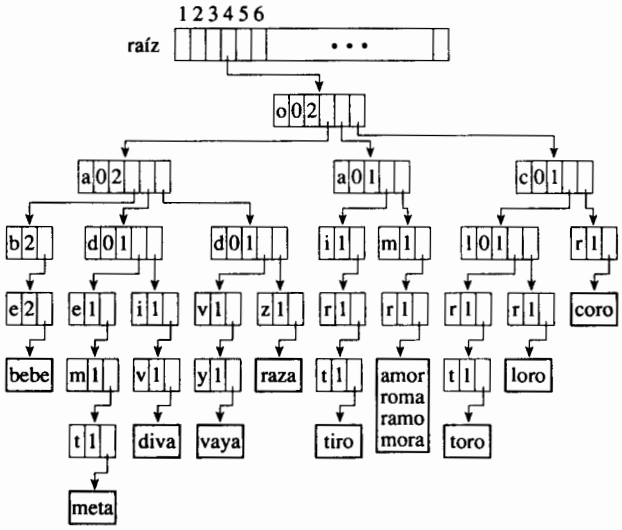
la búsqueda de las palabras más similares a una dada en grandes volúmenes de datos. El concepto de similitud se entiende en el sentido de la distancia de Levenshtein, DL. Dado un diccionario conjunto de palabras de algún alfabeto y una distancia, se recuperan todas las que se encuentran a distancia mínima de la proporcionada. Se construye como índice una estructura de datos que evita el recorrido secuencial del diccionario. El objetivo que se persigue es la optimización de los recursos de tiempo y espacio, de los esquemas de búsqueda y de la estructura de datos que los soporta.

4.1. *Una solución para la búsqueda de las palabras más similares*

Como aportación esencial se ha introducido una nueva distancia denominada distancia invariante trasposicional, DIT, que se define en función de la frecuencia de aparición de cada carácter en ambas cadenas. Se ha analizado su costo computacional obteniéndose que es sensiblemente inferior al de DL y además se ha demostrado que $DIT(X,Y) \leq 2*DL(X,Y)$. Estas características permiten su aplicación a la construcción de un filtro adaptativo DIT/DL que tiene por misión reducir el número de palabras del diccionario que han de soportar el cálculo de DL con la cadena de búsqueda.

El diccionario se organiza de manera genuina como un árbol en el cual se estructuran las componentes que intervienen en el cálculo de DIT —estructura S-D— con el fin de optimizar el cálculo DIT y a la vez disponer de una forma de uso más eficaz del filtro que seleccione las palabras con las que se evalúa DL. Se comprueba que la evolución de la relación ocupacional entre el índice y los datos con el incremento de la cardinalidad del diccionario, obteniéndose una tendencia a la igualdad en las respectivas ocupaciones, indica la bondad de la estructura.

El esquema de búsqueda de las palabras más similares desarrollado comienza con un valor del radio igual a infinito, y posteriormente va disminuyendo con el tratamiento de palabras cada vez más próximas. Hemos denominado a esta estructura de búsqueda Esquema Decreciente por contraposición al Esquema de búsqueda Creciente, donde el radio de búsqueda sigue una línea de modificación creciente. En el esquema creciente, a diferencia del decreciente, existen nodos que se visitan más de una vez; en principio esta característica no es deseable, pero queda suficientemente compensada por la aproximación más rápida a las palabras más próximas. La introducción de criterios de poda ha supuesto, para ambos esquemas, una optimización en el recorrido del índice.



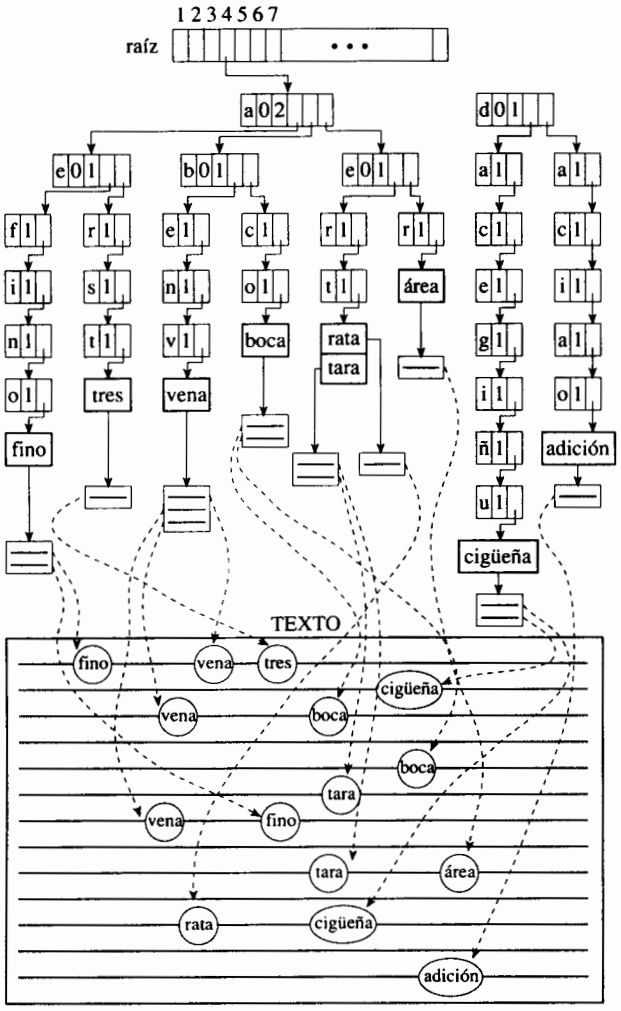
Se ha definido una nueva distancia, DS, calculada a partir de la longitud de la subsecuencia común más larga; tiene en cuenta características posicionales de las cadenas que el índice no considera y posee un bajo costo computacional inferior a DL pero superior a DIT. Además hemos demostrado que $DIT(X,Y) \leq 2*DS(X,Y) \leq DL(X,Y)$. Todo ello permite utilizar DS como filtro de DL, con lo que se obtiene una mejor realización.

Hemos estudiado diversas formas de paginación de la estructura S-D para resolver la situación en que debido a su tamaño no sea posible ubicarla en la memoria interna de un ordenador personal. Con la paginación en "preorden" se han obtenido los mejores resultados para los dos esquemas de búsqueda.

4.2. *Aplicación para peticiones complejas de localización de palabras en archivos documentales*

Se aplicó la estructura S-D a la recuperación de información en texto libre con resultados altamente satisfactorios. La estructura se construyó con el conjunto de palabras diferentes consideradas "no vacías", procedentes de un texto de más de un millón de palabras. Tal índice permite además de la recuperación de las

palabras más similares otros tipos de búsqueda habituales en texto libre, como son: la búsqueda con *máscara* que permite localizar palabras de las que se desconocen caracteres en posiciones determinadas, la búsqueda con *truncamiento* que permite recuperar artículos que contienen palabras que empiezan, terminan o presentan alguna coincidencia central con la palabra búsqueda, la búsqueda con *operadores booleanos*, que



combinen varias palabras utilizando como conectores los operadores lógicos, búsqueda con *condiciones topológicas* que delimita el entorno donde se aplica, y la búsqueda de frases que permite la localización de artículos en los que se encuentra la frase especificada.

La continuidad de este trabajo está en la extensión a la recuperación de información en bases de datos documentales de gran dinamicidad. El objetivo consiste en desarrollar un sistema que basándose en una reestructuración parcial de la estructura permita la recuperación de información en un archivo de documentos que se actualizan con frecuencia; la implementación se lleva a cabo en un "ambiente multitarea" de forma que puedan efectuarse tales reestructuraciones mientras el usuario realiza consultas. Otro aspecto a tratar en futuros trabajos consiste en llevar a cabo una generalización de los diversos tipos de búsqueda, teniendo en cuenta la organización interna de los diferentes documentos y la adaptación de la estructura a esta nueva situación.

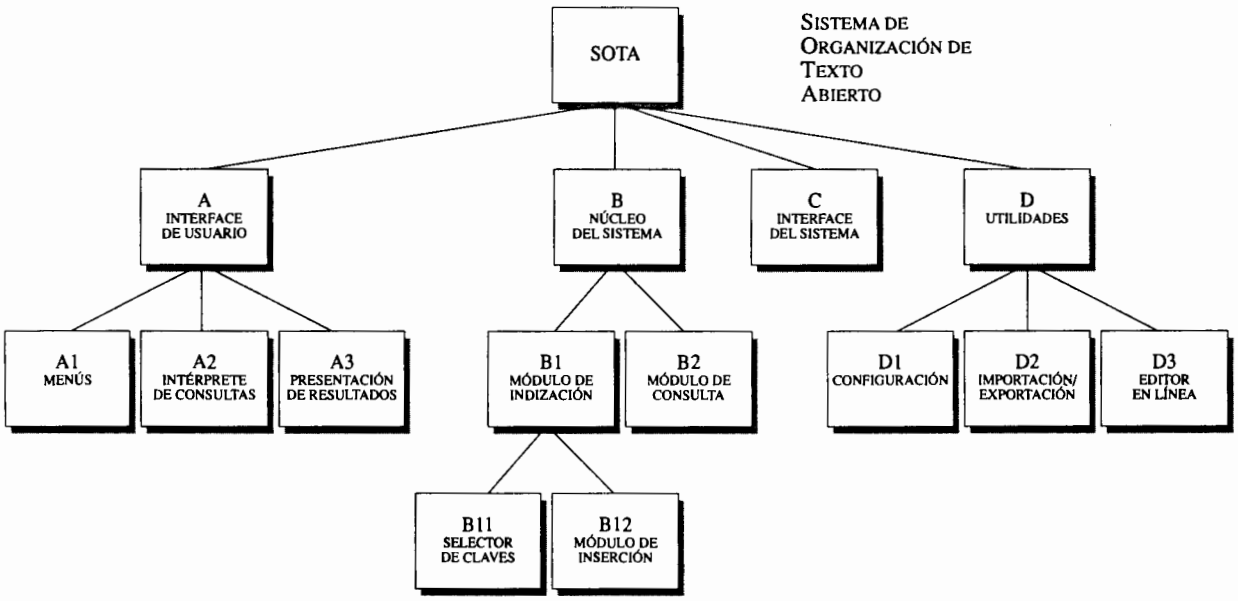
5. Proyectos de desarrollo de software

Toda institución, pública o privada, maneja un volumen elevado de documentos. En los sistemas informáticos que están sustentados en bases de datos de formato fijo, la mecanización ha de realizarse mediante un proceso de recopilación y posterior codificación de sus documentos. Este proceso es demasiado lento y costoso, lo que ha motivado muchos de los problemas vinculados con la sistematización de las instituciones. Una solución que evita los procesos de recopilación y codificación consiste en utilizar bases de datos que almacenen los documentos tal y como se presentan en la realidad.

5.1. *Gestión de Bases de Datos Documentales de Texto Libre. (Proyecto SOTA)*

Se pretende desarrollar un sistema para la indización de documentos textuales débilmente estructurados, o incluso sin estructura definida, que presente un alto grado de flexibilidad en cuanto a los formatos de los documentos permitidos; en cuanto a las modalidades de interrogación posibles, aspiramos a que: 1) sea adaptable a una amplia gama de configuraciones de recursos informáticos, y 2) transportable entre los entornos operativos más populares con un mínimo esfuerzo de programación.

Es necesario conjugar objetivos divergentes, tales como el conseguir un sistema que proporcione una gran potencia y flexibilidad de interrogación, sin que al mismo tiempo resulte demasiado avaricioso en cuanto a consumo de recursos. De ahí la importancia de que sea capaz de adaptar sus prestaciones a una amplia gama de configuraciones de recursos.



Otro objetivo fundamental es que al usuario le resulte cómodo trabajar con el producto, lo que como contrapartida supone la necesidad de aplicar un gran esfuerzo en el desarrollo de todas aquellas partes que interactúan con el usuario.

Debe preverse la inclusión de imágenes asociadas con los textos.

A efectos de estructura programática, el proyecto puede considerarse despiezado en cuatro módulos entre los que ha de conseguirse un máximo grado de independencia. Estos cuatro bloques serían los siguientes:

A) Interface de usuario.

B) Núcleo del sistema.

C) Interface del sistema.

D) Módulo de herramientas.

A continuación se describen con mayor detalle cada una de las partes y sus principales subapartados.

Módulo A. *Interface de usuario.* Centraliza la relación del usuario con el sistema fruto del proyecto. Se prefiere que sea lo más amigable posible, con menús desplegables, ratón, ventanas...

A1. *Menús.* Será un módulo abierto que llevará el peso de gestionar el aspecto de la interface.

A2. *Intérprete de consultas.* Se encargará de recibir las peticiones del usuario, formuladas en un lenguaje de interrogación, interpretarlas y pasarlas al correspondiente sector de consulta del núcleo. Debe permitirse al usuario formular sus interrogaciones de una manera flexible (si es posible "casi" en lenguaje natural).

A3. *Módulo de presentación de resultados.* Opciones de pantalla y de impresora, resultados totales y parciales, selección rápida de resultados a exponer y formas de presentación, hipertexto...

Modulo B. *Núcleo.* Es la parte central del proyecto que abarca todo lo relacionado con la indización, desde la creación a la utilización de los índices.

B1. *Indización.* Es el módulo al que se le encomienda la creación de los índices, que habrán de estar orientados a beneficiar los tipos de consulta más frecuentes, pero sin imposibilitar otros. Si los recursos lo admiten podrá haber múltiples índices. La estructura de indización constará de dos secciones: la que almacena los términos clave que consiente su búsqueda para las distintas peticiones, y otra, asociada a cada término que establece los lugares de aparición de estos en los documentos.

B1.1. *Selector de claves.* Analizará los documentos examinando los elementos indizables para pasárselos al módulo de inserción.

B1.2. *Módulo de inserción.* Se encargará de llevar a cabo la inserción efectiva de los términos clave en la estructura de índice.

B2. *Módulo de consulta.* Es el que se ocupa de recibir las peticiones tratadas por el intérprete de consultas y realizar el rastreo en los índices; pasa los resultados al módulo de presentación ubicado en la interface de usuario.

Módulo C. *Interface del sistema.* En este ámbito debe concentrarse todo lo que dependa del sistema, de modo que sea la única parte que precise ser modificada al trasladarse de un entorno a otro.

Módulo D. *Utilidades.* Este bloque incluye un conjunto diverso de herramientas de interés por aportar grados de apertura o de comodidad, o quizá por otras razones, pero que no influyen en el funcionamiento básico del proyecto ni en su presentación ante el usuario; tales instrumentos se emplean en la importación y exportación de formatos, edición en línea, tratamiento de imágenes...

D1. *Configuración.* Agrupa a todos los artificios de configuración del sistema que posibilitan adaptarlo a los recursos disponibles.

D2. *Importación/Exportación.* Está formado por módulos encargados de abrir este proyecto a múltiples generadores de documentos.

D3. *Editor en línea.* Compatible ASCII y dotado de las funciones básicas.

5.2. *Herramientas de Ayuda al Análisis de Documentos. (Proyecto Frectext)*

La elaboración de documentos es un proceso creativo que exige del escritor una notable capacidad de abstracción y estructuración para asegurar una calidad aceptable en el resultado. El autor no es autosuficiente sino que debe asistirse de la consulta de textos complementarios. Hasta el momento, algunos de estos procesos se continúan desarrollando manualmente con los inconvenientes propios de tal procedimiento de trabajo. La introducción de los ordenadores en el campo de la edición documental ha variado en cierta forma este concepto, ya que hoy día se pueden encontrar herramientas que agilicen parte de estas tareas. El inconveniente que presentan, entre otros, la corrección ortográfica o los diccionarios de sinónimos es que están hechos y pensados para el lenguaje inglés y no para el castellano, lo que hace que sus adaptaciones no resuelvan correctamente el problema.

Lo que se pretende con este trabajo es aportar algunas herramientas de soporte en la producción documental en castellano que incorporen características novedosas y eficientes. En el momento actual de su desarrollo integra varios instrumentos como son: búsquedas complejas en diccionario de significados, conjugación de todos los verbos de la gramática castellana, así como información estadística del documento en edición; este último apartado se puede desglosar en: ordenación de las palabras del documento con su frecuencia, alfabéticamente y por frecuencia creciente; muestra, además, las palabras más parecidas a una dada según distintos criterios de similitud; selección de palabras y destacado de las mismas; agrupamiento de las formas verbales presentes en el texto, etc. Todo se halla integrado en un entorno de trabajo amigable, basado en menús desplegables y ventanas configurables, con un sencillo procesador de textos, al tiempo que se continúa trabajando en el desarrollo de otras herramientas como el tratamiento de la sinonimia y la corrección de errores tipográficos.

Frectext define un entorno de trabajo interactivo, que responde en cada instante a las especificaciones indicadas por el usuario.

Las características más relevantes que presenta *Frectext* son las siguientes:

- Análisis del documento, presentando estadísticas donde se reflejan las frecuencias de las distintas palabras del mismo, ordenadas por frecuencia y alfabéticamente.

- Capacidad de seleccionar palabras del documento, así como las formas verbales existentes en el texto de un verbo señalado, reuniéndolas en un conjunto para su tratamiento posterior.

- Agrupamiento de las formas verbales presentes en el documento.

- Se muestran las palabras más parecidas a una dada según distintos criterios de similitud: Levenshtein, Secuencia Común Más Larga (SCML), y la secuencia Común Más Larga Contigua (SCMLC).

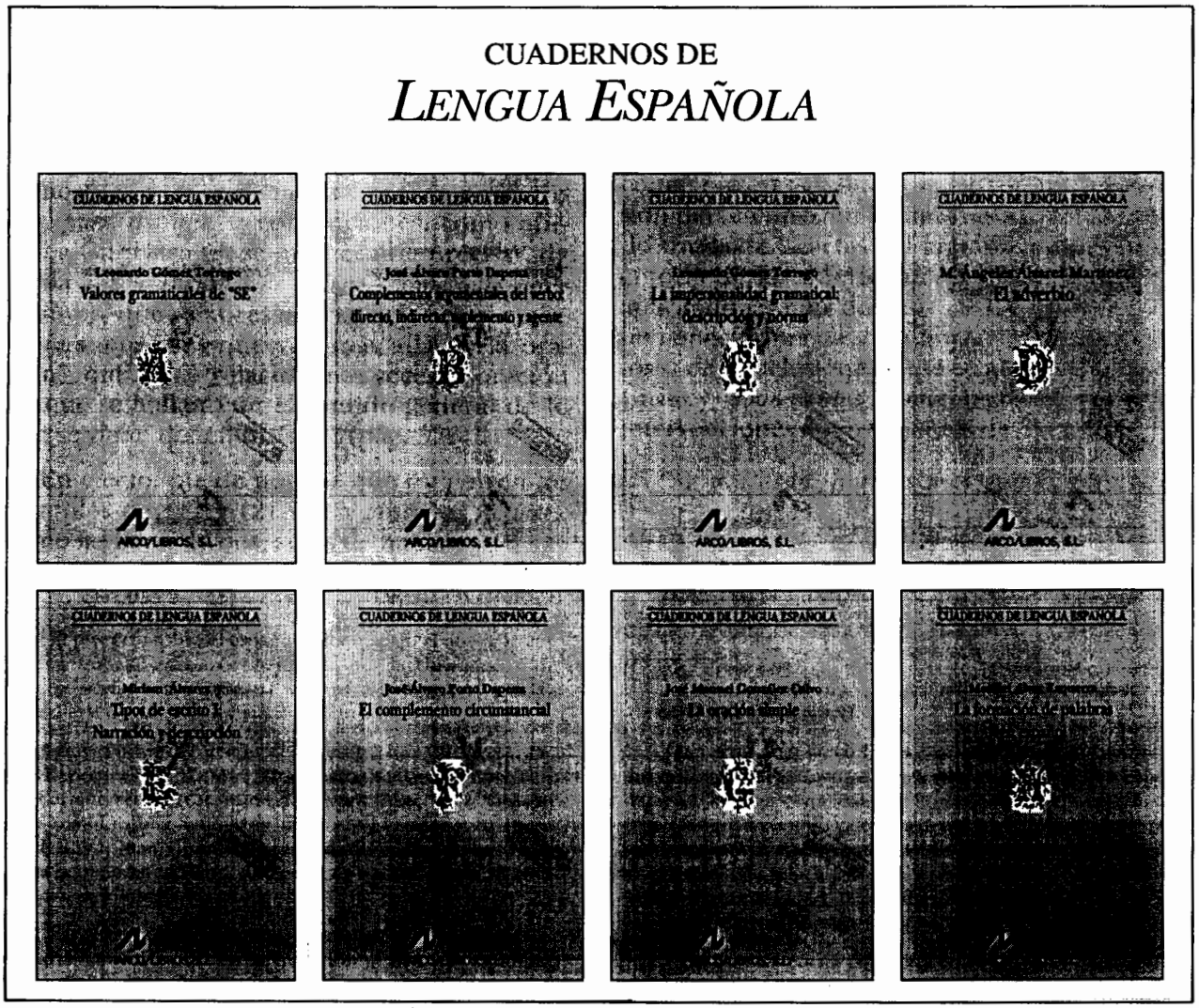
- Definición de un conjunto de palabras "vacías" cuyo tratamiento se excluye de las áreas especificadas por el usuario.

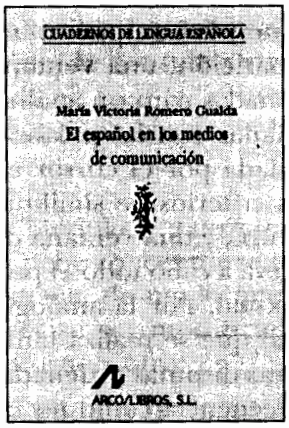
• Edición del documento con las características básicas de un tratamiento de texto: inserción y borrado, búsqueda y reemplazamiento, y las operaciones más comunes sobre bloques.

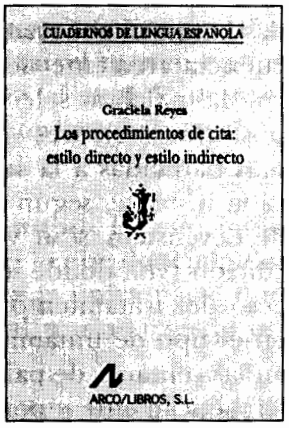
• Procesamiento dinámico de las palabras señaladas o insertadas en cada instante sobre las que se realiza la operación de búsqueda de las más parecidas.

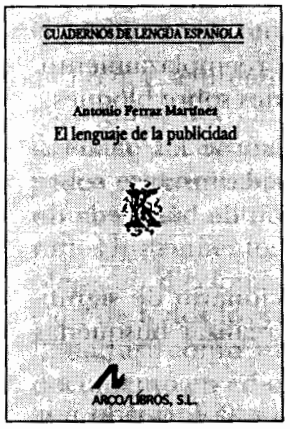
• Integración de un diccionario de significados en el que se pueden realizar búsquedas complejas.

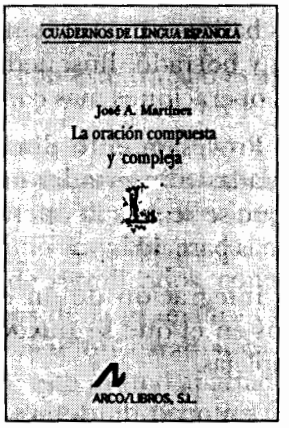
El entorno de desarrollo está formado por un conjunto de ventanas configurables, que se visualizan en la pantalla y que constituyen el soporte para el proceso del documento. Las ventanas se identifican por un título situado en el borde superior y un número de ventana localizado en la esquina inferior derecha —presentan distintos aspectos del análisis—. Así, hay dos ventanas que contienen todas las palabras distintas con sus frecuencias, ordenadas la una alfabéticamente y la otra por su frecuencia creciente de aparición; una ventana con las palabras seleccionadas para un análisis posterior (VS); tres ventanas con las palabras más parecidas a la señalada por el cursor en cada instante, según los criterios de similitud de Levenstein, SCML y SCMLC; y una ventana de Edición (VE), donde se lleva a cabo todo el proceso del tratamiento textual. Por la analogía en el tipo de tratamiento que se realiza tanto en las ventanas de palabras distintas ordenadas alfabéticamente y por frecuencia como en las ventanas de palabras ordenadas por similitud, se identifican como ventanas de Tratamiento (VT). Además, en todo momento hay una ventana Activa (VA), caracterizada por tener el recuadro del borde con líneas dobles —diferenciándose del resto de las ventanas que lo tienen simple— afectando a las demás las acciones que se realizan sobre ella.

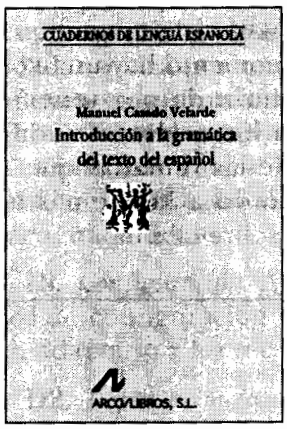


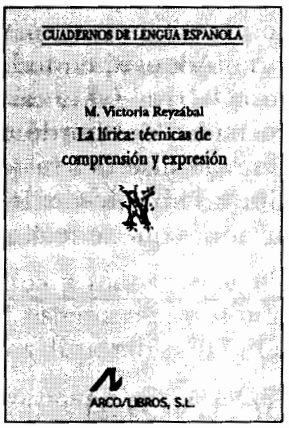

CUADERNOS DE LENGUA ESPAÑOLA
María Victoria Romero Gualda
El español en los medios de comunicación
ARCO/LIBROS, S.L.

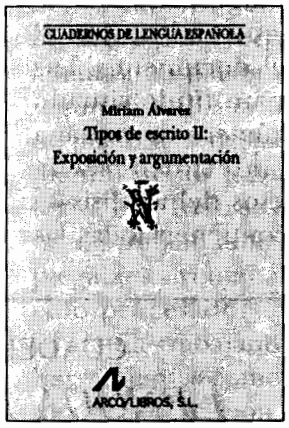

CUADERNOS DE LENGUA ESPAÑOLA
Graciela Reyes
Los procedimientos de cita: estilo directo y estilo indirecto
ARCO/LIBROS, S.L.

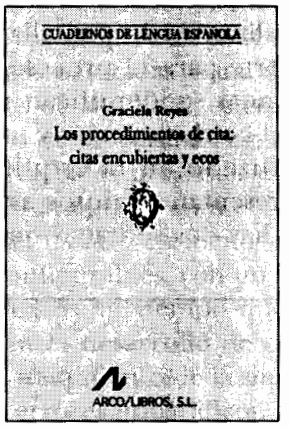

CUADERNOS DE LENGUA ESPAÑOLA
Antonio Ferraz Martínez
El lenguaje de la publicidad
ARCO/LIBROS, S.L.

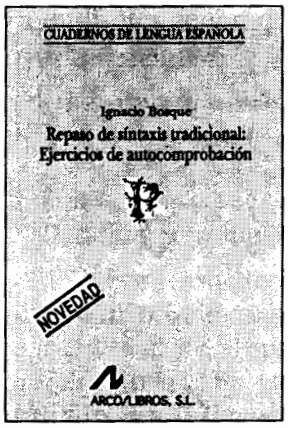

CUADERNOS DE LENGUA ESPAÑOLA
José A. Martínez
La oración compuesta y compleja
ARCO/LIBROS, S.L.

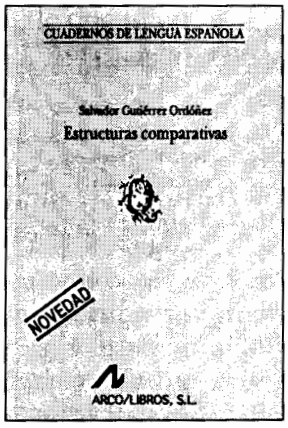

CUADERNOS DE LENGUA ESPAÑOLA
Manuel Casado Velarde
Introducción a la gramática del texto del español
ARCO/LIBROS, S.L.

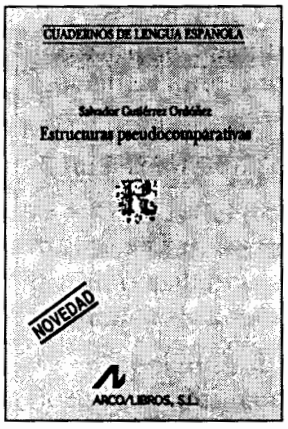

CUADERNOS DE LENGUA ESPAÑOLA
M. Victoria Reyzábal
La lírica: técnicas de comprensión y expresión
ARCO/LIBROS, S.L.

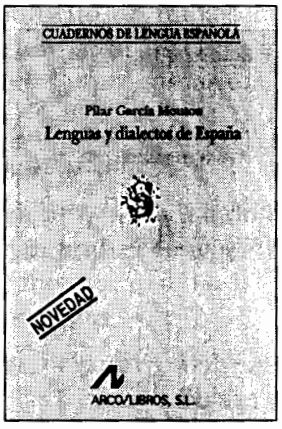

CUADERNOS DE LENGUA ESPAÑOLA
Miriam Álvarez
Tipos de escrito II: Exposición y argumentación
ARCO/LIBROS, S.L.

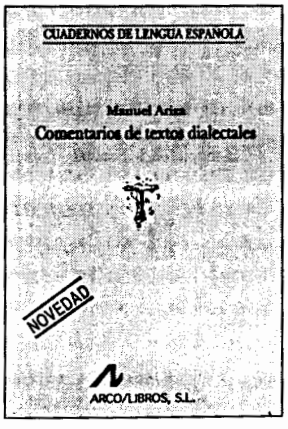

CUADERNOS DE LENGUA ESPAÑOLA
Graciela Reyes
Los procedimientos de cita: citas encubiertas y ecos
ARCO/LIBROS, S.L.

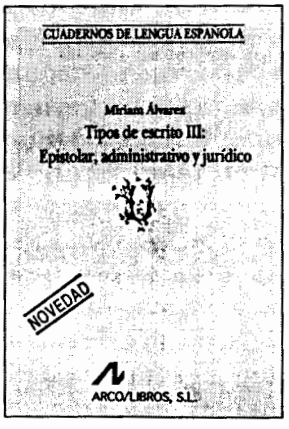

CUADERNOS DE LENGUA ESPAÑOLA
Ignacio Bosque
Repaso de sintaxis tradicional: Ejercicios de autocomprobación
NOVEDAD
ARCO/LIBROS, S.L.

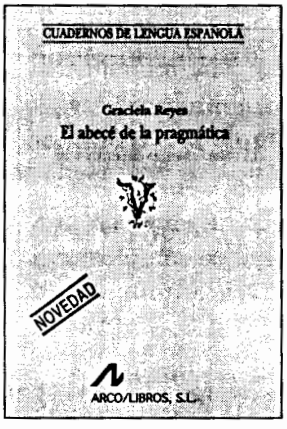

CUADERNOS DE LENGUA ESPAÑOLA
Salvador Gutiérrez Ordóñez
Estructuras comparativas
NOVEDAD
ARCO/LIBROS, S.L.

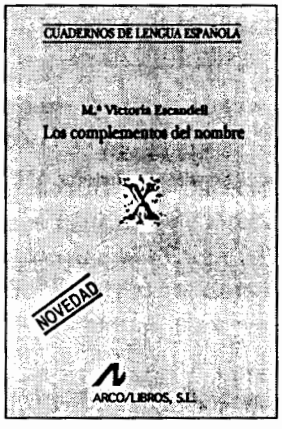

CUADERNOS DE LENGUA ESPAÑOLA
Salvador Gutiérrez Ordóñez
Estructuras pseudocomparativas
NOVEDAD
ARCO/LIBROS, S.L.


CUADERNOS DE LENGUA ESPAÑOLA
Pilar García Mouton
Lenguas y dialectos de España
NOVEDAD
ARCO/LIBROS, S.L.


CUADERNOS DE LENGUA ESPAÑOLA
Manuel Ariza
Comentarios de textos dialectales
NOVEDAD
ARCO/LIBROS, S.L.


CUADERNOS DE LENGUA ESPAÑOLA
Miriam Álvarez
Tipos de escrito III: Epistolar, administrativo y jurídico
NOVEDAD
ARCO/LIBROS, S.L.


CUADERNOS DE LENGUA ESPAÑOLA
Graciela Reyes
El abecé de la pragmática
NOVEDAD
ARCO/LIBROS, S.L.


CUADERNOS DE LENGUA ESPAÑOLA
M.ª Victoria Escandell
Los complementos del nombre
NOVEDAD
ARCO/LIBROS, S.L.